L'UNIVERSITÉ
DE MÉDECINE ET DE PHARMACIE
„VICTOR BABEŞ" DE TIMIŞOARA

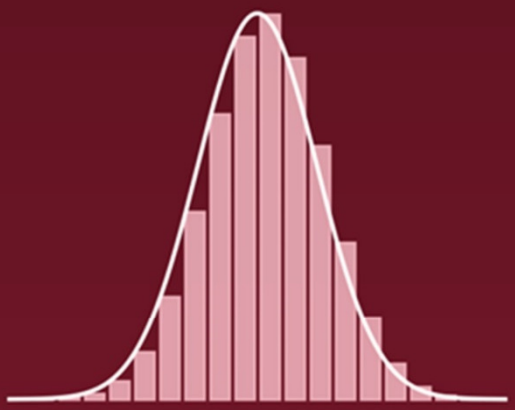
CĂLIN MUNTEAN

BIOSTATISTIQUE APPLIQUÉE

EN MÉDECINE GÉNÉRALE

Données, méthodes, analyse et décision. Horizon 2026

Cours et travaux pratiques pour les étudiants en médecine





ROC

IC 95%	128,96-133,30
RR	2,62
AUC	0,85

MANUALE

Éditions „Victor Babeş"
Timişoara, 2026

Editura „Victor Babeş”

Piața Eftimie Murgu nr. 2, cam. 316, 300041 Timișoara

Tel./Fax 0256 495 210

e-mail: evb@umft.ro

https://www.umft.ro/ro/organizare-evb/

Director: Prof. univ. dr. Sorin Ursoniu

Colecția: MANUALE

Coordonatori colecție: **Prof. univ. dr. Codruța Șoica**

Prof. univ. dr. Daniel Lighezan

Referent științific: Prof. univ. dr. Daliborca Vlad

Indicativ CNCSIS: 324

© 2026

Toate drepturile asupra acestei ediții sunt rezervate. Reproducerea parțială sau integrală a textului, pe orice suport, fără acordul scris al autorilor este interzisă și se va sancționa conform legilor în vigoare.

ISBN 978-606-786-654-4

Les patients des exemples sont fictifs. Le jeu de données pédagogique "mg2026" a été généré par simulation et ne provient d'aucun dossier de patient. Les chiffres tirés d'articles publiés sont reproduits tels qu'ils figurent dans les résumés indexés dans PubMed. Les informations sur les logiciels, les versions et la réglementation reflètent la situation au 9 octobre 2026. Cet ouvrage ne remplace ni les recommandations de pratique clinique ni la réglementation en vigueur, qui doivent être consultées dans leur version actualisée.

Table des matières

Préface 11

Remerciements 12

Comment lire cet ouvrage 12

Liste des abréviations 14

Partie I. Cours 16

Chapitre 1. Introduction à la biostatistique : concepts fondamentaux, rôle dans la recherche médicale et éléments de probabilités 17

1.1 Pourquoi un médecin a besoin de la biostatistique 17

1.2 Concepts de base : unité statistique, population, échantillon 19

1.3 Le jeu de données pédagogique mg2026 20

1.4 L'échantillonnage : représentativité et méthodes 21

1.5 Les biais (erreurs systématiques) 24

1.6 Statistique descriptive et statistique inférentielle 26

1.7 Les étapes d'une étude statistique et la question de recherche 28

1.8 Les données à l'ère du dossier électronique : le lien avec l'informatique médicale 29

1.9 Éléments de théorie des probabilités 33

1.10 Les logiciels utilisés dans ce cours 38

1.11 Problèmes résolus 40

1.12 Exercices proposés 43

Chapitre 2. Description et organisation des données : types de variables, indicateurs descriptifs et représentations graphiques 48

2.1 Du patient au tableau de données 48

2.2 Les tableaux de fréquences 53

2.3 Les indicateurs de tendance centrale 55

2.4 Les indicateurs de dispersion 58

2.5 Quantiles, percentiles, quartiles et résumé en cinq valeurs 60

2.6 La forme de la distribution : un premier aperçu de l'asymétrie et de l'aplatissement 62

2.7 Les représentations graphiques 62

2.8 Le tableau descriptif d'un article : le "Tableau 1" 66

2.9 L'analyse descriptive complète du jeu de données mg2026 68

2.10 Problèmes résolus 70

2.11 Exercices proposés 74

Chapitre 3. Les lois statistiques : loi normale, vérification de la normalité, asymétrie, aplatissement, score z et transformation des données 79

3.1 Variables aléatoires et lois de probabilité 79

3.2 Deux lois discrètes utiles en médecine 81

3.3 La loi normale 82
3.4 La loi normale centrée réduite et le score z 85
3.5 Les intervalles de référence en biologie médicale 87
3.6 Les distributions d'échantillonnage : t, khi-deux et F 89
3.7 Le théorème central limite 90
3.8 La vérification de la normalité 91
3.9 La transformation des données 96
3.10 Les valeurs extrêmes 99
3.11 Distributions qui se chevauchent : du score z aux décisions cliniques 101
3.12 Problèmes résolus 102
3.13 Exercices proposés 106
Chapitre 4. Estimation des paramètres et intervalles de confiance ; principes du calcul de la taille de l'échantillon 110
4.1 Estimation ponctuelle et estimation par intervalle 110
4.2 La distribution d'échantillonnage de la moyenne et l'erreur standard 111
4.3 L'intervalle pour les valeurs individuelles et l'intervalle pour la moyenne 112
4.4 L'intervalle de confiance de la moyenne avec la loi t 113
4.5 L'interprétation correcte d'un intervalle de confiance 115
4.6 La largeur de l'intervalle de confiance et la taille de l'échantillon 116
4.7 L'intervalle de confiance d'une proportion 118
4.8 L'intervalle de confiance de la différence de deux moyennes et de deux proportions 120
4.9 Intervalles de confiance pour des données asymétriques 121
4.10 Intervalles de confiance des rapports (OR, RR) et de l'aire ROC 121
4.11 Le bootstrap expliqué simplement 123
4.12 L'intervalle de confiance et la valeur p 124
4.13 Principes du calcul de la taille de l'échantillon 125
4.14 Problèmes résolus 128
4.15 Exercices proposés 131
Chapitre 5. Les bases des tests statistiques : hypothèses, seuil de signification, erreurs de type I et II, décision et interprétation 135
5.1 De la question clinique à l'hypothèse testable 135
5.2 La logique du test : la statistique de test et sa distribution sous H0 137
5.3 La valeur p 138
5.4 Le seuil de signification α et la règle de décision 139
5.5 Les erreurs de type I et II, la puissance du test 141
5.6 La taille de l'effet et la signification clinique 144
5.7 Trois voies vers la même décision : la valeur critique, la valeur p et l'intervalle de confiance 147

5.8 L'algorithme d'un test statistique 147
5.9 Les tests z et t pour une seule moyenne 148
5.10 Les comparaisons multiples et l'inflation de l'erreur de type I 150
5.11 p-hacking, HARKing et préenregistrement 152
5.12 Équivalence et non-infériorité, en bref 154
5.13 Comment rapporter un test statistique 154
5.14 Problèmes résolus 155
5.15 Exercices proposés 158
Chapitre 6. Tests statistiques paramétriques usuels : test t pour échantillons appariés et indépendants, analyse de la variance (ANOVA) 162
6.1 La famille des tests paramétriques 162
6.2 Comparaison d'une moyenne à une valeur donnée : test z et test t 163
6.3 Le test t pour échantillons appariés 163
6.4 Le test t pour échantillons indépendants 167
6.5 Analyse de la variance (ANOVA) à un facteur 169
6.6 Les tests post hoc 173
6.7 ANOVA à mesures répétées 175
6.8 ANOVA à deux facteurs et ANCOVA, en bref 177
6.9 Rapporter les tests t et les ANOVA 178
6.10 Problèmes résolus 178
6.11 Exercices proposés 182
Chapitre 7. Tests statistiques non paramétriques usuels : U de Mann-Whitney, Wilcoxon, Kruskal-Wallis, Friedman et test du khi-deux 186
7.1 Quand on ne peut plus utiliser les tests paramétriques 186
7.2 Les rangs : la matière première des tests non paramétriques 188
7.3 Le test U de Mann-Whitney 189
7.4 Le test de Wilcoxon pour données appariées et le test des signes 193
7.5 Le test de Kruskal-Wallis 195
7.6 Le test de Friedman 198
7.7 Le test du khi-deux pour les tableaux de contingence 200
7.8 Variantes du test du khi-deux 205
7.9 Au-delà de la valeur p : phi, V de Cramér et autres mesures d'association 207
7.10 Rapporter les résultats des tests non paramétriques et du test du khi-deux 208
7.11 Problèmes résolus 209
7.12 Exercices proposés 214
Chapitre 8. Sélection et application des tests statistiques : critères de choix, résolution de problèmes et études de cas 218
8.1 Le choix du test est une décision de conception, non de calcul 218

8.2 L'algorithme de choix du test 219
8.3 Le tableau de synthèse des tests 221
8.4 L'unité d'analyse et l'indépendance des observations 223
8.5 Le test se fixe avant de voir les données 224
8.6 Erreurs fréquentes dans le choix et l'application du test 225
8.7 Rapporter un résultat selon un format standardisé (a-g) 226
8.8 Problèmes résolus 227
8.9 Reconnaître le test erroné : dix descriptions de type "Méthodes" 236
8.10 Lecture critique de la section Méthodes : articles de collègues de la discipline 238
8.11 Exercices proposés 243
Chapitre 9. Analyse des relations entre variables : la corrélation et la régression linéaire pour les variables numériques et ordinales 247
9.1 L'analyse bivariée : deux variables mesurées chez le même patient 247
9.2 La covariance : un premier pas vers une mesure de l'association 249
9.3 Le coefficient de corrélation de Pearson 249
9.4 La corrélation pour des données ordinales ou asymétriques : les coefficients de Spearman et de Kendall 254
9.5 Corrélation ne signifie pas causalité 257
9.6 Corrélation et concordance : la méthode de Bland-Altman 258
9.7 La régression linéaire simple 259
9.8 La régression linéaire multiple : introduction 263
9.9 La régression logistique : une passerelle vers le chapitre 10 267
9.10 Problèmes résolus 268
9.11 Exercices proposés 271
Chapitre 10. Applications statistiques en épidémiologie : types d'études observationnelles et analyse du risque 275
10.1 Le facteur de risque 275
10.2 La mesure de la fréquence de la maladie 276
10.3 Les types d'études épidémiologiques 278
10.4 Le tableau de contingence 2 x 2 282
10.5 Le risque relatif 282
10.6 Le rapport de cotes 283
10.7 Intervalles de confiance et tests statistiques 284
10.8 Mesures d'impact : le risque attribuable, la fraction attribuable, le NST et le NNH 286
10.9 La confusion, la stratification et la modification de l'effet 287
10.10 De l'association à la causalité 290
10.11 Problèmes résolus 292
10.12 Exercices proposés 297

Chapitre 11. Méthodes statistiques dans la décision médicale : l'analyse bayésienne et la médecine fondée sur les preuves 301
11.1 L'incertitude dans la décision clinique 301
11.2 La probabilité conditionnelle revisitée 302
11.3 Le théorème de Bayes 304
11.4 Cotes, rapports de vraisemblance et forme de Bayes avec les cotes 306
11.5 Les tests séquentiels et la combinaison des tests 310
11.6 Le diagnostic différentiel avec plusieurs maladies 312
11.7 Les seuils de décision : tester, traiter 314
11.8 Le dépistage en médecine générale : l'exemple du FINDRISC 315
11.9 L'inférence bayésienne par rapport à l'inférence fréquentiste 316
11.10 Médecine fondée sur les preuves 319
11.11 Erreurs cognitives dans le raisonnement clinique 321
11.12 Communication du risque au patient 321
11.13 Problèmes résolus 322
11.14 Exercices proposés 326
Chapitre 12. Évaluation de la performance des tests diagnostiques : sensibilité, spécificité, valeurs prédictives et courbe ROC 331
12.1 Test à évaluer et standard de référence 331
12.2 Le tableau 2 x 2 et les indicateurs de performance 332
12.3 Reconstitution d'un tableau 2 x 2 336
12.4 Dépendance des valeurs prédictives à l'égard de la prévalence 336
12.5 Biais dans les études d'exactitude diagnostique 337
12.6 Tests à résultat continu et choix du seuil 339
12.7 La courbe ROC 340
12.8 Accord entre évaluateurs 345
12.9 Évaluation des classificateurs d'intelligence artificielle 346
12.10 Publication des études d'exactitude diagnostique : STARD 2015 et STARD-AI 346
12.11 Problèmes résolus 347
12.12 Exercices proposés 352
Chapitre 13. Les essais cliniques : classification, conception, protocole, outils de recueil des données et analyse intégrée des données biomédicales 356
13.1 Qu'est-ce qu'un essai clinique et qu'exige-t-il de la statistique 356
13.2 Classification des études cliniques 357
13.3 Les principes méthodologiques d'un essai randomisé 361
13.4 Le protocole, l'enregistrement et le cadre éthique 365
13.5 Les outils de recueil des données 367
13.6 La gestion et le nettoyage des données 372

13.7 La publication des résultats : CONSORT 2025 373
13.8 Analyse intégrée de l'étude mg2026 : un exemple complet 374
13.9 Problèmes résolus 381
13.10 Exercices proposés 385
Chapitre 14. Exemples récapitulatifs : interprétation intégrée des résultats statistiques 389
14.1 La carte de la discipline 389
14.2 Comment lire la statistique d'un article médical 391
14.3 Études de cas intégrées 393
14.4 Comment rédiger le paragraphe des résultats 405
14.5 Quarante affirmations : vrai ou faux 406
14.6 Tableau récapitulatif des formules 409
14.7 Problèmes résolus 411
14.8 Exercices proposés 413
Partie II. Travaux pratiques 420
Travaux pratiques 1. Introduction à la biostatistique et aux outils numériques : Google Sheets, JASP et MedCalc 421
Travaux pratiques 2. Description et organisation des données : le questionnaire médical, les tableaux de fréquences et les graphiques 430
Travaux pratiques 3. Distributions et test de la normalité 438
Travaux pratiques 4. Estimation et intervalles de confiance : l'effet de la taille de l'échantillon 446
Travaux pratiques 5. Test des hypothèses statistiques : formulation des hypothèses, erreurs et décision 453
Travaux pratiques 6. Tests paramétriques : le test t et l'ANOVA 461
Travaux pratiques 7. Tests non paramétriques : Mann-Whitney, Wilcoxon, Kruskal-Wallis, Friedman et khi-deux 469
Travaux pratiques 8. Choix du test statistique et présentation standardisée des résultats 477
Travaux pratiques 9. Analyse de corrélation et de régression : Pearson, Spearman et régression linéaire simple 483
Travaux pratiques 10. Analyses épidémiologiques : le risque relatif, le rapport de cotes et l'interprétation des études observationnelles 490
Travaux pratiques 11. Analyse bayésienne : le raisonnement probabiliste dans le diagnostic et le dépistage 497
Travaux pratiques 12. Performance des tests diagnostiques : sensibilité, spécificité, valeurs prédictives et courbe ROC 503
Travaux pratiques 13. Conception d'une mini-étude clinique et analyse intégrée d'un jeu de données complet 510
Travaux pratiques 14. L'examen pratique : sujet type, barème et plan de préparation 517
Annexe A. Réponses aux questions d'autoévaluation et aux exercices proposés 527
Chapitre 1 527

Chapitre 2 528
Chapitre 3 529
Chapitre 4 530
Chapitre 5 530
Chapitre 6 531
Chapitre 7 533
Chapitre 8 533
Chapitre 9 536
Chapitre 10 537
Chapitre 11 538
Chapitre 12 539
Chapitre 13 540
Chapitre 14 541
Annexe B. Glossaire 543
Annexe C. Le jeu de données pédagogique mg2026 : scénario, dictionnaire des données et erreurs du fichier brut 567
C.1 Le scénario 567
C.2 Les fichiers 567
C.3 Le dictionnaire des données 568
C.4 Les erreurs du fichier brut et leur correction 570
C.5 Comment le jeu de données a été généré 572
C.6 Comment citer le jeu de données dans les exposés et les devoirs 573
Annexe D. Fiche de formules 575
D.1 Chapitre 1. Probabilités 575
D.2 Chapitre 2. Description des données 576
D.3 Chapitre 3. Loi normale, score z, transformations 577
D.4 Chapitre 4. Estimation, intervalles de confiance, taille de l'échantillon 578
D.5 Chapitre 5. Bases des tests d'hypothèses 579
D.6 Chapitre 6. Tests paramétriques : t et ANOVA 580
D.7 Chapitre 7. Tests non paramétriques et test du khi-deux 581
D.8 Chapitre 8. Choix du test (repères numériques) 582
D.9 Chapitre 9. Corrélation et régression 582
D.10 Chapitre 10. Épidémiologie et analyse du risque 583
D.11 Chapitre 11. Bayes et médecine fondée sur les preuves 584
D.12 Chapitre 12. Tests diagnostiques et courbe ROC 585
D.13 Chapitre 13. Essais cliniques (formules de planification) 586
Annexe E. Tables statistiques 587

E.1 Loi normale centrée réduite : fonction de répartition Φ(z) 587
E.2 Valeurs critiques de la loi normale centrée réduite 588
E.3 Valeurs critiques de la loi de Student (t) 588
E.4 Valeurs critiques de la loi du khi-deux (χ^2) 589
E.5 Valeurs critiques de la loi de Fisher (F) pour $\alpha = 0,05$ 590
E.6 Valeurs critiques du coefficient de corrélation de Pearson (r) 591
E.7 Valeurs critiques approximatives du coefficient de corrélation de Spearman (rho) 592
E.8 Table de 1 000 chiffres aléatoires 593
Annexe F. QCM d'entraînement pour l'examen 595
Comment les questions sont construites et comment elles sont notées 595
Test 1 596
Test 2 604
Test 3 612
Réponses et explications 620
Bibliographie 634

Préface

Un médecin de famille reçoit le matin les résultats de quarante analyses, lit à midi de nouvelles recommandations sur l'hypertension et, le soir, un article sur un antidiabétique. À chacun de ces moments, il fait de la statistique sans s'en rendre compte. Il se demande si une glycémie de 118 mg/dL est un écart dû au hasard ou un signal, si la baisse de pression artérielle rapportée dans l'étude compterait pour ses patients et dans quelle mesure il peut se fier à un score de risque rempli en cinq minutes. La biostatistique n'est pas une discipline de calculs. C'est la manière dont la médecine mesure son incertitude.

Le présent ouvrage est le cours de biostatistique destiné aux étudiants de première année de la Faculté de médecine de l'Université de médecine et de pharmacie "Victor Babeș" de Timișoara. Il prolonge le manuel d'informatique médicale du premier semestre [1], dont il conserve la forme et les règles de travail, et suit la fiche de la discipline (syllabus) pour l'année universitaire 2026-2027 : 14 thèmes de cours et 14 séances de travaux pratiques, à raison d'une paire par semaine, avec un examen écrit de type QCM, au format du concours de résidanat roumain, et un examen pratique en salle de travaux pratiques. J'ai choisi le titre "Biostatistique appliquée en médecine générale" parce que tous les exemples viennent de la médecine que pratiquera tout diplômé, quelle que soit la spécialité qu'il choisira plus tard : le cabinet de médecine générale, la médecine interne, le diabète et les maladies métaboliques, l'hypertension, les maladies respiratoires, l'hépatologie, la cardiologie, la pédiatrie, l'obstétrique, les urgences et les maladies infectieuses.

La première partie du cours, les chapitres 1-4, pose les fondations : ce que sont les données et d'où elles viennent, comment on les décrit, ce qu'est une distribution et comment on estime un paramètre de la population à partir d'un échantillon. Le chapitre 1 fait aussi le lien avec l'informatique médicale du premier semestre, car une analyse statistique ne peut pas être meilleure que les données qu'elle traite, et les données de santé exigent des règles d'éthique et de protection que le médecin doit connaître. Les chapitres 5-8 traitent des tests d'hypothèses : la logique du test, les tests paramétriques et non paramétriques courants et, surtout, le choix du test approprié, la source d'erreurs la plus fréquente à l'examen comme dans les articles publiés. Les chapitres 9-12 portent sur l'association et la décision : la corrélation et la régression, les mesures du risque en épidémiologie, le raisonnement bayésien et l'évaluation des tests diagnostiques. Le chapitre 13 montre comment on conçoit une étude clinique et comment on construit un formulaire de recueil des données, et le chapitre 14 rassemble le tout dans des exemples d'interprétation intégrée.

La biostatistique s'apprend difficilement à partir de définitions et beaucoup plus facilement à partir d'exemples que le lecteur reconnaît. C'est pourquoi l'ouvrage utilise deux sources d'exemples. La première est le jeu de données pédagogique "mg2026", un essai randomisé simulé dans des cabinets de médecine générale, portant sur 160 patients à risque cardiométabolique suivis pendant six mois. Ce même jeu de données revient dans tous les chapitres et dans toutes les séances de travaux pratiques, de sorte que l'étudiant puisse recalculer la même moyenne, le même intervalle de confiance et le même test avec les logiciels gratuits de la salle de travaux pratiques, puis comparer son résultat à celui du livre.

La seconde source est la recherche publiée par les collègues de la discipline, le plus souvent en collaboration avec les cliniques de l'université. Les études sur l'élastographie hépatique et l'échographie de contraste [2,3], sur les comportements alimentaires et l'indice de masse corporelle chez les étudiants [4-6], sur l'implication numérique des patients diabétiques [7], sur l'évolution de la COVID-19 chez les patients atteints de diabète de type 1 et de type 2 [8], sur les différences entre femmes et hommes dans les urgences cardiovasculaires [9] ou sur la validation

de questionnaires en langue roumaine [10,11] sont lues dans le livre comme des études de cas : quelle question elles ont posée, quel schéma d'étude elles ont adopté, quel test elles ont utilisé, ce que disent leurs chiffres et ce qu'ils ne disent pas. L'étudiant qui apprend à lire ces articles le crayon à la main saura lire de la même façon n'importe quel article de la spécialité qu'il choisira.

Les outils sont ceux de la fiche de la discipline, tous gratuits ou sous licence académique : Google Forms et Google Sheets pour le recueil et l'organisation des données, JASP et MedCalc pour l'analyse, Epi Info pour les tableaux épidémiologiques et pour le travail hors ligne. Chaque séance de travaux pratiques indique les menus exacts et les résultats attendus, afin que l'étudiant puisse vérifier lui-même son travail.

J'ai conservé les règles de travail de l'ouvrage d'informatique médicale. Chaque référence a été effectivement retrouvée dans PubMed, dans CrossRef ou sur le site de l'institution émettrice avant d'être citée. Les chiffres des études publiées sont repris tels qu'ils figurent dans le résumé indexé. Les graphiques construits à partir du jeu de données pédagogique sont signalés comme tels. Les patients des exemples sont fictifs. Lorsqu'une information n'a pas pu être vérifiée, j'ai préféré ne pas l'écrire.

Remerciements

Ce livre doit beaucoup à mes collègues de département, avec qui j'enseigne la même discipline depuis de nombreuses années, avec qui j'ai rédigé les guides de travaux pratiques de la discipline [12-14] et aux côtés desquels j'ai travaillé à l'analyse statistique de nombreuses études cliniques. Pour la collaboration pédagogique et scientifique dont est né ce cours, je remercie M. Radu Moleriu, maître de conférences, docteur, Mme Anca Tudor, maître de conférences, docteur, Mme Mirela Frandeș, chef de travaux, docteur, Mme Lavinia Moleriu, chef de travaux, docteur, M. Flavius Cioca, assistant universitaire, docteur, et Mme Raluca Lupușoru, assistante universitaire, docteur. Bon nombre des exemples de l'ouvrage proviennent d'articles auxquels ils ont contribué, et bon nombre des problèmes de travaux pratiques ont été peaufinés lors de nos discussions en fin de séance. Je remercie aussi les étudiants de ces dernières années, dont les questions, parfois dérangeantes, ont déterminé ce qui méritait d'être expliqué plus en détail. Les erreurs qui subsistent sont les miennes.

Timișoara, octobre 2026

Călin Muntean

Comment lire cet ouvrage

L'ouvrage peut être lu dans l'ordre, semaine après semaine, mais ce n'est pas obligatoire. Les chapitres renvoient les uns aux autres, et la carte de la discipline du chapitre 14 montre comment ils s'articulent. Qui a besoin rapidement d'un thème, par exemple du calcul du rapport de cotes ou du choix entre le test t et le test de Mann-Whitney, peut partir directement du chapitre concerné et de la séance de travaux pratiques qui lui correspond.

Chaque chapitre de cours est construit de la même façon. Il commence par des objectifs opérationnels, formulés comme des actions que l'étudiant saura accomplir à la fin. Le contenu va des notions aux méthodes, avec des exemples calculés à la main, puis sur le jeu de données pédagogique. Les études de cas tirées de la recherche de la discipline sont placées dans des encadrés séparés. Viennent ensuite des problèmes résolus pas à pas et des exercices proposés, un encadré consacré aux pièges fréquents, c'est-à-dire aux erreurs observées à maintes reprises en travaux pratiques et à l'examen, puis un encadré avec les idées à retenir. Le chapitre se termine par dix questions de type QCM, au format de l'examen : question à réponse unique (QRU), avec une

seule bonne réponse, et question à réponses multiples (QRM), avec deux à quatre bonnes réponses. Les réponses et les solutions des exercices se trouvent à l'annexe A.

Les séances de travaux pratiques comportent un objectif, les outils avec la version utilisée, les étapes de travail avec les résultats attendus, des exercices supplémentaires, ce qu'il faut rendre avec un barème indicatif et une variante pour les salles sans réseau. Toutes utilisent le jeu de données "mg2026", fourni en deux versions : le fichier brut, où les erreurs de saisie ont été laissées volontairement, et le fichier nettoyé. Le dictionnaire des données se trouve à l'annexe C, le recueil de formules à l'annexe D, les tables statistiques à l'annexe E et trois tests QCM complets d'entraînement à l'annexe F.

Pour la préparation courante, je recommande trois étapes. On lit d'abord les objectifs. On parcourt ensuite le texte avec les figures, les tableaux et les problèmes résolus, en refaisant les calculs sur papier ou dans la feuille de calcul. Enfin, on résout les questions et les exercices sans regarder les réponses de l'annexe A. Les pièges fréquents méritent d'être lus deux fois, une fois avant la séance de travaux pratiques et une fois avant l'examen.

Les citations sont numérotées selon le style de Vancouver, dans l'ordre de première apparition, et la liste complète se trouve à la fin de l'ouvrage. Les abréviations sont regroupées dans une liste séparée, et les principaux termes sont brièvement définis dans le glossaire de l'annexe B, avec un renvoi au chapitre où ils sont expliqués. Le texte utilise la virgule décimale, comme en français (0,05 ; 131,13 mmHg), même si les logiciels d'analyse affichent souvent le point décimal.

Liste des abréviations

Liste des abréviations utilisées dans l'ouvrage, par ordre alphabétique. Pour les termes repris de l'anglais, la dénomination originale figure entre parenthèses.

Abréviation	Dénomination complète
ANOVA	analyse de la variance (angl. analysis of variance)
AUC	aire sous la courbe ROC (angl. area under the curve)
BMI	voir IMC (angl. body mass index)
BPCO	bronchopneumopathie chronique obstructive
CI	voir IC (angl. confidence interval)
CONSORT	recommandations pour la publication des essais randomisés (angl. Consolidated Standards of Reporting Trials)
CRF	cahier d'observation (angl. case report form)
CV	coefficient de variation
ddl	degrés de liberté (angl. degrees of freedom, df)
DOR	rapport de cotes diagnostique (angl. diagnostic odds ratio)
EBM	médecine fondée sur les preuves (angl. evidence-based medicine)
ECR	essai contrôlé randomisé (angl. randomized controlled trial, RCT)
EIQ	écart interquartile (angl. interquartile range, IQR)
ES	erreur standard (angl. standard error, SE)
ÉT	écart-type (angl. standard deviation, SD)
FINDRISC	score finlandais de risque de diabète (angl. Finnish Diabetes Risk Score)
FN	faux négatif
FP	faux positif
GRADE	système d'évaluation de la certitude des preuves (angl. Grading of Recommendations Assessment, Development and Evaluation)
H0, H1	hypothèse nulle, hypothèse alternative
HbA1c	hémoglobine glyquée
HGPO	hyperglycémie provoquée par voie orale (angl. oral glucose tolerance test, OGTT)
HR	rapport des risques instantanés (angl. hazard ratio)
HTA	hypertension artérielle
IC	intervalle de confiance (angl. confidence interval)
ICC	coefficient de corrélation intraclasse (angl. intraclass correlation coefficient)
IMC	indice de masse corporelle
ITT	analyse en intention de traiter (angl. intention to treat)
NNH	nombre de patients à exposer pour provoquer un événement indésirable (angl. number needed to harm)
NST	nombre de sujets à traiter pour prévenir un événement (angl. number needed to treat, NNT)
OR	rapport de cotes (angl. odds ratio)
PAD	pression artérielle diastolique
PAS	pression artérielle systolique
PICO	population, intervention, comparateur, critère de jugement (angl. population, intervention, comparison, outcome)

Abréviation	Dénomination complète
PRISMA	recommandations pour la publication des revues systématiques
QRM	question à réponses multiples (type de question de QCM)
QRU	question à réponse unique (type de question de QCM)
RAR	réduction absolue du risque
RGPD	Règlement général sur la protection des données, règlement (UE) 2016/679 (angl. GDPR)
ROC	courbe caractéristique de fonctionnement du récepteur (angl. receiver operating characteristic)
RP	rapport de prévalences (angl. prevalence ratio)
RR	risque relatif (angl. relative risk)
RRR	réduction relative du risque
RV+, RV-	rapport de vraisemblance positif et négatif (angl. likelihood ratio, LR)
Se	sensibilité
SEE	erreur standard de l'estimation en régression (angl. standard error of the estimate)
Sp	spécificité
STARD	recommandations pour la publication des études d'exactitude diagnostique
STROBE	recommandations pour la publication des études observationnelles
VN	vrai négatif (angl. true negative, TN)
VP	vrai positif (angl. true positive, TP)
VPN	valeur prédictive négative
VPP	valeur prédictive positive

Partie I. Cours

Chapitre 1. Introduction à la biostatistique : concepts fondamentaux, rôle dans la recherche médicale et éléments de probabilités

Objectifs

À la fin de ce chapitre, l'étudiant sera capable de :

- expliquer pourquoi la variabilité biologique rend la statistique nécessaire en médecine et donner des exemples tirés de la pratique de cabinet ;
- définir l'unité statistique, la population, l'échantillon et leur effectif, et les identifier dans un article publié ;
- comparer le recensement, le dépistage et l'échantillonnage, et choisir une méthode d'échantillonnage adaptée à une question donnée ;
- reconnaître les principaux types de biais (de sélection, d'information, de confusion) dans un scénario clinique ;
- distinguer la statistique descriptive de la statistique inférentielle et utiliser les notations des paramètres (μ, σ, π) et des statistiques (m, s, p_i) ;
- décrire les étapes d'une étude statistique et formuler une question de recherche au format PICO ;
- énumérer les sources de données biomédicales ainsi que les principes éthiques et juridiques de leur traitement (RGPD, loi roumaine 190/2018, anonymisation, pseudonymisation, consentement, avis du comité d'éthique) ;
- calculer des probabilités simples, conditionnelles et totales à partir d'un tableau de contingence et vérifier l'indépendance de deux événements ;
- calculer des probabilités binomiales simples et reconnaître les situations dans lesquelles on utilise la loi de Poisson.

1.1 Pourquoi un médecin a besoin de la biostatistique

Le premier contact de nombreux étudiants avec la biostatistique est une question sincère : "À quoi cela va-t-il me servir, si je veux devenir chirurgien ?". La réponse ne vient pas des mathématiques, mais de la journée ordinaire d'un médecin. Le matin, le médecin de famille a sur son bureau les résultats de laboratoire des patients vus la semaine précédente. Une glycémie à jeun de 118 mg/dL chez un patient qui avait 104 mg/dL il y a un an est-elle une aggravation réelle ou une fluctuation sans importance ? À midi, il lit le résumé d'une étude montrant qu'un nouvel antidiabétique réduit le risque d'hospitalisation pour insuffisance cardiaque "de 30%". Le soir, il remplit un score de risque pour un patient et doit lui expliquer ce que signifie un résultat "élevé". Chacune de ces situations exige un jugement sur l'incertitude, et la biostatistique est la langue dans laquelle ce jugement s'exprime.

La médecine fondée sur les preuves a été définie comme l'utilisation consciencieuse, explicite et judicieuse des meilleures preuves disponibles dans les décisions concernant la prise en charge de chaque patient, intégrées à l'expérience clinique du médecin et aux valeurs du patient [15]. Le mot "preuves" cache presque toujours un calcul statistique : un intervalle de confiance, une valeur p, un risque relatif, une sensibilité. Les guides de pratique clinique que vous utiliserez toute votre vie classent leurs recommandations selon la qualité des études dont elles proviennent, et la qualité d'une étude se juge en grande partie sur son schéma et son analyse statistique. Un médecin qui ne

comprend pas ce que signifie "IC 95% 0,62-0,91" ne peut pas vérifier par lui-même si une recommandation s'appuie sur des données solides ou sur un résultat fragile.

Il existe aussi une raison très concrète et proche dans le temps : le concours de résidanat. Sa bibliographie contient des chapitres d'épidémiologie, d'évaluation des tests diagnostiques et d'interprétation des études cliniques. Les questions sur la sensibilité et la valeur prédictive, sur le risque relatif ou sur les types d'études reviennent d'année en année. L'examen de cette discipline est construit dans le même style, avec des questions à réponse unique et à réponses multiples, précisément pour vous habituer tôt à ce type de raisonnement.

Que pensent les étudiants en médecine de ces disciplines ? Un sondage anonyme mené auprès de 121 étudiants de notre université leur a demandé de noter plusieurs cours sur une échelle de 1 à 10 [16]. Le cours de médecine d'urgence a obtenu une médiane de 9 (EIQ 7-10), celui d'informatique biomédicale 7 (5-9) et la biostatistique 8 (5-9). Le résultat le plus intéressant est pourtant ailleurs : la corrélation entre les connaissances mesurées par un test et celles autoévaluées par les étudiants était pratiquement nulle ($R = 0,107$, $p = 0,246$). Autrement dit, les étudiants ne sont pas capables d'évaluer correctement leur niveau de connaissances dans ce domaine ; certains croient en savoir plus qu'ils n'en savent, d'autres se sous-estiment. Remarquez aussi comment les résultats ont été rapportés : médiane et écart interquartile, car une note sur une échelle de 1 à 10 est une variable ordinale, et non une mesure proprement dite (chapitre 2). Sur le coût et les possibilités de l'enseignement assisté par les technologies de l'information pendant les années de pandémie, les collègues de la discipline ont publié une analyse distincte [17] ; bon nombre des outils numériques utilisés dans les travaux pratiques de cet ouvrage se sont consolidés à cette époque.

Enfin et surtout, beaucoup d'entre vous feront de la recherche : un mémoire de fin d'études, un projet dans un cercle scientifique étudiant, une thèse de doctorat. Sans notions de biostatistique, la collaboration avec un statisticien devient un monologue. Le statisticien ne peut pas choisir seul le test approprié s'il ne comprend pas la question clinique, et le médecin ne peut pas formuler correctement la question s'il ne sait pas ce qu'une analyse peut offrir et ce qu'elle ne peut pas offrir. Le présent livre tente de vous préparer à ce dialogue, comme le font aussi les manuels de référence de la discipline [14,18,19].

1.1.1 La variabilité, matière première de la biologie

Deux patients de même âge, de même sexe et porteurs du même diagnostic n'ont jamais la même pression artérielle. Le même patient n'a pas non plus la même pression lors de deux mesures consécutives, faites à cinq minutes d'intervalle. Cette variabilité est la caractéristique fondamentale des phénomènes biologiques et elle a plusieurs sources. Il existe une variabilité entre les individus (génétique, âge, mode de vie, maladies associées), une variabilité dans le temps chez un même individu (rythme circadien, effort, stress, alimentation) et une variabilité de la mesure (l'appareil, la technique, l'observateur, les conditions de laboratoire).

Si toutes les valeurs étaient identiques, la statistique ne serait pas nécessaire : une seule mesure dirait tout sur l'ensemble de la population. C'est précisément parce que les valeurs diffèrent que nous avons besoin de méthodes qui décrivent dans quelle mesure elles diffèrent (la dispersion), autour de quelle valeur elles se regroupent (la tendance centrale) et à quel point les conclusions tirées d'un nombre limité d'observations sont sûres. Dans le jeu de données pédagogique mg2026, présenté à la section 1.3, la pression artérielle systolique à l'inclusion varie entre 102 et 169 mmHg, bien que les 160 patients aient tous été recrutés selon les mêmes critères. La moyenne est de 131,13 mmHg et l'écart-type de 13,88 mmHg ; nous apprendrons au chapitre 2 ce que signifient ces chiffres.

En travaux pratiques, les étudiants demandent souvent si la variabilité est "une erreur". La réponse est non. La variabilité biologique réelle est une information : elle nous dit à quel point les patients diffèrent les uns des autres. L'erreur de mesure s'y ajoute, et nous pouvons la réduire grâce à des appareils étalonnés, des protocoles standardisés et des mesures répétées. La première ne peut pas être éliminée, et elle ne doit pas l'être. La biostatistique l'accepte et apprend à travailler avec elle.

1.1.2 Statistique et biostatistique

La statistique est la science qui s'occupe du recueil, de l'organisation, de la description, de l'analyse et de l'interprétation des données relatives aux phénomènes de masse, c'est-à-dire aux phénomènes qui se manifestent chez un grand nombre d'individus et qui ne peuvent pas être compris à partir d'un seul cas. La biostatistique est l'application de ces méthodes aux données de la biologie, de la médecine et de la santé publique. La différence ne tient pas aux formules, qui sont les mêmes, mais au type de problèmes : études cliniques, tests diagnostiques, facteurs de risque, survie, scores de gravité, questionnaires de qualité de vie.

La biostatistique présente quelques particularités. Les données proviennent de personnes, avec leur consentement, et sont protégées par des règles éthiques et juridiques strictes. Les mesures ont souvent des unités cliniques (mmHg, mg/dL, kg/m²) et des seuils ayant une signification médicale (140 mmHg pour l'hypertension, 126 mg/dL pour le diabète). De nombreuses variables sont asymétriques (les triglycérides, la durée d'hospitalisation, les titres d'anticorps), et de nombreux résultats sont binaires (guéri ou non, décédé ou non). Surtout, les conclusions se transforment en décisions concernant des patients, si bien que les erreurs ont un coût réel.

1.2 Concepts de base : unité statistique, population, échantillon

Toute étude commence par une question simple : "De qui parlons-nous ?". La réponse précise à cette question définit quelques notions que vous utiliserez dans tout le reste du cours.

1.2.1 L'unité statistique et la population

L'unité statistique (l'individu statistique) est l'élément sur lequel portent les observations, situé dans l'espace et dans le temps. Dans la plupart des études médicales, l'unité statistique est le patient. Ce n'est toutefois pas toujours le cas. Dans une étude de prescription, l'unité peut être l'ordonnance ; dans une étude d'ophtalmologie, l'œil ; dans une étude d'odontologie, la dent ou la surface dentaire ; dans une étude sur les services de santé, le cabinet ou l'hôpital. Le choix de l'unité a des conséquences : les deux yeux d'un même patient ne sont pas des observations indépendantes, et l'analyse doit en tenir compte.

La population est l'ensemble de toutes les unités statistiques qui ont au moins une propriété commune et qui sont, elles aussi, situées dans l'espace et dans le temps. "Les adultes prédiabétiques du département de Timiș en 2026" constituent une population bien définie ; "les patients prédiabétiques" non, car on ne dit ni où ni quand. Le nombre d'unités de la population s'appelle l'effectif de la population et se note habituellement N.

En recherche, on distingue la population cible, à laquelle on veut généraliser les conclusions, et la population accessible (ou population source), dans laquelle on peut effectivement recruter. Pour une étude menée dans des cabinets de médecine générale de l'ouest du pays, la population cible peut être "les adultes à risque cardiométabolique en Roumanie", mais la population accessible est "les adultes inscrits sur les listes des 12 cabinets participants". L'écart entre les deux est l'une des premières questions que doit se poser le lecteur d'un article.

1.2.2 L'échantillon

L'échantillon est un sous-ensemble de la population, choisi pour être étudié à sa place. Le nombre d'unités de l'échantillon est l'effectif de l'échantillon et se note n. L'échantillon de 160 patients du jeu de données mg2026 provient d'une population accessible de quelques milliers d'adultes inscrits dans les cabinets participants ; les analyses portent sur les 160, mais la question d'intérêt concerne l'ensemble de la population.

L'effectif de l'échantillon ne se choisit ni "le plus grand possible" ni "autant qu'on peut". Il se calcule avant le début de l'étude, en fonction de la précision souhaitée ou de la différence que l'on veut détecter. Nous verrons au chapitre 4 que, pour estimer la moyenne de la pression systolique avec une marge de ±3 mmHg, lorsque l'écart-type est d'environ 14 mmHg, 84 patients suffisent, mais que pour estimer une proportion de 20% avec une marge de ±5%, il en faut 246. La différence tient à la nature de la variable et à la précision demandée, et non au souhait du chercheur.

1.2.3 Recensement, dépistage, échantillonnage

Il existe trois façons principales de recueillir des données sur une population. Le recensement observe toutes les unités de la population à un moment donné. Il est exact (il n'y a pas d'erreur d'échantillonnage), bien situé dans le temps et dans l'espace, mais très coûteux, lent et rarement possible en médecine. Le recensement de la population de la Roumanie a lieu tous les dix ans ; un registre national exhaustif d'une maladie rare se rapproche d'un recensement des cas.

Le dépistage examine systématiquement une population (ou une large partie de celle-ci) afin d'identifier les personnes atteintes d'une affection donnée, souvent encore non diagnostiquée. Les exemples classiques sont le dépistage du cancer du col de l'utérus, du cancer colorectal, de la surdité chez le nouveau-né ou du diabète de type 2 à l'aide du score FINDRISC au cabinet de médecine générale. Le dépistage n'exige pas une localisation stricte dans le temps (il peut s'étendre sur plusieurs années), a un coût relativement élevé et vise un bénéfice individuel, et pas seulement une estimation statistique.

L'échantillonnage (la sélection) n'étudie qu'une partie de la population et étend les conclusions à l'ensemble. Le coût est réduit, les données peuvent être recueillies plus soigneusement et le résultat est disponible plus rapidement. Le prix à payer est l'incertitude : le résultat obtenu sur l'échantillon diffère inévitablement de la vraie valeur dans la population. La statistique inférentielle existe précisément pour quantifier cette incertitude. Presque toutes les études que vous lirez sont des études sur échantillons.

1.3 Le jeu de données pédagogique mg2026

Tout au long du cours, nous travaillerons sur le même jeu de données, afin que vous puissiez suivre comment les mêmes variables sont décrites, estimées, comparées et corrélées. Ce jeu s'appelle mg2026 et il est disponible sur Moodle en deux versions : mg2026_curat.csv (160 lignes, prêt pour l'analyse) et mg2026_brut.csv (161 lignes, avec des erreurs laissées volontairement, que vous nettoierez lors de la séance de travaux pratiques 1).

Le scénario est celui d'un essai randomisé pragmatique simulé, mené dans 12 cabinets de médecine générale de l'ouest du pays. Ont été inclus 160 adultes à risque cardiométabolique : des personnes en surpoids, prédiabétiques ou atteintes d'un diabète non diagnostiqué auparavant. Les patients ont été répartis aléatoirement entre trois bras : A, conseils standard sur le mode de vie (53 patients) ; B, conseils et application mobile d'automesure (53) ; C, conseils, application et téléconsultation mensuelle avec le médecin de famille (54). Les mesures ont été faites à l'inclusion, à 3 mois et à 6 mois. Le diabète a été confirmé à l'inclusion par l'hyperglycémie

provoquée par voie orale (HGPO), qui est le standard de référence par rapport auquel on évalue le score FINDRISC. Les patients sont fictifs et les valeurs ont été générées par simulation, de manière à ressembler aux données réelles du cabinet sans provenir d'aucune personne.

Le tableau 1.1 résume les variables du jeu de données. Le dictionnaire complet se trouve à l'annexe C et sera réutilisé au chapitre 2, lorsque nous aborderons les types de variables.

Tableau 1.1. Les variables du jeu de données pédagogique mg2026.

Variable	Signification	Type (voir le chapitre 2)
id	code du patient (P001-P160)	identifiant
brat	bras de traitement (A, B, C)	nominale
sex	F, M	nominale dichotomique
varsta	âge à l'inclusion, en années	numérique
mediu	urbain, rural	nominale dichotomique
educatie	1 collège, 2 lycée ou école professionnelle, 3 études universitaires, 4 études postuniversitaires	ordinale
fumat	non-fumeur, ancien fumeur, fumeur	nominale
activitate	activité physique autoévaluée, 1-5	ordinale
inaltime, greutate0, imc0, talie	taille (cm), poids (kg), IMC (kg/m²), tour de taille (cm)	numériques
tas0, tad0, fc0	pression systolique, diastolique (mmHg), fréquence cardiaque (battements/min)	numériques
colesterol, hdl, trigliceride	bilan lipidique (mg/dL)	numériques
glicemie0, glicemie3, glicemie6	glycémie à jeun à 0, 3 et 6 mois (mg/dL)	numériques
hba1c0	hémoglobine glyquée (%)	numérique
findrisc	score de risque de diabète (0-26)	numérique discrète (score)
diabet_ogtt	diabète confirmé par HGPO (0/1)	nominale dichotomique
greutate6, tas6	poids et PAS à 6 mois	numériques
scadere5	perte de poids ≥ 5% à 6 mois (0/1)	nominale dichotomique
satisfactie	satisfaction à l'égard du programme, 1-5 (Likert)	ordinale

Quelques chiffres vous aideront à vous orienter dès le départ. Sur les 160 patients, 86 sont des femmes et 74 des hommes ; 95 vivent en milieu urbain et 65 en milieu rural ; 74 sont non-fumeurs, 33 anciens fumeurs et 53 fumeurs. L'âge moyen est de 56,29 ans, avec des valeurs extrêmes de 35 et 79 ans. Le diabète a été confirmé par HGPO chez 29 patients, soit 18,1%. Ces chiffres reviendront, sous différentes formes, dans presque chaque chapitre.

Un détail de schéma d'étude mérite d'être retenu dès maintenant : à l'inclusion, les données mg2026 peuvent être analysées comme une étude transversale (par exemple, l'association entre obésité et diabète), alors que le suivi à 3 et 6 mois dans des bras randomisés en fait un essai clinique contrôlé. Le même jeu de données peut donc répondre à des questions de nature différente, à condition de savoir quel type de conclusion chacune de ses parties autorise.

1.4 L'échantillonnage : représentativité et méthodes

Un échantillon n'est utile que s'il ressemble à la population dont il provient. On dit qu'un échantillon est représentatif lorsqu'il reproduit, dans des proportions voisines, la structure de la population pour les caractéristiques qui comptent dans l'étude : âge, sexe, milieu d'origine, gravité de la maladie, etc. La représentativité ne peut pas être garantie par l'intuition. Elle s'obtient par la manière dont les individus sont choisis.

1.4.1 Les conditions d'un échantillon représentatif

Deux conditions fondent une sélection correcte. La première est l'équiprobabilité : chaque individu de la population doit avoir la même probabilité d'être inclus dans l'échantillon (ou, dans les variantes plus complexes, une probabilité connue et non nulle). La seconde est l'indépendance : l'inclusion d'un individu ne doit pas influencer celle d'un autre. Si un médecin invite à l'étude, en plus du patient, tous les membres de sa famille, l'indépendance est perdue, car les membres d'une même famille se ressemblent par leur alimentation, leurs habitudes et leurs gènes.

Toute condition qui donne à certains individus plus de chances d'être sélectionnés déforme l'échantillon. Le résultat est un biais de sélection, dont nous parlerons à la section 1.5. Une règle pratique : si vous ne pouvez pas décrire exactement comment chaque patient est arrivé dans l'échantillon, vous ne pouvez pas affirmer que l'échantillon est représentatif.

1.4.2 L'échantillonnage aléatoire simple

L'échantillonnage aléatoire simple est le modèle théorique dont partent presque toutes les formules de la statistique. On part d'une liste complète de la population (la base de sondage), chaque individu reçoit un numéro, et les numéros inclus sont choisis à l'aide d'un générateur de nombres aléatoires. Dans Google Sheets, par exemple, on peut ajouter une colonne contenant =RAND() à côté de la liste des patients, trier selon cette colonne et conserver les n premières lignes.

Supposons qu'un cabinet compte sur sa liste 1 850 adultes de 40 à 75 ans et souhaite en inviter 120 à une évaluation FINDRISC. Avec un échantillonnage aléatoire simple, chaque patient a la même probabilité d'être invité, 120/1 850 = 0,065 (6,5%). La méthode est correcte, mais elle présente deux points faibles en pratique : elle exige une liste complète et à jour, et elle peut produire, par hasard, un échantillon déséquilibré (par exemple, avec trop peu d'hommes), surtout lorsque n est petit.

1.4.3 L'échantillonnage systématique

Dans l'échantillonnage systématique, on calcule un pas $k = \frac{N}{n}$, on choisit au hasard un point de départ entre 1 et k, puis on prend chaque k-ième individu de la liste. Pour 1 200 patients et un échantillon de 80, le pas est k = 15. Si le numéro de départ tiré est 7, on inclut les patients portant les numéros 7, 22, 37, 52, et ainsi de suite jusqu'à 1 192.

La méthode est simple et s'accorde très bien avec le flux du cabinet : "un patient sur cinq parmi ceux qui se présentent en consultation au mois de mars". Le danger apparaît lorsque la liste présente une périodicité qui coïncide avec le pas. Si les rendez-vous sont organisés de telle sorte que chaque cinquième rendez-vous de la journée est réservé aux consultations urgentes, un pas de 5 peut ne sélectionner que ce type de patients. Avant d'utiliser la méthode, vérifiez comment la liste est ordonnée.

1.4.4 L'échantillonnage stratifié

La population est divisée en strates homogènes selon une caractéristique importante (sexe, classe d'âge, milieu urbain ou rural), puis on tire un échantillon aléatoire dans chaque strate. Dans la variante proportionnelle, chaque strate contribue par un nombre d'individus proportionnel à sa taille dans la population. Si, sur 1 200 patients, 720 vivent en milieu urbain (60%) et 480 en milieu rural (40%), un échantillon stratifié de 80 comprendra 48 patients urbains et 32 ruraux.

La stratification augmente la précision des estimations lorsque la caractéristique de stratification influe sur la variable étudiée, et elle garantit que les petits groupes ne manquent pas dans l'échantillon. On utilise parfois délibérément une stratification non proportionnelle, avec une

surreprésentation d'un groupe rare (par exemple, les patients de plus de 75 ans), suivie d'une pondération des résultats lors de l'analyse.

1.4.5 L'échantillonnage en grappes

Lorsqu'il n'existe pas de liste de tous les individus, mais qu'il existe une liste de groupes naturels (cabinets, écoles, villages, services hospitaliers), on choisit au hasard quelques groupes et on étudie tous les individus des groupes retenus, ou un échantillon dans chacun d'eux (échantillonnage à deux degrés). Pour une étude nationale sur le contrôle de la pression artérielle, il est beaucoup plus facile de choisir au hasard 40 cabinets de médecine générale et d'inclure 25 patients dans chacun que de choisir 1 000 patients dispersés dans tout le pays.

L'avantage logistique a un coût statistique. Les patients d'un même cabinet se ressemblent (même médecin, même zone, mêmes habitudes de prescription), de sorte que 1 000 patients issus de 40 cabinets apportent moins d'information que 1 000 patients choisis individuellement. L'analyse doit tenir compte de ce regroupement. Le jeu de données mg2026 a une structure semblable, avec 12 cabinets participants, même si, dans ce cours, nous l'analyserons par souci de simplicité comme un échantillon d'individus.

1.4.6 L'échantillonnage de convenance

L'échantillon de convenance comprend les individus les plus faciles à recruter : les patients qui viennent en consultation pendant une période donnée, les volontaires qui répondent à une annonce, les personnes qui remplissent un questionnaire en ligne diffusé sur les réseaux sociaux. C'est de loin la méthode la plus utilisée en recherche clinique, car elle est rapide et peu coûteuse. C'est aussi la plus exposée aux biais de sélection, car rien ne garantit que les personnes recrutées ressemblent à la population cible.

Un échantillon de convenance n'est pas automatiquement inutile. Les essais cliniques randomisés utilisent presque toujours des patients recrutés consécutivement, donc un échantillon de convenance, et la validité de la comparaison entre les bras vient de la randomisation, non de la représentativité. Le problème surgit lorsque le résultat obtenu sur un échantillon de convenance (par exemple, une prévalence) est présenté comme valable pour l'ensemble de la population. Un médecin qui lit une telle étude doit se demander qui a répondu, qui n'a pas répondu, et pourquoi.

Tableau 1.2. Les méthodes d'échantillonnage en bref, avec des exemples tirés de la médecine générale.

Méthode	Principe	Exemple	Avantage	Limite
Aléatoire simple	chaque individu a la même chance	120 patients tirés avec RAND() dans la liste du cabinet	base théorique des formules	exige une liste complète ; déséquilibres dus au hasard
Systématique	chaque k-ième individu, départ aléatoire	un patient sur 15 de la liste, départ au 7e	simple, facile à appliquer au triage	périodicité de la liste
Stratifiée	échantillon aléatoire dans chaque strate	48 urbains + 32 ruraux, proportionnel	précision accrue, petits groupes présents	exige des informations sur les strates
En grappes	on choisit des groupes au hasard	40 cabinets, 25 patients chacun	logistique facile	observations corrélées au sein de la grappe
De convenance	les personnes disponibles	les patients vus en consultation en mars	rapide, peu coûteuse	risque élevé de biais de sélection

Étude de cas : la même cohorte d'étudiants, deux questions différentes

> Deux études publiées en 2026 par des collègues de la discipline ont utilisé le même groupe de 921 étudiants âgés de 18 à 24 ans, en première année d'études. Dans l'étude transversale sur les comportements alimentaires orientés vers la commodité, 24,1% des participants ont été classés en surpoids ou obèses, dont 19,4% en surpoids et 4,7% obèses [4]. Dans l'étude prospective sur la chrononutrition, ces mêmes étudiants ont été suivis pendant la première année universitaire, et la variation moyenne de l'IMC a été de -0,45 ± 1,07 kg/m² (IC 95% -0,52 à -0,38 ; $p < 0,001$) ; 48,2% ont vu leur IMC diminuer, 38,7% sont restés stables et 13,1% ont vu leur IMC augmenter, tandis qu'une prise de poids cliniquement pertinente est survenue chez 7,2% de la cohorte [5]. Une troisième étude, portant sur 693 étudiants du même âge, a rapporté une prévalence du prédiabète de 21,1% (diabète : 1,4%) [6]. La question que doit se poser le lecteur n'est pas de savoir si les chiffres sont exacts, mais à quelle population ils peuvent être étendus. Les étudiants d'une université de médecine et de pharmacie d'un seul centre ne représentent pas l'ensemble des jeunes de 18-24 ans de Roumanie : ils ont un niveau d'instruction plus élevé, un meilleur accès à l'information médicale et un mode de vie particulier. Les conclusions sont valides pour la population source et, avec prudence, pour les étudiants d'universités semblables. Ce qu'il faudrait rapporter en plus pour juger de la représentativité : combien d'étudiants ont été invités, combien ont accepté (le taux de participation) et comment les participants se comparent à ceux qui ont refusé, au moins selon le sexe et la faculté.

1.5 Les biais (erreurs systématiques)

Tout résultat obtenu sur un échantillon diffère de la vraie valeur dans la population. Cette différence a deux composantes. L'erreur aléatoire (d'échantillonnage) apparaît parce que nous n'avons étudié qu'une partie de la population ; elle diminue quand n augmente et peut être quantifiée par des intervalles de confiance. L'erreur systématique (le biais, bias en anglais) déplace toujours le résultat dans le même sens et ne disparaît pas quand on augmente l'échantillon. Une balance qui affiche 2 kg de trop donnera un poids moyen faux de 2 kg, que l'on pèse 10 ou 10 000 patients.

Cette distinction est fondamentale pour la lecture critique. Une grande étude a une erreur aléatoire faible, c'est-à-dire des intervalles de confiance étroits, mais elle peut avoir une erreur systématique importante. Précision ne signifie pas exactitude. Les biais sont habituellement classés en trois grandes familles : les biais de sélection, les biais d'information (de mesure) et la confusion.

1.5.1 Le biais de sélection

Le biais de sélection survient lorsque la manière dont les participants entrent dans l'étude (ou y restent) dépend à la fois de l'exposition et du résultat, de sorte que l'échantillon ne reflète plus la population. Quelques formes fréquentes :

L'effet "volontaire en bonne santé" : les personnes qui se proposent de participer à un programme de prévention sont, en moyenne, plus soucieuses de leur santé, fument moins et font plus d'exercice que la population générale. Un programme de dépistage évalué uniquement sur des volontaires paraîtra plus efficace qu'il ne l'est en réalité.

La non-réponse : si seulement 40% des personnes invitées remplissent un questionnaire sur la consommation d'alcool, il est probable que les gros consommateurs soient sous-représentés parmi les répondants. La prévalence estimée sera trop faible.

L'autosélection dans les sondages en ligne : un questionnaire diffusé sur un réseau social est rempli par ceux qui utilisent ce réseau, s'intéressent au sujet et ont du temps. L'âge, le sexe et le niveau d'instruction des répondants peuvent être très différents de ceux de la population.

Deux études récentes illustrent bien le problème. Un sondage en ligne de convenance sur les comportements alimentaires a inclus 753 jeunes de 18 à 30 ans, dont 93,5% de femmes ; 11,6% étaient en insuffisance pondérale et 29,3% en surpoids ou obèses [20]. Les auteurs ont correctement décrit le schéma d'étude dès le titre et ont rapporté les prévalences avec des intervalles de confiance de Wilson (une méthode que nous apprendrons au chapitre 4), mais une proportion de 93,5% de femmes montre que les résultats concernent en fait les jeunes femmes qui ont choisi de répondre. Un autre sondage en ligne, auprès de personnes diabétiques, a obtenu 329 réponses validées, et 44,07% des répondants ont déclaré utiliser des compléments à base de plantes [21]. Il est fort possible que les patients intéressés par les compléments aient été plus disposés à remplir un questionnaire sur ce sujet, ce qui gonflerait la prévalence. Par contraste, une étude transversale portant sur 2 976 étudiants du département de Timiș, avec 62,5% de filles et 37,5% de garçons âgés de 18 à 25 ans, disposait d'un grand échantillon, mais même un effectif important ne garantit pas l'absence de biais de sélection : la question reste de savoir qui a été invité et qui a répondu [22].

Les perdus de vue (dans les études prospectives) : si les patients qui ne répondent pas au traitement abandonnent l'étude plus souvent que les autres, ceux qui restent à la fin montrent un effet meilleur que l'effet réel.

Le biais de spectre (dans les études de tests diagnostiques) : un test évalué chez des patients atteints d'une maladie avancée, comparés à des volontaires en parfaite santé, paraît bien meilleur qu'il ne le sera au cabinet, où il doit distinguer des formes légères de la maladie d'autres affections semblables. Nous reviendrons sur ce biais au chapitre 12.

1.5.2 Le biais d'information (de mesure)

Le biais d'information survient lorsque les variables sont mesurées ou classées de manière erronée et que l'erreur diffère entre les groupes. Le biais de mémorisation (recall bias) en est l'exemple classique : les mères d'enfants nés avec une malformation se souviennent beaucoup plus en détail des médicaments pris pendant la grossesse que les mères d'enfants en bonne santé, non parce qu'elles en ont pris davantage, mais parce qu'elles ont davantage réfléchi aux causes possibles. Dans une étude cas-témoins, cette différence peut créer une association qui n'existe pas.

D'autres formes sont le biais de l'observateur (le médecin qui sait quel traitement a reçu le patient tend, involontairement, à juger son évolution plus favorablement), le biais instrumental (un tensiomètre non étalonné), le biais de déclaration (les patients sous-déclarent le tabagisme ou la consommation d'alcool face au médecin) et l'effet de "préférence pour le chiffre zéro", très fréquent lors de la mesure de la pression artérielle : une proportion anormalement élevée de valeurs se terminant par 0 (130, 140, 150) montre que les mesures ont été arrondies. L'insu (le masquage) des évaluateurs et les protocoles de mesure standardisés sont les principaux moyens de prévention.

1.5.3 La confusion

La confusion (angl. confounding) survient lorsque l'association entre une exposition et un résultat est faussée par une troisième variable, liée à la fois à l'exposition et au résultat, sans être une étape intermédiaire entre elles. L'exemple de manuel : dans certaines études anciennes, les buveurs de café avaient plus souvent un cancer du poumon. Le café ne provoque pas de cancer du poumon ; les fumeurs boivent davantage de café, et le tabagisme provoque le cancer du poumon. Le tabagisme est le facteur de confusion.

En médecine générale, l'âge est le facteur de confusion le plus fréquent. Les patients qui reçoivent davantage de médicaments sont plus souvent hospitalisés, mais ils sont aussi plus âgés et ont davantage de maladies. Une association brute "plus de médicaments, plus

d'hospitalisations" ne dit rien de l'effet des médicaments tant qu'on ne tient pas compte de l'âge et des comorbidités. La confusion se prévient par le schéma d'étude (randomisation, restriction, appariement) et se contrôle lors de l'analyse par stratification ou par des modèles multivariables, que nous rencontrerons aux chapitres 9 et 10.

Une remarque qui revient souvent à l'examen : la randomisation est la seule méthode qui équilibre entre les groupes à la fois les facteurs de confusion connus et inconnus. C'est pourquoi les essais randomisés se situent au sommet de la hiérarchie des preuves pour les questions portant sur l'effet des traitements.

> **Étude de cas : quand une association disparaît après ajustement**
>
> Dans une étude prospective portant sur 278 patients hospitalisés pour insuffisance cardiaque décompensée et suivis pendant 12 mois, on a analysé le rapport entre l'hémoglobine et l'indice de distribution des globules rouges (HRR) comme marqueur pronostique [23]. La médiane du HRR était de 0,89 et l'événement principal est survenu chez 167 patients (60,1%). Les courbes de survie ont montré une différence entre les patients à HRR bas et ceux à HRR élevé (log-rank p = 0,027), mais après ajustement sur l'âge, le sexe, le NT-proBNP, la fraction d'éjection et le débit de filtration glomérulaire, l'association n'était plus statistiquement significative (p = 0,240). L'interprétation correcte est qu'une part importante de l'association brute s'expliquait par les variables d'ajustement : les patients à HRR bas étaient probablement plus âgés, avec une fonction rénale et cardiaque moins bonne. Le HRR reste un marqueur facile à obtenir, mais son information pronostique recoupe en grande partie celle des marqueurs classiques. Les auteurs ont procédé correctement en rapportant les deux analyses ; le lecteur doit retenir qu'une association brute n'est pas la preuve d'un effet indépendant. Ce qui aiderait en plus : le rapport des risques instantanés (HR) avec son intervalle de confiance pour les deux modèles, afin de voir de combien l'effet a diminué, et pas seulement s'il a franchi le seuil de 0,05.

1.6 Statistique descriptive et statistique inférentielle

La biostatistique comporte deux grandes branches, qui correspondent à deux questions différentes. La statistique descriptive répond à la question "Qu'avons-nous observé ?" : elle organise les données en tableaux de fréquences, les résume par des indicateurs (moyenne, médiane, écart-type, proportions) et les présente sous forme graphique. Elle ne concerne que l'échantillon étudié et ne fait aucune affirmation sur la population. Le chapitre 2 lui est entièrement consacré.

La statistique inférentielle répond à la question "Que pouvons-nous dire de la population à partir de ce que nous avons observé ?". Elle utilise la théorie des probabilités pour estimer les paramètres de la population (par des intervalles de confiance) et pour tester des hypothèses (par des tests statistiques). La figure 1.1 présente schématiquement ce cheminement : on extrait un échantillon de la population, on calcule des statistiques sur l'échantillon, puis on tire de ces statistiques des conclusions, assorties d'une certaine probabilité d'erreur, sur les paramètres de la population.

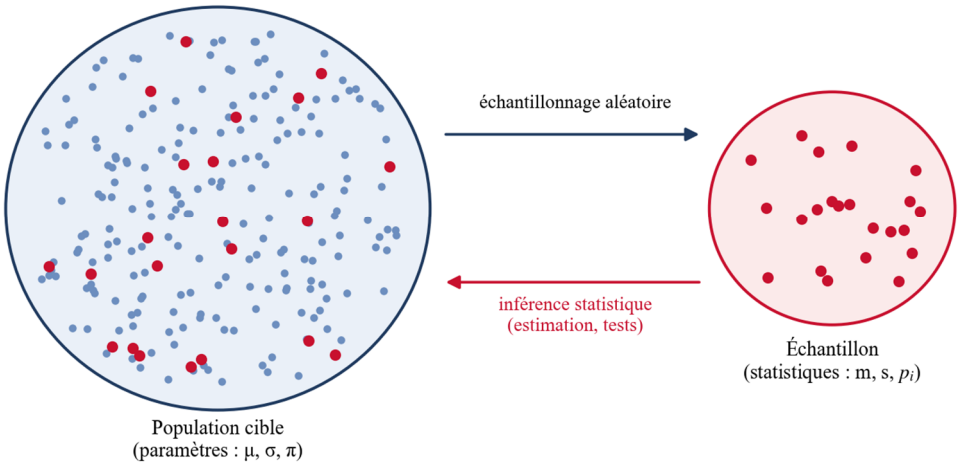


Figure 1.1. La population, l'échantillon et l'inférence statistique. De la population cible (dont les paramètres μ, σ, π sont inconnus) on extrait un échantillon sur lequel on calcule des statistiques (m, s, p_i) ; l'inférence va en sens inverse, de l'échantillon vers la population, et s'accompagne toujours d'une mesure de l'incertitude.

1.6.1 Paramètres et statistiques : les notations

Une source constante de confusion à l'examen est le mélange des notations. Les caractéristiques de la population s'appellent des paramètres et se notent, par convention, par des lettres grecques : la moyenne de la population μ (mu), l'écart-type de la population σ (sigma), la proportion dans la population π (pi, qui n'a ici aucun rapport avec 3,14). Les paramètres sont fixes, mais généralement inconnus.

Les caractéristiques calculées sur l'échantillon s'appellent des statistiques (ou estimateurs) et se notent par des lettres latines : la moyenne de l'échantillon m (dans certains ouvrages $\bar{x}$), l'écart-type de l'échantillon s, la proportion dans l'échantillon p_i (p avec l'indice inférieur i), c'est-à-dire la proportion des sujets de la classe i de la variable parmi l'ensemble des sujets étudiés : $p_i = \frac{n_i}{n}$, où n_i est le nombre de sujets de la classe i. Les statistiques peuvent être calculées exactement, mais elles varient d'un échantillon à l'autre. Le tableau 1.3 résume cette convention, avec les valeurs du jeu de données mg2026.

Tableau 1.3. Paramètres et statistiques : notations et exemples sur le jeu de données mg2026.

Caractéristique	**Paramètre (population)**	**Statistique (échantillon)**	**Valeur dans mg2026 (n = 160)**
moyenne de la PAS à l'inclusion	μ (inconnue)	m	131,13 mmHg
écart-type de la PAS	σ (inconnu)	s	13,88 mmHg
proportion des patients dont le diabète est confirmé	π (inconnue)	p_i	$\frac{29}{160} = 0{,}181 (18{,}1\%)$
effectif	N (des milliers de patients)	n	160

La valeur $p_i = 0{,}181$, où la classe i est "diabète confirmé", est une estimation ponctuelle de π. Un autre échantillon de 160 patients issus des mêmes cabinets aurait presque certainement donné une proportion un peu différente, peut-être 15% ou 21%. Nous verrons au chapitre 4 que l'intervalle de confiance à 95% de π, calculé par la méthode de Wilson, est de 12,9-24,8%. Cet intervalle dit quelque chose de la population ; le chiffre de 18,1% dit quelque chose de l'échantillon.

Attention à un piège de vocabulaire : c'est pour cette raison que la proportion dans l'échantillon s'écrit avec un indice, p_i, et non avec la seule lettre p. Sans indice, la lettre p désigne,

dans les chapitres consacrés aux tests statistiques, la "valeur p", c'est-à-dire une probabilité calculée par le test, et ces deux notions ne doivent pas être confondues.

1.7 Les étapes d'une étude statistique et la question de recherche

Une étude ne commence ni par les données ni par le logiciel de statistique. Elle commence par une question. La figure 1.2 montre l'enchaînement des étapes, de la question clinique à la décision ; chaque étape dépend de la qualité de la précédente.

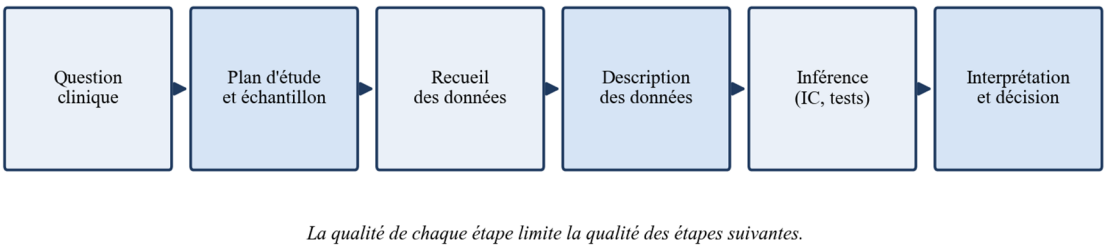


Figure 1.2. Les étapes d'une étude statistique : la question clinique, le schéma d'étude et l'échantillon, le recueil des données, la description des données, l'inférence (intervalles de confiance et tests) et l'interprétation avec la décision clinique. Les erreurs commises aux premières étapes ne peuvent plus être corrigées par l'analyse.

1.7.1 La question de recherche et le format PICO

Une bonne question de recherche est précise, pertinente sur le plan clinique et vérifiable à l'aide de données. Le format PICO aide à la structurer en quatre éléments : P (la population ou les patients), I (l'intervention ou l'exposition), C (la comparaison) et O (le critère de jugement, angl. outcome) [24]. On ajoute parfois T (le temps) et S (le type d'étude).

Pour le scénario mg2026, une question formulée de façon vague serait : "Les applications mobiles aident-elles les patients prédiabétiques ?". Au format PICO, elle devient : "Chez les adultes à risque cardiométabolique suivis en cabinet de médecine générale (P), l'ajout d'une application d'automesure et d'une téléconsultation mensuelle aux conseils standard (I), par rapport aux conseils standard seuls (C), augmente-t-il la proportion de patients qui perdent au moins 5% de leur poids en 6 mois (O) ?". La seconde formulation dit exactement quelles données il faut recueillir, quels groupes sont comparés et quelle variable sera analysée. La réponse, que nous calculerons au chapitre 10, est que cette proportion a été de 29,6% dans le bras C contre 11,3% dans le bras A.

Le format PICO n'est pas une formalité. Il oblige le chercheur à choisir un seul critère de jugement principal avant de voir les données, ce qui prévient l'un des problèmes les plus fréquents de la littérature : tester des dizaines de variables et ne rapporter que celles qui sont "significatives".

1.7.2 De la question au schéma d'étude et au recueil des données

Une fois la question formulée, on choisit le schéma d'étude (transversal, cas-témoins, cohorte, essai clinique randomisé ; voir les chapitres 10 et 13), on définit la population, on fixe les critères d'inclusion et d'exclusion, on calcule l'effectif de l'échantillon et on construit l'instrument de recueil : la fiche d'observation, le formulaire électronique ou le questionnaire. Chaque variable

reçoit une définition opérationnelle, une unité de mesure et un mode de codage. Ce n'est qu'ensuite que commence le recrutement.

L'analyse statistique se planifie elle aussi avant le recueil, dans un plan d'analyse statistique. Celui-ci précise le critère de jugement principal, les tests qui seront utilisés, le seuil de signification et la manière de traiter les données manquantes. L'analyse "après avoir vu les données" n'est pas interdite, mais elle doit être présentée comme une exploration, et non comme la vérification d'une hypothèse. Les recommandations pour la publication, comme STROBE pour les études observationnelles, exigent que ces éléments soient décrits explicitement dans l'article [25].

> **Étude de cas : les étapes d'une étude par questionnaire, de la traduction au rapport**
>
> Un questionnaire utilisé dans une langue différente de celle dans laquelle il a été créé ne peut pas être simplement traduit par un seul locuteur bilingue. Le protocole ALSATION, publié par une équipe dont fait partie l'auteur de ce cours, décrit le plan de traduction en roumain de trois questionnaires de littératie en santé [26]. Le fait que la traduction d'un instrument ait justifié un protocole publié séparément en dit long : l'étape de "construction de l'instrument" exige la même rigueur que l'analyse. En général, l'adaptation d'un questionnaire suppose des traductions indépendantes, leur réconciliation, une rétrotraduction, une évaluation par des experts et un essai auprès d'un petit groupe de répondants, et chacune de ces étapes peut modifier le sens d'une question. Par ailleurs, une étude par questionnaire menée dans 10 pharmacies en milieu urbain et rural a inclus 189 personnes âgées de 40 ans et plus, dont 78,8% de femmes et 69,3% de citadins ; 75,1% ont déclaré savoir ce qu'est l'ostéoporose, mais 77,3% n'avaient jamais passé d'ostéodensitométrie DXA, et 42,9% de ceux qui ont déclaré un diagnostic d'ostéoporose ne suivaient aucun traitement [27]. Les deux études montrent que le résultat d'un questionnaire dépend d'étapes qui n'apparaissent pas dans le tableau final : comment les questions ont été formulées, où elles ont été posées, qui a accepté de répondre. Dans l'étude menée en pharmacie, la prédominance des femmes (78,8%) et des répondants urbains (69,3%) reflète en partie qui entre dans une pharmacie et accepte un questionnaire ; les prévalences doivent donc être lues comme valables pour ce type de répondants, et non pour l'ensemble des Roumains de plus de 40 ans.

1.8 Les données à l'ère du dossier électronique : le lien avec l'informatique médicale

Au premier semestre, en informatique médicale, vous avez appris comment les données de santé sont représentées, stockées et transmises : le dossier électronique du patient, les standards d'interopérabilité, les classifications et les nomenclatures, la sécurité de l'information [1]. La biostatistique commence exactement là où s'arrêtait ce cours. Une analyse statistique ne peut pas être meilleure que les données qu'elle traite, et la plupart des problèmes que vous rencontrerez en pratique ne tiennent pas au choix du test, mais à la qualité, à la structure et à la licéité des données.

1.8.1 D'où viennent les données biomédicales

Les sources de données peuvent être regroupées en quelques grandes catégories, chacune avec ses avantages et ses limites [28].

Le dossier électronique de santé et les systèmes informatiques des cabinets et des hôpitaux (angl. electronic health record, EHR) contiennent des diagnostics, des prescriptions, des résultats de laboratoire et des consultations, enregistrés au cours des soins courants [29]. L'avantage est le volume immense et le faible coût de données qui existent déjà. La limite est qu'elles ont été recueillies pour les soins et le remboursement, et non pour la recherche : les diagnostics peuvent être codés pour justifier une prescription, certaines valeurs manquent et les définitions diffèrent

d'un établissement à l'autre. Les diagnostics sont habituellement codés selon la Classification internationale des maladies [30], et les analyses de laboratoire peuvent être identifiées par des codes standard de type LOINC [31].

Les registres de maladies (registres du cancer, du diabète, de la dialyse, des maladies rares) et les données de surveillance épidémiologique transmises aux directions de santé publique suivent systématiquement tous les cas d'une zone donnée. Ils sont précieux pour estimer l'incidence et les tendances dans le temps, mais dépendent de l'exhaustivité des déclarations.

Les questionnaires et les formulaires de recherche (sur papier ou électroniques, par exemple dans Google Forms) recueillent exactement les variables dont l'étude a besoin, avec des définitions contrôlées. Le coût est plus élevé, et la qualité dépend de la formulation des questions, de la validation de l'instrument et du taux de réponse.

Les dispositifs portables et les applications mobiles (montres connectées, capteurs de glucose à mesure continue, tensiomètres connectés, applications d'automesure comme celle du bras B de l'étude mg2026) produisent de longues séries de mesures, à une fréquence impossible au cabinet. Leurs problèmes sont l'étalonnage, les données manquantes lorsque le patient ne porte pas le dispositif et le fait que les utilisateurs de ces technologies sont, en règle générale, plus jeunes et plus motivés que le patient moyen.

Les données de laboratoire et d'imagerie proviennent d'appareils et de systèmes informatiques qui respectent leurs propres standards ; elles sont généralement précises, mais les valeurs de référence et les méthodes peuvent différer d'un laboratoire à l'autre, ce qui compte lorsqu'on combine des données de plusieurs centres.

1.8.2 La qualité des données et le dictionnaire des données

La qualité des données a plusieurs dimensions : l'exhaustivité (combien de valeurs manquent), l'exactitude (les valeurs reflètent la réalité), la cohérence (la même information est codée de la même façon partout), la plausibilité (les valeurs sont biologiquement possibles) et l'actualité. Dans le fichier mg2026_brut.csv, vous trouverez des erreurs typiques : le sexe écrit tantôt "F", tantôt "f" ou "Feminin" ; une pression systolique de 1540 mmHg (un chiffre en trop) ; une taille de 1,72 saisie en mètres dans une colonne en centimètres ; un poids de 7,26 kg (virgule décimale déplacée) ; une glycémie écrite "87 mg/dl", donc du texte au lieu d'un nombre ; une ligne en double. Chacune d'elles fausserait une moyenne, une fréquence ou un test si elle n'était pas repérée à temps.

L'outil de base de la qualité des données est le dictionnaire des données (angl. data dictionary). Pour chaque variable, il précise le nom (court, sans espaces ni signes diacritiques, par exemple tas0), la description, le type (numérique, texte, date), l'unité de mesure, les valeurs autorisées ou l'intervalle plausible (par exemple, PAS entre 70 et 260 mmHg), les codes des variables catégorielles (1 = collège, 2 = lycée...) et le code utilisé pour les valeurs manquantes. Le dictionnaire s'écrit avant le recueil, et non après ; il devient ensuite la base des règles de validation du formulaire électronique et de la feuille de calcul. La séance de travaux pratiques 1 vous demande de construire un tel dictionnaire.

La règle d'or du nettoyage des données est simple : une valeur suspecte ne se corrige pas à l'intuition, elle se vérifie dans le document source. Une pression de 1540 mmHg ressemble de toute évidence à 154 mmHg avec un chiffre en trop, mais il peut tout aussi bien s'agir de 150 ou de 140 mal saisis, ou d'une valeur tout à fait différente ; l'hypothèse la plus probable doit malgré tout être confirmée. Si le document source n'est pas disponible, la valeur est marquée comme manquante et la décision est consignée dans un journal des modifications. Les données brutes d'origine ne sont jamais effacées ; on travaille sur une copie.

1.8.3 L'éthique et la protection des données dans le traitement des données biomédicales

La fiche de la discipline exige, pour la note 5, une familiarisation avec les principes de l'éthique dans le traitement des données biomédicales. Ce n'est pas une exigence formelle. Les données de santé font partie des informations les plus sensibles sur une personne, et une fuite de données peut nuire à l'emploi, à l'assurance ou aux relations sociales du patient.

Le Règlement général sur la protection des données (RGPD) s'applique dans toute l'Union européenne depuis 2018 et range les données concernant la santé, les données génétiques et les données biométriques parmi les catégories particulières de données, dont le traitement est en principe interdit, sauf exceptions expressément prévues par le règlement, parmi lesquelles le consentement explicite de la personne, la prestation de soins de santé et la recherche scientifique assortie de garanties appropriées [32]. En Roumanie, la loi n° 190/2018 fixe les mesures de mise en œuvre du règlement, y compris des règles spécifiques pour le traitement des données génétiques, biométriques et concernant la santé ainsi que pour le traitement du code numérique personnel (l'équivalent roumain du numéro d'identification national) [33]. Le règlement relatif à l'espace européen des données de santé, adopté en 2025, ajoute un cadre pour l'utilisation secondaire des données de santé en recherche, par l'intermédiaire d'organismes d'accès habilités [34].

Quelques principes du RGPD se traduisent directement en règles de travail pour un étudiant qui fait de la recherche. La limitation des finalités : les données recueillies pour les soins ne sont pas librement réutilisées à d'autres fins. La minimisation des données : on ne recueille que les variables nécessaires à la question de recherche ; on ne demande pas la date de naissance complète si l'âge en années suffit, on ne note pas l'adresse si le milieu urbain ou rural suffit. L'intégrité et la confidentialité : les fichiers sont conservés sur des supports protégés, et non sur une clé USB oubliée en salle de travaux pratiques ni dans des comptes personnels de stockage en nuage.

La différence entre anonymisation et pseudonymisation fait souvent l'objet de questions. La pseudonymisation remplace les identifiants directs (nom, code numérique personnel, numéro de téléphone) par un code, par exemple P001, mais conserve séparément une table de correspondance qui permet de réidentifier le patient. Les données pseudonymisées restent des données à caractère personnel et relèvent toujours du RGPD. L'anonymisation supprime de manière irréversible toute possibilité raisonnable de réidentification, y compris par le croisement de plusieurs variables ; les données véritablement anonymes ne sont plus des données personnelles. Attention : supprimer le nom n'est pas anonymiser. Un fichier qui contient la date de naissance, le sexe, la commune et le diagnostic d'une maladie rare peut identifier un patient même sans son nom. Le jeu de données mg2026 comporte des codes P001-P160, mais, étant simulé, il ne correspond à aucune personne réelle.

Outre la législation sur les données, la recherche sur l'être humain est régie par des principes éthiques. La Déclaration d'Helsinki de l'Association médicale mondiale, révisée en 2024, exige que tout protocole de recherche impliquant des sujets humains soit approuvé au préalable par un comité d'éthique indépendant et que les participants donnent leur consentement éclairé, librement exprimé, après avoir reçu des informations claires sur le but, les risques, les bénéfices et leur droit de se retirer à tout moment [35]. Pour les études rétrospectives sur des données existantes, le comité d'éthique peut autoriser l'utilisation des données sans consentement individuel, dans les conditions prévues par la loi, si le risque est minime et que les données sont pseudonymisées ou anonymisées. Les revues médicales exigent que l'approbation du comité d'éthique et les modalités d'obtention du consentement soient mentionnées dans la section des méthodes [36].

L'éthique du traitement des données ne s'arrête pas à la protection de l'identité. Elle englobe aussi l'honnêteté de l'analyse : on n'exclut pas des patients "gênants" sans règle fixée à l'avance, on ne teste pas des dizaines d'hypothèses en ne rapportant que les résultats favorables, on ne cache pas les données manquantes et on ne présente pas une association comme une relation causale. Ces règles, longuement discutées dans la littérature sur l'éthique de l'informatique biomédicale [37], reviendront dans les chapitres consacrés aux tests statistiques et aux études cliniques.

> **La règle qui évite l'erreur**
>
> Avant de charger un fichier contenant des données de patients dans un logiciel en ligne, un formulaire, un outil d'intelligence artificielle ou une plateforme de stockage, posez-vous trois questions : (1) Le fichier contient-il des identifiants directs ou des combinaisons de variables permettant d'identifier une personne ? (2) Ai-je l'autorisation (du comité d'éthique, de l'établissement) de traiter ces données à cet endroit ? (3) Ai-je conservé la table de correspondance des codes séparément des données, dans un lieu protégé ? Si la réponse à l'une de ces questions est "je ne sais pas", ne chargez pas le fichier. En travaux pratiques, vous n'utiliserez que le jeu de données simulé mg2026 et les fichiers pédagogiques de Moodle, qui ne contiennent aucune donnée d'identification réelle.

1.8.4 Les données de routine en recherche : exemples tirés de notre région

De nombreuses études publiées par les collègues de la discipline utilisent des données existantes, recueillies au départ à d'autres fins. Une étude transversale des prescriptions électroniques en milieu urbain dans les départements de Timiș et d'Arad a analysé 1 498 ordonnances destinées à des personnes âgées, à l'aide des critères STOPP/START [38]. Les ordonnances avaient été établies à 57% pour des femmes, l'âge moyen des patients était de 74,1 ± 6,95 ans et le nombre moyen de médicaments par ordonnance de 4,7 ± 1,51 ; les statines figuraient sur 47,4% des ordonnances. La prévalence des médicaments potentiellement inappropriés était de 18,58% et celle des omissions potentielles de prescription de 49,2%. Remarquez que l'unité statistique est l'ordonnance, et non le patient : si un patient a reçu plusieurs ordonnances pendant la période étudiée, il est compté plusieurs fois, ce qu'il faut garder à l'esprit en lisant "57% de femmes".

Les données des systèmes de surveillance et des registres des centres de traitement constituent un autre type de données de routine. Une étude sur l'hémodialyse a comparé deux cohortes transversales issues de 10 centres roumains : 1 401 patients en 2015 et 1 698 en 2019, avant et après le programme national de traitement par antiviraux à action directe [39]. La prévalence de l'infection par le virus de l'hépatite C a diminué de 16,9% à 10,5% ($p < 0,0001$), tandis que la prévalence de l'infection par le virus de l'hépatite B est restée pratiquement inchangée (4,7% contre 4,8%, $p = 0,604$), et la mortalité à un an n'a pas différé significativement (9,9% contre 10,7%, $p = 0,46$).

> **Étude de cas : les données de surveillance et leurs limites**
>
> Une étude rétrospective a utilisé les données de surveillance de la Direction de santé publique de Timiș pour tous les cas confirmés de COVID-19 parmi le personnel de santé, entre le 1er janvier 2021 et le 31 mars 2022 [40]. L'incidence chez le personnel auxiliaire et le personnel paramédical a suivi de très près l'évolution dans la population générale, avec des coefficients de corrélation de Spearman de 0,852 et 0,821 ($p < 0,001$), et la couverture vaccinale a dépassé 70%. C'est un bel exemple de valorisation de données existantes, pour un coût très faible. Ce que dit le résultat : les vagues d'infection chez le personnel de santé ont coïncidé dans le temps avec les vagues dans la population. Ce qu'il ne dit pas : la corrélation entre deux séries temporelles agrégées ne montre pas où les soignants se sont infectés (à l'hôpital ou dans la communauté) et ne constitue pas une preuve de causalité. Il s'agit d'une association écologique, au niveau du groupe, qui peut s'expliquer par le fait que les deux séries

> dépendent du même facteur, la circulation du virus. L'étude sur l'hémodialyse évoquée plus haut présente une limite voisine : les deux cohortes, de 2015 et de 2019, ne sont pas constituées des mêmes patients, de sorte que la baisse de la prévalence du VHC décrit la population dialysée à ces deux moments, et non la guérison de patients suivis individuellement. Les cohortes différaient aussi par l'âge (59,7 ± 12,92 contre 61 ± 12,93 ans, p = 0,005) ; une différence de 1,3 an est statistiquement significative dans un échantillon de plus de 3 000 patients, mais cliniquement négligeable, une distinction que nous discuterons en détail au chapitre 5. Dans les deux cas, les données de routine permettent d'aborder des questions qui seraient autrement impossibles à traiter, à condition que l'interprétation respecte la manière dont elles ont été recueillies.

1.9 Éléments de théorie des probabilités

La statistique inférentielle s'appuie sur la théorie des probabilités. Vous n'avez pas besoin de démonstrations rigoureuses, mais vous devez maîtriser quelques règles simples, car elles sont à la base des valeurs p, des intervalles de confiance, du risque relatif et des valeurs prédictives des tests diagnostiques [41,42]. Tous les exemples de cette section utilisent le jeu de données mg2026.

1.9.1 Expérience aléatoire, issues, événements

Une expérience aléatoire est un processus dont le résultat ne peut pas être prédit avec certitude, bien que l'ensemble des résultats possibles soit connu. Le lancer d'un dé en est l'exemple classique. En médecine, l'expérience aléatoire peut être le choix au hasard d'un patient dans une liste, la mesure de la glycémie chez un patient, le résultat d'un test rapide ou l'évolution d'un patient après traitement.

L'ensemble de tous les résultats possibles s'appelle l'univers des possibles (ou espace échantillon) et se note Ω. Si nous choisissons au hasard un patient de mg2026 et notons son statut tabagique, Ω = {non-fumeur, ancien fumeur, fumeur}. Un événement est un sous-ensemble de l'univers des possibles, par exemple "le patient a déjà fumé" = {ancien fumeur, fumeur}. L'événement certain est Ω lui-même (il se réalise toujours), et l'événement impossible est l'ensemble vide.

1.9.2 Définition classique et définition fréquentiste

La définition classique de la probabilité s'applique lorsque tous les résultats sont également possibles :

P(A) = nombre de cas favorables / nombre de cas possibles

Si nous choisissons de façon totalement aléatoire un patient parmi les 160 du jeu de données mg2026, chaque patient a la même chance d'être choisi, et la probabilité qu'il s'agisse d'une femme est $P(F) = \frac{86}{160} = 0{,}5375$. La probabilité qu'il ait un diabète confirmé par HGPO est $P(D) = \frac{29}{160} = 0{,}18125$, soit environ 18,1%.

La définition fréquentiste (statistique) part de l'observation : la probabilité d'un événement est la limite vers laquelle tend la fréquence relative de cet événement lorsque l'expérience est répétée un très grand nombre de fois, dans les mêmes conditions. Si, sur 10 000 nouveau-nés d'une maternité, 5 140 sont des garçons, nous estimons la probabilité de naissance d'un garçon à 0,514. La loi des grands nombres garantit qu'à mesure que le nombre de répétitions augmente, la fréquence relative se rapproche de la vraie probabilité. Sur 10 naissances, la proportion de garçons peut être de 0,3 ou de 0,7 ; sur 10 000, elle sera très proche de la valeur réelle.

En médecine, nous utilisons presque toujours la définition fréquentiste. Lorsque nous disons que le risque de diabète dans une population est de 18%, nous faisons référence à une fréquence

relative observée sur un échantillon et utilisée comme estimation d'une probabilité. La proportion dans l'échantillon (p) estime la probabilité dans la population (π).

1.9.3 Les axiomes de la probabilité et l'événement contraire

Toute probabilité respecte trois règles de base (les axiomes de Kolmogorov) : (1) la probabilité de tout événement est un nombre compris entre 0 et 1, $0 \leq P(A) \leq 1$; (2) la probabilité de l'événement certain est 1, $P(\Omega) = 1$; (3) si deux événements ne peuvent pas se produire simultanément (ils sont mutuellement exclusifs), la probabilité que l'un d'eux se produise est la somme de leurs probabilités. Toutes les autres règles découlent de ces trois-là.

L'événement contraire de A, noté "non-A", se réalise lorsque A ne se réalise pas. Sa probabilité est :

$$P(\text{non} - A) = 1 - P(A)$$

La probabilité qu'un patient choisi au hasard dans mg2026 n'ait pas de diabète est de 1 - 0,18125 = 0,81875. La règle semble banale, mais elle est extrêmement utile pour les calculs du type "au moins un" : il est beaucoup plus simple de calculer la probabilité qu'aucun événement ne se produise et de la soustraire de 1.

1.9.4 La règle de l'addition

Pour deux événements mutuellement exclusifs (qui ne peuvent pas se produire ensemble), la probabilité que l'un ou l'autre se produise est :

$$P(A \text{ ou } B) = P(A) + P(B)$$

Un patient ne peut pas être à la fois fumeur et ancien fumeur. La probabilité qu'un patient choisi au hasard ait déjà fumé est donc $P(\text{ fumeur }) + P(\text{ancien fumeur }) = \frac{53}{160} + \frac{33}{160} = \frac{86}{160} = 0{,}5375$.

Si les événements peuvent se produire ensemble, la somme ci-dessus compterait deux fois les individus qui présentent les deux caractéristiques, et la formule devient :

$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ et } B)$$

Sur les 160 patients, 53 sont fumeurs, 54 sont obèses (IMC ≥ 30 kg/m²) et 22 sont à la fois fumeurs et obèses. La probabilité qu'un patient choisi au hasard soit fumeur ou obèse (ou les deux) est :

$$P(\text{ fumeur ou obèse}) = \frac{53}{160} + \frac{54}{160} - \frac{22}{160} = \frac{85}{160} = 0{,}531$$

Si nous avions simplement additionné, nous aurions obtenu 107/160 = 0,669, c'est-à-dire que nous aurions compté deux fois les 22 fumeurs obèses. À l'examen, c'est l'erreur la plus fréquente dans les problèmes portant sur la règle de l'addition.

1.9.5 La probabilité conditionnelle

La probabilité conditionnelle de l'événement A, sachant que l'événement B s'est produit, se note P(A|B) et se lit "probabilité de A sachant B". Elle se calcule en restreignant l'univers des possibles aux individus qui présentent B :

$$P(A \mid B) = \frac{P(A \text{ et } B)}{P(B)}$$

Ce concept est probablement le plus important de tout le chapitre, car c'est sur lui que reposent le risque relatif, la sensibilité, la spécificité et les valeurs prédictives. Appliquons-le au tableau 1.4, qui croise l'obésité (IMC ≥ 30 kg/m²) et le diabète confirmé par HGPO à l'inclusion dans l'étude mg2026.

Tableau 1.4. Obésité et diabète confirmé par HGPO à l'inclusion, jeu de données mg2026 (n = 160).

	Diabète (D)	**Sans diabète**	**Total**
Obèses (O), IMC ≥ 30	18	36	54
Non obèses, IMC < 30	11	95	106
Total	29	131	160

Le tableau permet de calculer plusieurs types de probabilités, et la différence entre elles doit être très bien comprise.

Les probabilités marginales concernent une seule variable et se calculent à partir des totaux situés en marge du tableau : $P(O) = \frac{54}{160} = 0{,}3375$; $P(D) = \frac{29}{160} = 0{,}181$.

La probabilité conjointe (de l'intersection) concerne les deux caractéristiques simultanément et se calcule à partir de la cellule correspondante rapportée au total général : $P(O \text{ et } D) = \frac{18}{160} = 0{,}1125$.

Les probabilités conditionnelles se calculent en rapportant la cellule au total de la ligne ou de la colonne qui exprime la condition. La probabilité d'avoir un diabète si le patient est obèse est $P(D \mid O) = \frac{18}{54} = 0{,}333 (33{,}3\%)$. La probabilité d'avoir un diabète si le patient n'est pas obèse est $P(D \mid \text{non} - O) = \frac{11}{106} = 0{,}104 (10{,}4\%)$. On peut aussi le vérifier avec la formule : $P(D \mid O) = \frac{P(O \text{ et } D)}{P(O)} = \frac{0{,}1125}{0{,}3375} = 0{,}333$.

Remarquez que P(D|O) n'est pas la même chose que P(O|D). La probabilité d'être obèse si le patient a un diabète est $P(O \mid D) = \frac{18}{29} = 0{,}621 (62{,}1\%)$. La première probabilité répond à la question "Combien d'obèses ont un diabète ?", la seconde à la question "Combien de diabétiques sont obèses ?". La confusion entre les deux, parfois appelée erreur de la conditionnelle inversée, est à l'origine de nombreux raisonnements erronés, en médecine et ailleurs. Au chapitre 11, vous verrez que la différence entre la sensibilité d'un test, P(test positif|maladie), et sa valeur prédictive positive, P(maladie|test positif), est exactement la différence entre P(O|D) et P(D|O).

De la formule de la probabilité conditionnelle découle la règle de multiplication, valable pour deux événements quelconques :

$$P(A \text{ et } B) = P(A) \times P(B \mid A) = P(B) \times P(A \mid B)$$

Sur le tableau 1.4 : $P(O \text{ et } D) = P(O) \times P(D \mid O) = 0{,}3375 \times 0{,}333 = 0{,}1125$, le même résultat que 18/160.

1.9.6 Les événements indépendants

Deux événements sont indépendants si la réalisation de l'un ne modifie pas la probabilité de l'autre, c'est-à-dire si P(A|B) = P(A). Pour des événements indépendants, la règle de multiplication se simplifie :

$P(A \text{ et } B) = P(A) \times P(B)$ (uniquement pour des événements indépendants)

Cette relation fournit aussi un moyen de vérifier l'indépendance. Si l'obésité et le diabète étaient indépendants dans mg2026, la probabilité conjointe devrait être $P(O) \times P(D) = 0{,}3375 \times 0{,}18125 = 0{,}0612$. La probabilité observée est pourtant de 0,1125, presque le double. De façon équivalente, P(D|O) = 0,333 est très différente de P(D) = 0,181 et de P(D|non-O) = 0,104. Nous concluons que, dans cet échantillon, les deux événements ne sont pas indépendants : l'obésité et le diabète apparaissent ensemble plus souvent que ne le voudrait le hasard. La question de savoir si la différence observée peut s'expliquer par le hasard (par la fluctuation d'échantillonnage) relève du test du khi-deux, au chapitre 7 ; la réponse, anticipée ici, est non ($\chi^2 = 12{,}70$, $p < 0{,}001$).

Un exemple d'indépendance approximative tiré du même jeu de données : la probabilité de diabète est de 15/86 = 0,174 chez les femmes et de 14/74 = 0,189 chez les hommes, toutes deux très proches de la probabilité globale de 0,181. Le sexe et le diabète semblent indépendants dans mg2026.

Ne confondez pas événements indépendants et événements mutuellement exclusifs. Deux événements mutuellement exclusifs, de probabilités non nulles, ne peuvent pas être indépendants : si nous savons que l'un s'est produit, nous savons avec certitude que l'autre ne s'est pas produit, donc sa probabilité est tombée à zéro. "Fumeur" et "ancien fumeur" sont mutuellement exclusifs et, précisément pour cette raison, fortement dépendants.

La règle de multiplication pour des événements indépendants est celle que nous utilisons lorsque les patients sont choisis indépendamment les uns des autres. Si la probabilité de diabète chez un patient est de 0,18, la probabilité que deux patients choisis indépendamment aient tous deux un diabète est de 0,18 x 0,18 = 0,0324, et la probabilité qu'aucun de dix patients n'ait de diabète est de 0,82 à la puissance dix, soit 0,137. Il en découle que la probabilité qu'au moins un des dix ait un diabète est de 1 - 0,137 = 0,863.

1.9.7 La formule des probabilités totales et un premier aperçu du théorème de Bayes

Nous connaissons parfois la probabilité d'un événement dans chacun des sous-groupes qui partagent la population et nous voulons la probabilité globale. Si les événements B1, B2,..., Bk sont mutuellement exclusifs et couvrent ensemble toute la population, alors :

$$P(A) = P(A \mid B_1) \times P(B_1) + P(A \mid B_2) \times P(B_2) + \cdots + P(A \mid B_k) \times P(B_k)$$

Il s'agit d'une moyenne pondérée des probabilités conditionnelles, avec des poids égaux à la taille de chaque sous-groupe. Sur le tableau 1.4 :

$$\begin{aligned} P(D) &= P(D \mid O) \times P(O) + P(D \mid \text{non} - O) \times P(\text{non} - O) \\ &= 0{,}333 \times 0{,}3375 + 0{,}104 \times 0{,}6625 = 0{,}1125 + 0{,}0688 = 0{,}181 \end{aligned}$$

Nous retrouvons, comme prévu, 29/160. La formule devient vraiment utile lorsque nous ne disposons pas du tableau complet, mais seulement des probabilités par sous-groupe. Supposons qu'un médecin de famille travaille avec deux listes : le cabinet en ville regroupe 60% des patients, avec une prévalence du diabète de 15% ; l'antenne de la commune rurale regroupe 40% des patients, avec une prévalence de 25%. La probabilité qu'un patient quelconque de ce médecin ait un diabète est de 0,15 x 0,60 + 0,25 x 0,40 = 0,09 + 0,10 = 0,19.

Le théorème de Bayes répond à la question inverse : si nous savons que le patient a un diabète, quelle est la probabilité qu'il vienne de la commune rurale ? En utilisant la règle de multiplication et la formule des probabilités totales :

$$P(\text{ commune } \mid D) = \frac{P(D \mid \text{ commune }) \times P(\text{ commune })}{P(D)} = \frac{0{,}10}{0{,}19} = 0{,}526$$

Bien que seulement 40% des patients viennent de la commune rurale, plus de la moitié des diabétiques (52,6%) en proviennent, car la prévalence est plus élevée dans ce sous-groupe. Sur le tableau mg2026, le même théorème donne $P(O \mid D) = \frac{P(D|O) \times P(O)}{P(D)} = \frac{0{,}333 \times 0{,}3375}{0{,}181} = 0{,}621$, c'est-à-dire la valeur calculée directement plus haut. La logique bayésienne, qui actualise une probabilité initiale (a priori) à partir d'une information nouvelle pour obtenir une probabilité finale (a posteriori), est le fondement de l'interprétation des tests diagnostiques et sera traitée en détail au chapitre 11.

1.9.8 Les lois de probabilité : un premier aperçu

Une variable aléatoire est une variable dont la valeur dépend d'une expérience aléatoire, par exemple le nombre de patients diabétiques parmi dix patients choisis au hasard. Sa loi de probabilité indique quelles valeurs elle peut prendre et avec quelle probabilité. Les lois se divisent en lois discrètes (pour les dénombrements : 0, 1, 2,...) et lois continues (pour les mesures, comme la pression artérielle ou la glycémie). La loi continue la plus importante, la loi normale, fait l'objet du chapitre 3. Nous présentons ici deux lois discrètes que vous rencontrerez souvent.

La loi binomiale décrit le nombre de "succès" obtenus en n épreuves indépendantes, chacune ayant la même probabilité de succès p. Le mot "succès" est conventionnel et désigne seulement l'événement compté, même s'il s'agit d'une maladie. Les conditions sont : un nombre fixe d'épreuves, deux issues possibles à chaque épreuve, une probabilité constante, l'indépendance entre les épreuves. La probabilité d'obtenir exactement k succès est :

$$P(X = k) = C(n,\ k) \times p^k \times (1-p)^{n-k}$$

où $C(n,\ k) = \frac{n!}{k! \times (n-k)!}$ est le nombre de façons de choisir les k succès parmi les n épreuves (combinaisons de n éléments pris k à k), et p^k signifie p à la puissance k.

Prenons un exemple tiré directement de notre scénario. La proportion de patients dont le diabète est confirmé dans mg2026 est d'environ 18%. Si un médecin réalise une HGPO chez 10 patients au profil similaire, choisis indépendamment, quelle est la probabilité qu'exactement 2 d'entre eux aient un diabète ? Nous avons n = 10, p = 0,18, k = 2 :

$$C(10,\ 2) = \frac{10 \times 9}{2} = 45$$

$$P(X = 2) = 45 \times 0{,}18^2 \times 0{,}82^8 = 45 \times 0{,}0324 \times 0{,}2044 = 0{,}298$$

Le tableau 1.5 donne la loi complète, et la figure 1.3A la représente graphiquement. La moyenne de la loi binomiale est n x p = 10 x 0,18 = 1,8 patient, et l'écart-type $\sqrt{[n \times p \times (1-p)]} = \sqrt{10 \times 0{,}18 \times 0{,}82} = \sqrt{1{,}476} = 1{,}21$.

Tableau 1.5. Loi binomiale du nombre de patients diabétiques sur 10 (p = 0,18).

k	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**≥ 7**
$P(X = k)$	0,137	0,302	0,298	0,174	0,067	0,018	0,003	< 0,001
P(X ≤ k)	0,137	0,439	0,737	0,912	0,979	0,996	1,000	1,000

Le tableau permet de lire directement les réponses à des questions pratiques. La probabilité qu'aucun patient n'ait de diabète est de 0,137 ; la probabilité qu'au moins un ait un diabète est de 1 - 0,137 = 0,863 ; la probabilité qu'au moins trois aient un diabète est $1 - P(X \leq 2) = 1 - 0{,}737 = 0{,}263$. Si le médecin trouve 6 diabétiques sur 10, ce résultat serait très peu probable (0,003) dans l'hypothèse où p = 0,18, ce qui suggère que ses patients ont un risque plus élevé que ceux de mg2026. Ce raisonnement, "quelle serait la probabilité du résultat observé si notre hypothèse était vraie", est exactement la logique des tests statistiques du chapitre 5.

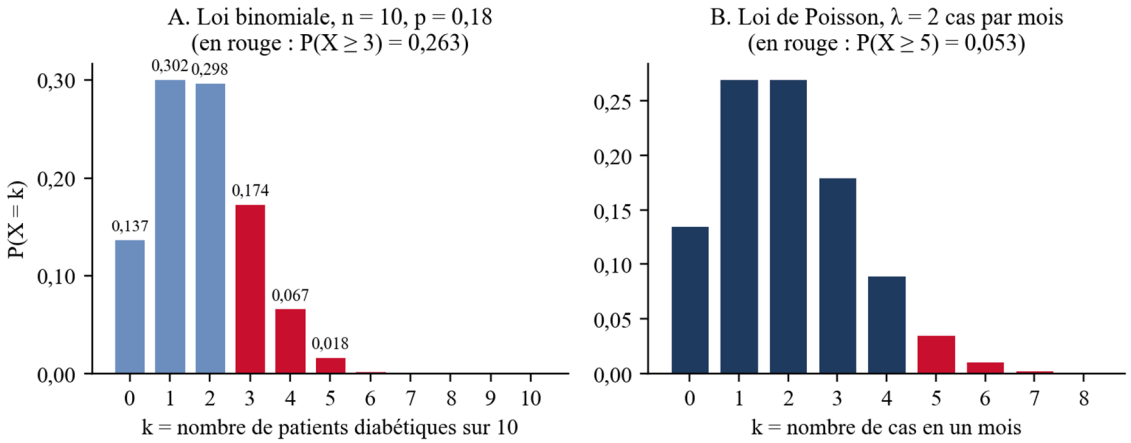


Figure 1.3. Deux lois discrètes. A : loi binomiale du nombre de patients diabétiques parmi 10 patients évalués, lorsque la probabilité individuelle est de 0,18 ; les barres rouges correspondent à l'événement "au moins 3 patients diabétiques" (probabilité totale 0,263). B : loi de Poisson pour un événement rare dont la moyenne est de 2 cas par mois ; les barres rouges correspondent à l'événement "au moins 5 cas" (probabilité 0,053). Calcul théorique.

La loi de Poisson décrit le nombre d'événements rares qui se produisent dans un intervalle de temps ou un espace donné, lorsque les événements surviennent indépendamment les uns des autres, avec un taux moyen constant λ (lambda). Les exemples typiques sont le nombre de nouveaux cas d'une maladie rare déclarés chaque mois dans un département, le nombre d'effets indésirables graves d'un vaccin pour 100 000 doses ou le nombre d'hospitalisations en urgence au cours d'une nuit dans un petit hôpital. La probabilité d'observer exactement k événements est :

$$P(X = k) = \frac{e^{-\lambda} \times \lambda^{k}}{k!}$$

où e ≈ 2,718 est la base des logarithmes naturels. La particularité de la loi de Poisson est que sa moyenne et sa variance sont égales, toutes deux à λ. Si une direction de santé publique reçoit en moyenne 2 déclarations par mois d'une maladie rare, la probabilité d'un mois sans aucun cas est $e^{-2} = 0{,}135$, et la probabilité d'au moins 5 cas en un mois est de 0,053 (figure 1.3B). Un mois avec 5 cas ne prouve pas l'existence d'un foyer, mais il survient assez rarement (environ un mois sur 19) pour justifier une vérification. C'est ainsi que fonctionnent, pour l'essentiel, les systèmes d'alerte épidémiologique.

La loi de Poisson est aussi une bonne approximation de la loi binomiale lorsque n est grand et p très petit, avec λ = n x p. Pour un effet indésirable de probabilité 1/10 000 chez 20 000 patients traités, le nombre d'effets suit approximativement une loi de Poisson de paramètre λ = 2.

1.10 Les logiciels utilisés dans ce cours

Les calculs à la main sont indispensables pour comprendre ce que fait une méthode, mais personne ne calcule à la main un écart-type sur 160 valeurs. La fiche de la discipline prévoit quelques logiciels, choisis parce qu'ils sont gratuits ou disposent d'une version gratuite, qu'ils fonctionnent sur les ordinateurs de la salle de travaux pratiques et qu'ils couvrent les méthodes du cours. Chacun a son rôle, et vous les utiliserez tous en travaux pratiques.

Google Sheets est le tableur de la suite Google, accessible depuis le navigateur avec le compte universitaire. Nous l'utilisons pour la saisie et l'organisation des données, pour les règles de validation, pour le nettoyage des données, pour les calculs descriptifs avec des fonctions comme AVERAGE, MEDIAN, STDEV, COUNTIF, pour les tableaux croisés dynamiques et pour les graphiques simples [43,44]. C'est l'outil le plus proche des feuilles de calcul que vous rencontrerez sur n'importe quel lieu de travail. Avec les paramètres régionaux France (ou

Roumanie), les décimales s'écrivent avec une virgule et les arguments des fonctions sont séparés par un point-virgule.

Google Forms permet de construire rapidement un questionnaire ou un formulaire de recueil de données, avec des types de questions qui correspondent aux types de variables et des règles de validation des réponses ; les réponses sont enregistrées automatiquement dans une feuille Google Sheets [45,46]. Vous l'utiliserez lors de la séance de travaux pratiques 2 et lors de la séance consacrée à la conception des études cliniques.

JASP est un logiciel statistique gratuit et open source, développé à l'Université d'Amsterdam [47]. Il dispose d'une interface à menus, les résultats s'affichent immédiatement dans des tableaux mis en forme selon les conventions académiques, et les analyses se mettent à jour automatiquement lorsque l'on modifie les options. Il couvre la statistique descriptive, les tests paramétriques et non paramétriques, la corrélation, la régression et bien d'autres méthodes, y compris des variantes bayésiennes des tests. Il existe des guides destinés aux étudiants, par exemple pour la statistique médicale [48,49].

MedCalc est un logiciel statistique orienté vers la recherche médicale, très utilisé pour les courbes ROC, l'analyse des tests diagnostiques, la comparaison de méthodes de mesure et la méta-analyse [50]. La version complète est commerciale, mais le site de l'éditeur propose gratuitement des calculateurs en ligne pour de nombreux problèmes du cours : intervalles de confiance des proportions, risque relatif, rapport de cotes, sensibilité et spécificité, nombre de sujets à traiter. Vous les utiliserez surtout aux chapitres 10-12.

Epi Info est développé par les Centers for Disease Control and Prevention (CDC) des États-Unis et il est destiné à l'épidémiologie de terrain : construction de formulaires, saisie des données, tableaux de fréquences, tableaux 2x2 avec risque relatif et rapport de cotes, cartes [51]. Il a l'avantage de fonctionner aussi sans connexion Internet, ce qui le rend utile pour des enquêtes épidémiologiques menées dans des lieux isolés. La dernière version est la 7.2.7 ; les CDC ont arrêté le développement et le support du logiciel en septembre 2025, de sorte que nous l'utilisons pour des exercices classiques d'épidémiologie, en sachant qu'il ne recevra plus de mises à jour.

Tableau 1.6. Les logiciels de la fiche de la discipline et leur utilisation principale.

Logiciel	Type	Utilisation principale dans le cours
Google Sheets	tableur en ligne	saisie, validation et nettoyage des données ; statistiques descriptives ; tableaux croisés dynamiques ; graphiques
Google Forms	formulaires en ligne	questionnaires et formulaires de recueil de données
JASP	logiciel statistique gratuit	statistiques descriptives, normalité, tests t, ANOVA, tests non paramétriques, corrélation, régression
MedCalc	logiciel de statistique médicale, calculateurs en ligne	ROC, tests diagnostiques, risque relatif, rapport de cotes, intervalles de confiance
Epi Info	logiciel d'épidémiologie, fonctionne sans réseau	formulaires, fréquences, tableaux 2x2, analyse du risque

Les supports de cours, les fichiers de données et les devoirs se trouvent sur la plateforme Moodle de l'université (Emediqual), où l'on dépose aussi les travaux à rendre [52]. Un dernier conseil pratique : quel que soit le logiciel, vérifiez toujours le nombre d'observations (n) indiqué dans chaque tableau de résultats. Si vous attendez 160 et que le logiciel en indique 158, c'est que deux valeurs manquent ou ont été lues comme du texte, et ce point doit être éclairci avant toute interprétation.

1.11 Problèmes résolus

Problème résolu 1.1. Le tabagisme et l'obésité dans le jeu de données mg2026

Énoncé. Le tableau 1.7 croise le statut tabagique et l'obésité (IMC ≥ 30 kg/m²) à l'inclusion, pour les 160 patients du jeu de données mg2026. On choisit un patient au hasard. Calculez : (a) la probabilité que le patient soit fumeur ; (b) la probabilité que le patient soit fumeur et obèse ; (c) la probabilité que le patient soit fumeur ou obèse ; (d) la probabilité d'obésité chez les fumeurs, les anciens fumeurs et les non-fumeurs ; (e) vérifiez si les événements "fumeur" et "obèse" sont indépendants.

Tableau 1.7. Statut tabagique et obésité à l'inclusion, jeu de données mg2026.

	Obèses	**Non obèses**	**Total**
Fumeurs	22	31	53
Anciens fumeurs	9	24	33
Non-fumeurs	23	51	74
Total	54	106	160

Solution. Tous les patients ont la même probabilité d'être choisis, nous utilisons donc la définition classique, avec 160 cas possibles.

(a) $P(\text{ fumeur }) = \frac{53}{160} = 0{,}331.$

(b) La probabilité conjointe se lit dans la cellule située à l'intersection : $P(\text{ fumeur et obèse}) = \frac{22}{160} = 0{,}1375.$

(c) Les événements ne sont pas mutuellement exclusifs (22 patients présentent les deux caractéristiques), nous utilisons donc la règle générale de l'addition : $P(\text{ fumeur ou obèse}) = \frac{53}{160} + \frac{54}{160} - \frac{22}{160} = \frac{85}{160} = 0{,}531.$

(d) Les probabilités conditionnelles se calculent par ligne : $P(\text{obèse} \mid \text{ fumeur }) = \frac{22}{53} = 0{,}415$; $P(\text{obèse} \mid \text{ancien fumeur}) = \frac{9}{33} = 0{,}273$; $P(\text{obèse} \mid \text{ non-fumeur}) = \frac{23}{74} = 0{,}311.$

(e) Si les deux événements étaient indépendants, nous aurions $P(\text{ fumeur et obèse}) = P(\text{ fumeur }) \times P(\text{obèse}) = 0{,}331 \times 0{,}3375 = 0{,}112$. La valeur observée est de 0,1375, un peu plus élevée. De façon équivalente, P(obèse|fumeur) = 0,415 est supérieure à P(obèse) = 0,3375. Les événements ne sont pas parfaitement indépendants dans l'échantillon.

Interprétation. Dans le jeu de données mg2026, l'obésité est un peu plus fréquente chez les fumeurs (41,5%) que chez les non-fumeurs (31,1%) et les anciens fumeurs (27,3%). Les différences sont toutefois bien plus faibles que dans le cas de l'obésité et du diabète, et, sur un échantillon de 160 patients, elles peuvent aussi être dues au hasard. Le fait que P(A et B) diffère de P(A) x P(B) dans un échantillon ne prouve pas la dépendance dans la population ; pour cela, nous avons besoin d'un test statistique (chapitre 7). Sur le plan clinique, l'association entre tabagisme et obésité n'est pas une relation causale directe ; tous deux peuvent être liés à d'autres facteurs du mode de vie.

Problème résolu 1.2. La probabilité binomiale : les patients hypertendus de la salle d'attente

Énoncé. Dans le jeu de données mg2026, 41 des 160 patients (25,6%) ont à l'inclusion une pression systolique ≥ 140 mmHg. Nous supposons que la probabilité qu'un patient au profil similaire ait une PAS ≥ 140 mmHg est p = 0,256 et que les patients sont indépendants. Dans la salle d'attente du cabinet se trouvent 5 patients de ce type. Quelle est la probabilité que : (a) aucun

n'ait une PAS ≥ 140 ; (b) au moins un ait une PAS ≥ 140 ; (c) exactement deux aient une PAS ≥ 140 ; (d) au moins deux aient une PAS ≥ 140 ?

Solution. Nous vérifions les conditions de la loi binomiale : un nombre fixe de patients (n = 5), deux issues possibles pour chacun (PAS ≥ 140 ou non), une probabilité constante (0,256), l'indépendance (supposée). Elles sont remplies.

(a) $P(X = 0) = (1 - 0{,}256)^5 = 0{,}744^5$. Calculons pas à pas : 0,744² = 0,5535 ; 0,744^4 = 0,5535² = 0,3064 ; $0{,}744^5 = 0{,}3064 \times 0{,}744 = 0{,}228$.

(b) Nous utilisons l'événement contraire : $P(X \geq 1) = 1 - P(X = 0) = 1 - 0{,}228 = 0{,}772$.

(c) $C(5,\ 2) = \frac{5 \times 4}{2} = 10$. $P(X = 2) = 10 \times 0{,}256^2 \times 0{,}744^3 = 10 \times 0{,}0655 \times 0{,}4118 = 0{,}270$.

(d) Nous avons aussi besoin de $P(X = 1) = 5 \times 0{,}256 \times 0{,}744^4 = 5 \times 0{,}256 \times 0{,}3064 = 0{,}392$. Ensuite $P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - 0{,}228 - 0{,}392 = 0{,}380$.

Interprétation. Avec une probabilité d'environ 77%, au moins un des cinq patients aura une pression systolique d'au moins 140 mmHg à la mesure, et avec une probabilité de 38%, au moins deux. En pratique, lors d'une journée ordinaire, le médecin doit s'attendre à des valeurs élevées dans presque chaque groupe de cinq patients à risque cardiométabolique. Attention : une seule mesure ≥ 140 mmHg ne signifie pas un diagnostic d'hypertension, qui exige des mesures répétées selon un protocole standardisé.

Problème résolu 1.3. Conception d'un échantillon dans un cabinet

Énoncé. Un cabinet de médecine générale compte sur sa liste 1 600 adultes de 40 à 75 ans : 1 000 habitant en ville et 600 dans les villages rattachés. Le médecin veut évaluer le risque de diabète à l'aide du score FINDRISC sur un échantillon de 80 patients. (a) Décrivez un échantillonnage systématique. (b) Décrivez un échantillonnage stratifié proportionnel selon le milieu de résidence. (c) Que se passerait-il si le médecin incluait les 80 premiers patients qui viennent au cabinet au mois de janvier ?

Solution. (a) Le pas d'échantillonnage est k = N/n = 1 600/80 = 20. On tire au hasard un nombre entre 1 et 20, par exemple 13 (dans Google Sheets, =RANDBETWEEN(1;20)). On inclut les patients portant les numéros d'ordre 13, 33, 53, 73, et ainsi de suite, c'est-à-dire 13 + 20 x (i - 1) pour i = 1,..., 80. Le dernier inclus est le patient numéro 13 + 20 x 79 = 1 593. Avant d'appliquer la méthode, le médecin doit vérifier que la liste n'est pas ordonnée selon un critère périodique.

(b) Les proportions dans la population sont de 1 000/1 600 = 62,5% pour la ville et de 600/1 600 = 37,5% pour les villages. L'échantillon de 80 comprendra 80 x 0,625 = 50 patients de la ville et 80 x 0,375 = 30 des villages, choisis par échantillonnage aléatoire simple dans chaque strate. La fraction de sondage est la même dans les deux strates : 50/1 000 = 30/600 = 5%.

(c) Les 80 premiers patients qui viennent au cabinet forment un échantillon de convenance. Les patients qui consultent souvent sont en général plus malades ou plus âgés (davantage de maladies chroniques, ordonnances mensuelles), et en janvier les infections respiratoires prédominent. Les patients en bonne santé, qui ne sont pas venus au cabinet depuis des mois, seraient sous-représentés. La prévalence du risque élevé serait très probablement surestimée. Il s'agit d'un biais de sélection.

Interprétation. L'échantillonnage stratifié garantit que les villageois, qui accèdent plus difficilement au cabinet et ont peut-être un profil de risque différent, sont présents exactement

dans leur proportion dans la population. La variante systématique est plus simple, mais la proportion de villageois dans l'échantillon pourrait varier par hasard autour de 37,5%.

Problème résolu 1.4. Reconnaître les biais

Énoncé. Pour chacune des situations ci-dessous, identifiez le type de biais et proposez une mesure de prévention. (1) Dans une étude cas-témoins sur la consommation d'anti-inflammatoires non stéroïdiens et l'hémorragie digestive, les cas sont interrogés à l'hôpital, quelques jours après l'épisode, tandis que les témoins sont interrogés par téléphone. (2) Un programme d'éducation thérapeutique pour le diabète est évalué en comparant les participants au programme aux patients qui ont refusé d'y participer. (3) Une étude montre que les personnes qui ont sur elles un briquet ont plus souvent une maladie coronarienne. (4) L'infirmière qui mesure la pression artérielle sait quels patients sont dans le bras avec téléconsultation et arrondit les valeurs à la valeur la plus proche se terminant par 0.

Solution. (1) Biais d'information, plus précisément biais de mémorisation : les cas, préoccupés par la cause de leur maladie, se souviennent mieux des médicaments pris que les témoins ; de plus, le mode d'entretien différent (en face à face pour les uns, par téléphone pour les autres) peut produire des différences de déclaration. Prévention : même mode d'entretien pour les cas et les témoins, questionnaire structuré, vérification de la consommation dans le dossier médical ou dans la base de données des prescriptions.

(2) Biais de sélection (autosélection, effet du volontaire en bonne santé) : les patients qui acceptent un programme éducatif sont habituellement plus motivés et plus observants. La différence d'équilibre glycémique pourrait s'expliquer par la motivation, et non par le programme. Prévention : randomisation des patients éligibles entre un groupe avec programme et un groupe sans programme.

(3) Confusion : le fait d'avoir un briquet sur soi est associé au tabagisme, et le tabagisme provoque la maladie coronarienne. Le briquet n'a aucun effet propre. Prévention : stratification selon le statut tabagique ou ajustement sur le tabagisme dans un modèle multivariable.

(4) Biais d'information : biais de l'observateur (la connaissance du bras peut influencer involontairement la mesure) et préférence pour le chiffre zéro. Prévention : évaluateurs qui ne connaissent pas le bras du patient (insu), tensiomètres automatiques, protocole de mesure standardisé, avec enregistrement de la valeur exacte affichée.

Interprétation. Le biais de sélection concerne qui entre dans l'étude, le biais d'information concerne la manière dont les variables sont mesurées, et la confusion concerne l'interprétation d'une association en présence d'une troisième variable. Tous trois déplacent le résultat de manière systématique et ne disparaissent pas quand on augmente l'échantillon.

Problème résolu 1.5. Probabilités totales et théorème de Bayes avec trois cabinets

Énoncé. Un réseau de trois cabinets de médecine générale participe à un programme de dépistage du diabète. Le cabinet 1 fournit 50% des patients évalués, le cabinet 2 30% et le cabinet 3 20%. La proportion de patients atteints d'un diabète non diagnostiqué auparavant est de 12% dans le cabinet 1, de 20% dans le cabinet 2 et de 30% dans le cabinet 3. (a) Quelle est la probabilité qu'un patient quelconque du programme ait un diabète ? (b) Un patient a un diabète confirmé ; quelle est la probabilité qu'il provienne du cabinet 3 ?

Solution. Notons C1, C2, C3 les événements "le patient provient du cabinet 1, 2, 3" ; ils sont mutuellement exclusifs et couvrent toute la population du programme.

(a) Formule des probabilités totales :

$$P(D) = P(D \mid C_1) \times P(C_1) + P(D \mid C_2) \times P(C_2) + P(D \mid C_3) \times P(C_3)\, P(D)$$
$$= 0{,}12 \times 0{,}50 + 0{,}20 \times 0{,}30 + 0{,}30 \times 0{,}20 = 0{,}06 + 0{,}06 + 0{,}06 = 0{,}18$$

(b) Théorème de Bayes :

$$P(C_3 \mid D) = \frac{P(D \mid C_3) \times P(C_3)}{P(D)} = \frac{0{,}06}{0{,}18} = 0{,}333$$

Interprétation. La prévalence globale du diabète dans le programme est de 18%, une moyenne pondérée des trois prévalences, et non leur moyenne simple (qui $\frac{\text{ar fi}\quad (12+20+30)}{3} = 20{,}7\%$). Bien que le cabinet 3 ne fournisse que 20% des patients, il fournit un tiers des cas de diabète, car sa prévalence est la plus élevée. Remarquez la symétrie : chaque cabinet contribue pour 6 points de pourcentage à la prévalence globale, donc chacun fournit un tiers des cas. La planification des ressources pour la confirmation du diagnostic (HGPO, consultations) devrait tenir compte de cette répartition, et non du nombre de patients évalués dans chaque cabinet.

Problème résolu 1.6. La loi de Poisson dans un système d'alerte

Énoncé. La direction de santé publique d'un département reçoit en moyenne 3 déclarations d'hépatite virale A par mois. Nous supposons que les déclarations sont indépendantes et surviennent à un taux constant. (a) Quelle est la probabilité qu'aucun cas ne soit déclaré au cours d'un mois ? (b) Quelle est la probabilité d'exactement 3 cas ? (c) Au cours d'un mois, 7 cas sont déclarés ; à quel point ce nombre est-il inhabituel ?

Solution. Le nombre mensuel de cas suit approximativement une loi de Poisson de paramètre $\lambda = 3$.

(a) $P(X = 0) = \frac{e^{-3} \times 3^0}{0!} = e^{-3} = 0{,}0498$. Environ un mois sur 20 ne comptera aucun cas.

(b) $P(X = 3) = \frac{e^{-3} \times 3^3}{3!} = \frac{0{,}0498 \times 27}{6} = 0{,}224$.

(c) Nous calculons la probabilité d'au moins 7 cas à l'aide de l'événement contraire. Les probabilités pour k = 0, 1,..., 6 sont 0,0498 ; 0,1494 ; 0,2240 ; 0,2240 ; 0,1680 ; 0,1008 ; 0,0504, dont la somme vaut 0,966. Donc $P(X \geq 7) = 1 - 0{,}9665 = 0{,}034$.

Interprétation. Un mois avec 7 cas ou plus survient, dans des conditions habituelles, avec une probabilité d'environ 3,4%, c'est-à-dire moins d'une fois tous les deux ans. Ce résultat ne prouve pas l'existence d'un foyer, mais il est suffisamment inhabituel pour justifier une enquête épidémiologique : les cas ont-ils une source commune, proviennent-ils de la même localité, de la même collectivité d'enfants ? C'est ainsi que l'on fixe, en principe, les seuils d'alerte dans la surveillance des maladies transmissibles.

1.12 Exercices proposés

1. Dans le jeu de données mg2026, 19 des 95 patients du milieu urbain et 10 des 65 patients du milieu rural ont un diabète confirmé par HGPO. Construisez le tableau 2x2 et calculez : P(D|urbain), P(D|rural), P(urbain et D), P(urbain) x P(D). Les événements "urbain" et "diabète" sont-ils indépendants dans l'échantillon ?
2. Dans mg2026, il y a 86 femmes et 53 fumeurs, et 30 des femmes sont fumeuses. Calculez P(fumeur|femme), P(femme|fumeur) et P(femme ou fumeur). Expliquez pourquoi les deux premières probabilités sont différentes.
3. La prévalence de l'hypertension artérielle dans une population d'adultes de plus de 40 ans est de 25%. On choisit 8 personnes au hasard. Calculez la probabilité que : (a) aucune ne

soit hypertendue ; (b) exactement une soit hypertendue ; (c) au moins deux soient hypertendues.

4. Un hôpital compte dans son registre des sorties 2 400 patients atteints d'insuffisance cardiaque au cours des trois dernières années. On souhaite constituer un échantillon systématique de 120 dossiers. Quel est le pas d'échantillonnage ? Si le numéro de départ tiré est 9, écrivez les cinq premiers numéros de dossier inclus ainsi que le dernier.
5. Un cabinet compte 1 500 adultes sur sa liste : 600 âgés de 18 à 39 ans, 600 de 40 à 64 ans et 300 de 65 ans et plus. Combien de patients de chaque classe d'âge faut-il inclure dans un échantillon stratifié proportionnel de 90 patients ? Quelle est la fraction de sondage ?
6. Classez le biais (sélection, information ou confusion) dans chacune des situations suivantes : (a) un questionnaire sur la consommation d'alcool rempli en présence du médecin de famille ; (b) une étude sur l'effet de la vitamine D sur les fractures, dans laquelle les personnes qui prennent des compléments sont aussi celles qui font le plus d'exercice ; (c) une étude sur l'efficacité d'un traitement de la BPCO dans laquelle 30% des patients du groupe traité abandonnent l'étude en raison d'effets indésirables et ne sont plus évalués ; (d) un nouveau test diagnostique du cancer de la prostate évalué chez des patients métastatiques et chez des étudiants volontaires en bonne santé.
7. Lisez la description de l'étude sur la prescription chez les personnes âgées à la section 1.8.4 [38]. Précisez : l'unité statistique, l'effectif de l'échantillon, la population cible probable, le type d'étude, une variable numérique et une variable qualitative. Quel problème peut se poser si un même patient a plusieurs ordonnances dans la base ?
8. Un service d'urgences pédiatriques reçoit en moyenne 1,5 cas par semaine d'intoxication médicamenteuse accidentelle. En supposant une loi de Poisson, calculez la probabilité qu'une semaine ne compte aucun cas et la probabilité d'au moins 4 cas.
9. Dans un cabinet, 30% des adultes sont fumeurs. La prévalence de la BPCO est de 15% chez les fumeurs et de 3% chez les non-fumeurs. Calculez la prévalence de la BPCO dans la population du cabinet et la probabilité qu'un patient atteint de BPCO soit fumeur.
10.

    (a) Formulez au format PICO la question : "La téléconsultation aide-t-elle à faire baisser la pression artérielle ?", en utilisant le scénario mg2026 et les variables tas0 et tas6. (b) Un fichier contient les colonnes : code patient (P001...), âge en années, sexe, IMC, glycémie ; la table qui relie les codes aux noms des patients est conservée par le médecin investigateur dans un dossier séparé et protégé. Les données sont-elles anonymisées ou pseudonymisées ? Quel régime juridique leur est applicable ?

Pièges fréquents

- Confondre P(A|B) avec P(B|A) : la probabilité de diabète chez les obèses (33,3%) n'est pas la probabilité d'obésité chez les diabétiques (62,1%).
- Appliquer la règle simple de l'addition, P(A) + P(B), à des événements qui peuvent se produire ensemble ; le résultat peut même dépasser 1.
- Confondre événements indépendants et événements mutuellement exclusifs.
- Croire qu'un grand échantillon est automatiquement représentatif ; un effectif important réduit l'erreur aléatoire, pas l'erreur systématique.
- Utiliser des lettres grecques (μ, σ) pour des valeurs calculées sur l'échantillon ; pour l'échantillon, on écrit m et s.
- Corriger "à l'intuition" une valeur aberrante au lieu de la vérifier dans le document source.
- Considérer un fichier comme anonyme simplement parce que les noms des patients ont été supprimés.

À retenir

- La variabilité biologique est la raison pour laquelle la médecine a besoin de la statistique ; elle se décrit, elle ne s'élimine pas.
- La population se définit dans l'espace et dans le temps ; l'échantillon en est une partie, choisie de manière à être représentative (équiprobabilité et indépendance).
- L'échantillonnage aléatoire simple, systématique, stratifié et en grappes sont des méthodes probabilistes ; l'échantillon de convenance est fréquent, mais exposé au biais de sélection.
- Les biais (sélection, information, confusion) ne diminuent pas quand l'échantillon augmente ; ils se préviennent par le schéma d'étude.
- La statistique descriptive résume l'échantillon ; la statistique inférentielle généralise à la population, avec une incertitude quantifiée.
- Les paramètres de la population (μ, σ, π) sont inconnus ; les statistiques de l'échantillon (m, s, p_i) les estiment.
- La probabilité conditionnelle $P(A \mid B) = \frac{P(A \text{ et } B)}{P(B)}$ est à la base du risque relatif et des tests diagnostiques.
- Les événements indépendants vérifient $P(A \text{ et } B) = P(A) \times P(B)$; la vérification sur l'échantillon ne remplace pas le test statistique.
- Les données de santé sont des données sensibles (catégorie particulière au sens du RGPD et de la loi roumaine 190/2018) ; les données pseudonymisées restent des données personnelles ; la recherche sur l'être humain exige l'avis du comité d'éthique et, en règle générale, un consentement éclairé.

Questions d'autoévaluation

1. (QRU) Dans une étude qui analyse 1 498 ordonnances électroniques établies pour des personnes âgées, l'unité statistique est :
    a) le patient âgé
    b) l'ordonnance
    c) la pharmacie
    d) le médecin prescripteur
    e) le département
2. (QRU) La notation correcte de la moyenne d'une variable dans la population est :
    a) m
    b) s
    c) μ
    d) p
    e) σ
3. (QRU) Dans le tableau 1.4 (obésité et diabète, mg2026), la probabilité qu'un patient obèse ait un diabète est :
    a) 0,113
    b) 0,181
    c) 0,333
    d) 0,621
    e) 0,104
4. (QRU) Pour un échantillonnage systématique de 60 patients à partir d'une liste de 900, le pas d'échantillonnage est :

a) 60
b) 15
c) 9
d) 6
e) 90

5. (QRU) Si la probabilité qu'un patient ait un diabète est de 0,18 et que les patients sont indépendants, la probabilité qu'aucun de 10 patients n'ait de diabète est d'environ :
   a) 0,18
   b) 0,82
   c) 0,137
   d) 0,863
   e) 0,298
6. (QRM) Au sujet de la pseudonymisation des données, les affirmations correctes sont :
   a) elle remplace les identifiants directs par des codes
   b) les données pseudonymisées restent des données à caractère personnel au sens du RGPD
   c) elle est irréversible, le patient ne peut plus être réidentifié d'aucune façon
   d) la table de correspondance entre les codes et l'identité est conservée séparément des données
   e) elle dispense l'étude de l'avis du comité d'éthique
7. (QRM) Au sujet des biais, les affirmations correctes sont :
   a) ils diminuent à mesure que l'effectif de l'échantillon augmente
   b) le biais de mémorisation est un biais d'information
   c) la randomisation équilibre entre les groupes les facteurs de confusion connus et inconnus
   d) l'effet du volontaire en bonne santé est un biais de sélection
   e) un intervalle de confiance étroit garantit l'absence de biais
8. (QRM) Soient A et B deux événements mutuellement exclusifs, avec P(A) > 0 et P(B) > 0. Les affirmations correctes sont :
   a) $P(A \text{ et } B) = 0$
   b) $P(A \text{ ou } B) = P(A) + P(B)$
   c) A et B sont indépendants
   d) $P(A \mid B) = 0$
   e) P(A) + P(B) peut être supérieur à 1
9. (QRM) La statistique inférentielle :
   a) estime les paramètres de la population par des intervalles de confiance
   b) teste des hypothèses sur la population
   c) porte exclusivement sur les individus de l'échantillon
   d) s'appuie sur la théorie des probabilités
   e) consiste à construire des tableaux de fréquences
10. (QRM) Au sujet de l'échantillon de convenance, les affirmations correctes sont :

a) il est constitué des individus les plus faciles à recruter
b) il garantit la représentativité si son effectif dépasse 500

c) il peut être utilisé dans un essai randomisé, où la validité de la comparaison entre les bras provient de la randomisation

d) un sondage en ligne diffusé sur les réseaux sociaux en est un exemple

e) il élimine le biais de sélection

Chapitre 2. Description et organisation des données : types de variables, indicateurs descriptifs et représentations graphiques

Objectifs

À la fin de ce chapitre, l'étudiant sera capable de :

- classer une variable selon son type (nominale, ordinale, numérique discrète ou continue) et selon son échelle de mesure (nominale, ordinale, d'intervalle, de rapport) ;
- organiser les données d'une étude dans un tableau comportant une ligne par patient et une colonne par variable, accompagné d'un dictionnaire des données, et reconnaître les erreurs de saisie fréquentes ;
- construire un tableau de fréquences absolues, relatives et cumulées, y compris pour des données regroupées en classes, et reconstituer les valeurs manquantes d'un tel tableau ;
- calculer à la main et avec le logiciel la moyenne, la médiane, le mode, la moyenne pondérée et la moyenne géométrique, et choisir l'indicateur adapté au type de données et à la forme de la distribution ;
- calculer l'étendue, l'écart moyen absolu, la variance, l'écart-type, le coefficient de variation et l'écart interquartile ;
- expliquer la différence entre l'écart-type et l'erreur standard de la moyenne ;
- déterminer les quartiles, les percentiles et le résumé en cinq valeurs, et identifier les valeurs extrêmes à l'aide de la règle 1,5 x EIQ ;
- choisir le graphique adapté à chaque type de variable et reconnaître les graphiques trompeurs ;
- lire et construire un tableau descriptif du type "Tableau 1" d'un article médical.

2.1 Du patient au tableau de données

Un patient qui entre dans le cabinet n'est pas une suite de chiffres. Pour pouvoir être analysé statistiquement, il doit toutefois être décrit par un ensemble de caractéristiques mesurées ou observées, les mêmes pour tous les patients de l'étude. Chaque caractéristique s'appelle une variable, parce qu'elle prend des valeurs différentes d'un patient à l'autre. L'âge, le sexe, la pression artérielle, le statut tabagique, le niveau d'instruction, la satisfaction à l'égard du traitement sont des variables. La valeur d'une variable chez un patient donné s'appelle une observation.

Le type d'une variable détermine presque tout ce qui suit : quels indicateurs descriptifs peuvent être calculés, quel graphique convient, quel test statistique peut être appliqué. C'est pourquoi la première étape de toute analyse, à l'examen comme en recherche, est le classement correct des variables. Les étudiants qui sautent cette étape finissent par calculer la moyenne du sexe ou l'écart-type du bras de traitement, des résultats que le logiciel affiche sans protester, mais qui n'ont aucun sens.

2.1.1 Variables qualitatives et quantitatives

La figure 2.1 résume la classification de base. Les variables qualitatives (catégorielles) placent chaque individu dans une catégorie. Les variables quantitatives (numériques) expriment une grandeur par un nombre, généralement accompagné d'une unité de mesure.

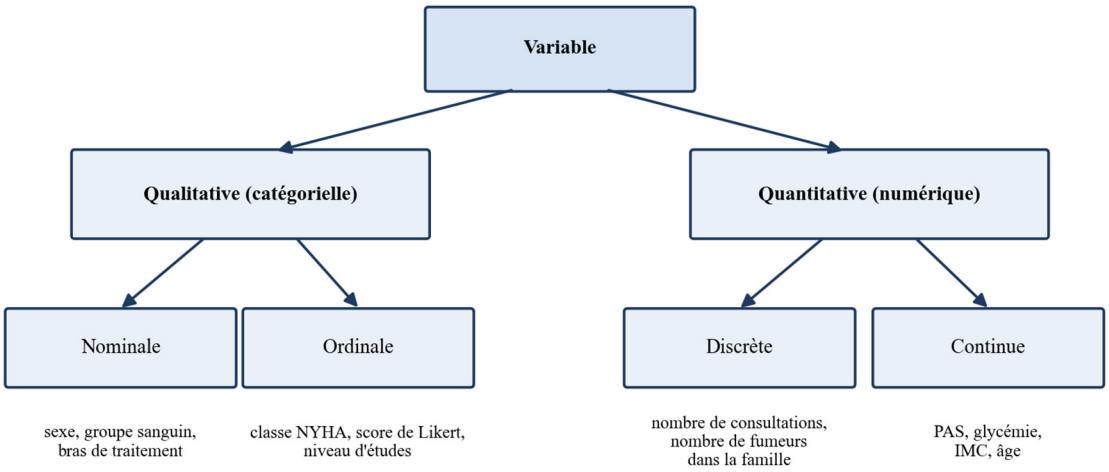


Figure 2.1. Classification des variables statistiques. Les variables qualitatives peuvent être nominales (catégories sans ordre) ou ordinales (catégories ordonnées) ; les variables quantitatives peuvent être discrètes (dénombrements) ou continues (mesures). Les exemples sont tirés de la médecine générale et du jeu de données mg2026.

Les variables nominales ont des catégories entre lesquelles il n'existe pas d'ordre naturel : le sexe (F, M), le groupe sanguin (0, A, B, AB), le bras de traitement (A, B, C), le statut tabagique (non-fumeur, ancien fumeur, fumeur), le milieu de résidence (urbain, rural). Même si nous les codons par des chiffres, par exemple 1 pour les femmes et 2 pour les hommes, ces chiffres ne sont que des étiquettes. La moyenne des codes n'a pas de sens. Une variable nominale à deux catégories seulement est dite dichotomique (ou binaire) : diabète confirmé oui/non, décédé oui/non, perte de poids d'au moins 5% oui/non.

Les variables ordinales ont des catégories ordonnées, mais les distances entre elles ne sont pas égales ou ne sont pas connues. Les stades NYHA de l'insuffisance cardiaque, les stades de fibrose hépatique F0-F4, les grades de stéatose S0-S3, le niveau d'instruction, les échelles de Likert ("pas du tout d'accord"... "tout à fait d'accord"), le score de douleur de 0 à 10 sont ordinaux. Nous savons que le stade NYHA III est plus sévère que le stade II, mais nous ne pouvons pas dire que la différence entre I et II est égale à la différence entre III et IV. Dans le jeu de données mg2026, le niveau d'instruction (1-4), l'activité physique autoévaluée (1-5) et la satisfaction (1-5) sont ordinaux.

Les variables quantitatives discrètes prennent des valeurs isolées, généralement des nombres entiers issus d'un dénombrement : le nombre de consultations par an, le nombre de médicaments par ordonnance, le nombre de grossesses, le nombre de dents cariées. Les variables quantitatives continues peuvent prendre, au moins en théorie, n'importe quelle valeur d'un intervalle, limitée seulement par la précision de l'instrument : la pression artérielle, la glycémie, le poids, la taille, l'âge, la concentration d'un médicament. Le fait qu'une variable continue soit arrondie à l'entier (PAS de 131 mmHg, âge de 56 ans) ne la rend pas discrète.

2.1.2 Les échelles de mesure

Une classification plus fine, très utilisée dans les manuels de méthodologie, distingue quatre échelles de mesure, par ordre croissant de l'information qu'elles contiennent [19,53].

L'échelle nominale permet seulement de vérifier l'égalité : deux patients ont ou n'ont pas le même groupe sanguin. L'échelle ordinale permet en plus d'ordonner : un patient a un stade plus avancé qu'un autre. L'échelle d'intervalle permet en plus de comparer les différences : la différence entre 36 °C et 38 °C est égale à la différence entre 38 °C et 40 °C. Son zéro est toutefois conventionnel, de sorte que les rapports n'ont pas de sens ; une température de 40 °C

n'est pas "deux fois plus élevée" que 20 °C. L'échelle de rapport (ou de proportion) possède un zéro absolu, qui signifie l'absence de la grandeur mesurée, et autorise toutes les opérations, y compris les rapports : une glycémie de 200 mg/dL est le double d'une glycémie de 100 mg/dL, et un patient de 90 kg pèse 1,5 fois plus qu'un patient de 60 kg.

La plupart des variables quantitatives en médecine relèvent d'une échelle de rapport : le poids, la taille, la glycémie, le cholestérol, la pression artérielle, la fréquence cardiaque, la durée d'hospitalisation. La distinction entre intervalle et rapport compte rarement en pratique pour le choix du test, mais elle a une conséquence que vous rencontrerez à la section 2.4 : le coefficient de variation n'a de sens que pour les variables mesurées sur une échelle de rapport.

Tableau 2.1. Les échelles de mesure et les opérations autorisées.

Échelle	Opérations qui ont un sens	Exemples	Tendance centrale adaptée
Nominale	=, ≠	sexe, groupe sanguin, bras de traitement, tabagisme	mode, proportions
Ordinale	=, ≠, <, >	stade NYHA, niveau d'instruction, échelle de Likert, score de douleur	médiane, mode
D'intervalle	en plus, différences	température en °C, année de naissance	moyenne, médiane
De rapport	en plus, rapports	glycémie, poids, PAS, âge, durée d'hospitalisation	moyenne, médiane, moyenne géométrique

2.1.3 Variables dérivées et scores

De nombreuses variables de l'analyse ne sont pas mesurées directement, mais calculées à partir d'autres variables. L'indice de masse corporelle en est l'exemple type :

IMC = poids (kg) / [taille (m)]²

Pour le patient P003 de mg2026, qui pèse 127,4 kg et mesure 183 cm (1,83 m), IMC = $\frac{127{,}4}{1{,}83^2} = \frac{127{,}4}{3{,}3489} = \frac{38{,}0\ \text{kg}}{m^2}$. D'autres variables dérivées fréquentes sont l'âge (calculé à partir de la date de naissance et de la date d'inclusion), le débit de filtration glomérulaire estimé (à partir de la créatinine, de l'âge et du sexe), la variation d'une variable dans le temps (glicemie0 - glicemie6) ou la perte de poids en pourcentage.

Les variables dérivées se calculent dans la feuille de calcul à l'aide d'une formule, elles ne se saisissent pas à la main. Si une taille est mal saisie, la formule produira un IMC absurde, facile à repérer ; si l'IMC était tapé séparément, l'erreur resterait cachée. Lors de la séance de travaux pratiques 1, vous verrez qu'une taille de 1,72 saisie dans une colonne en centimètres donne, lorsqu'on recalcule l'IMC avec la formule, une valeur de plus de 260 000 kg/m², qui saute immédiatement aux yeux. Inversement, une variable dérivée conservée dans le fichier peut aider à récupérer une valeur erronée : à partir de l'IMC et du poids, on peut recalculer la taille.

Les scores cliniques (FINDRISC, score d'Apgar, score de Glasgow, scores de risque cardiovasculaire) sont des sommes de points attribués pour différentes caractéristiques. Formellement, ce sont des variables discrètes, avec un grand nombre de valeurs possibles ; en pratique, ils sont souvent traités comme des variables numériques, mais avec prudence, car un point de plus n'a pas la même signification clinique dans toutes les zones de l'échelle. Le score FINDRISC de mg2026 prend des valeurs entre 0 et 25, avec une moyenne de 10,37 et une médiane de 10.

2.1.4 La dichotomisation et son coût

Les médecins aiment les seuils : hypertendu ou non (140 mmHg), obèse ou non (IMC 30 kg/m²), diabétique ou non (glycémie à jeun 126 mg/dL). La transformation d'une variable continue en variable dichotomique s'appelle la dichotomisation. Elle est utile pour les décisions cliniques et pour la communication avec le patient, mais elle a un coût dans l'analyse statistique.

Le premier coût est la perte d'information. Après dichotomisation, un patient dont l'IMC est de 30,0 et un autre dont l'IMC est de 48,3 sont tout aussi "obèses", alors qu'un patient dont l'IMC est de 29,9 est "non obèse", bien qu'il ne diffère du premier que de 0,1 kg/m². Dans le jeu de données mg2026, la pression systolique moyenne est de 136,87 mmHg chez les 54 patients obèses et de 128,21 mmHg chez les 106 non obèses, mais cette comparaison ignore le fait que, parmi les patients obèses, la PAS continue d'augmenter avec l'IMC. La régression du chapitre 9 montrera que chaque kg/m² supplémentaire est associé, en moyenne, à 1,435 mmHg de plus, information que la dichotomisation perd. Le deuxième coût est la baisse de la puissance statistique : pour détecter la même association avec une variable dichotomisée, il faut davantage de patients. Le troisième est l'arbitraire du seuil : si le seuil n'est pas fixé à l'avance, le chercheur peut être tenté de choisir celui qui donne le résultat "le plus joli".

Un bon exemple d'analyse qui conserve les deux perspectives vient de l'étude prospective portant sur 921 étudiants [5]. La variation de l'IMC a d'abord été analysée comme une variable continue (moyenne -0,45 ± 1,07 kg/m²), par régression linéaire ; puis, séparément, une prise de poids cliniquement pertinente a été définie, survenue chez 7,2% de la cohorte, et analysée par régression logistique. La première analyse utilise toute l'information ; la seconde répond à une question clinique ("qui prend nettement du poids ?"). Conclusion pratique : conservez la variable continue pour l'analyse principale et utilisez les seuils cliniques consacrés, fixés à l'avance, pour les analyses secondaires.

2.1.5 Le codage et le tableau de données

Les données d'une étude s'organisent dans un tableau rectangulaire, selon une règle simple : une ligne par unité statistique (patient) et une colonne par variable. La première ligne contient les noms des variables. Chaque cellule contient une seule valeur. Tous les logiciels statistiques (JASP, MedCalc, Epi Info) supposent cette structure, parfois appelée format "long" ou "ordonné" (angl. tidy data).

Quelques règles pratiques préviennent la plupart des problèmes. Les noms des variables sont courts, sans espaces, sans signes diacritiques et sans caractères spéciaux (tas0, et non "PA systolique initiale (mmHg)") ; la description complète figure dans le dictionnaire des données. Une colonne contient un seul type de données : dans la colonne de la glycémie, on n'écrit que des nombres, sans unité ("87", et non "87 mg/dl"). Les catégories s'écrivent toujours de la même façon ("F", et non tantôt "F", tantôt "f" ou "Feminin"). Les valeurs manquantes sont laissées vides ou marquées par un code unique, déclaré dans le dictionnaire ; on n'écrit pas 0 pour une valeur manquante, car 0 est une valeur réelle pour de nombreuses variables. Les mesures répétées (la glycémie à 0, 3 et 6 mois) sont placées dans des colonnes séparées, dont le nom indique le moment de la mesure.

Le tableau 2.2 présente un extrait du dictionnaire du jeu de données mg2026. Le dictionnaire complet se trouve à l'annexe C.

Tableau 2.2. Extrait du dictionnaire des données du jeu de données mg2026.

Nom	Description	Type	Unité / codes	Intervalle plausible
id	code du patient	texte	P001-P160	unique
sex	sexe	nominal	F, M	-
varsta	âge à l'inclusion	numérique entier	années	18-90

Nom	Description	Type	Unité / codes	Intervalle plausible
educatie	niveau d'instruction	ordinal	1 collège ; 2 lycée ou école professionnelle ; 3 études universitaires ; 4 études postuniversitaires	1-4
inaltime	taille	numérique	cm	130-210
tas0	pression systolique à l'inclusion	numérique entier	mmHg	70-260
glicemie0	glycémie à jeun à l'inclusion	numérique	mg/dL	40-400
hba1c0	hémoglobine glyquée	numérique, une décimale	%	3,5-15
diabet_ogtt	diabète confirmé par HGPO	dichotomique	0 non, 1 oui	0, 1
satisfactie	satisfaction à 6 mois	ordinal	1 très insatisfait... 5 très satisfait	1-5

2.1.6 Les erreurs de saisie fréquentes

Le fichier mg2026_brut.csv contient volontairement quelques-unes des erreurs que vous rencontrerez dans toute base de données réelle. Nous les présentons ici par type, et lors de la séance de travaux pratiques 1 vous les retrouverez et les corrigerez.

Tableau 2.3. Types d'erreurs de saisie, avec des exemples tirés de mg2026_brut.csv.

Type d'erreur	Exemple tiré du fichier brut	Comment la détecter
Codage incohérent	sexe "f" et "Feminin" au lieu de "F" ; milieu "Urban" au lieu de "urban"	tableau de fréquences de la variable (des catégories supplémentaires apparaissent)
Espaces invisibles	"Fumator " (avec une espace finale)	fréquences ; fonction TRIM
Chiffre en trop	PAS 1540 mmHg	minimum et maximum ; intervalle plausible
Unité ou décimale erronée	taille 1,72 (mètres) dans une colonne en cm ; poids 7,26 kg au lieu de 72,6	minimum et maximum ; variable dérivée IMC
Texte dans une colonne numérique	glycémie "87 mg/dl" ; HbA1c "5,2" écrite avec une virgule dans un fichier à point décimal, donc lue comme du texte	nombre de valeurs numériques (n) inférieur au nombre de lignes
Valeurs manquantes	âge manquant ; glycémie à 6 mois manquante	COUNTBLANK, n indiqué par le logiciel
Ligne en double	le même patient saisi deux fois	comptage des codes uniques ; tri selon id

L'ordre de travail recommandé est le suivant : d'abord les fréquences des variables qualitatives (les catégories supplémentaires apparaissent immédiatement), puis le minimum, le maximum et le nombre de valeurs des variables numériques, enfin la vérification des codes uniques. Une demi-heure de nettoyage épargne des jours d'analyse à refaire.

> **Étude de cas : une même variable, trois façons de la rapporter**
>
> La manière dont un article rapporte une variable indique comment les auteurs ont compris son type. Dans un sondage auprès de 121 étudiants en médecine, les notes attribuées aux cours, sur une échelle de 1 à 10, ont été rapportées sous forme de médiane et d'écart interquartile : 9 (7-10) pour la médecine d'urgence, 7 (5-9) pour l'informatique biomédicale et 8 (5-9) pour la biostatistique [16]. Ce choix est correct : une note donnée par un étudiant est ordinale, et une médiane dit "la moitié des étudiants ont donné au moins 8", sans supposer que la distance entre 7 et 8 est égale à celle entre 9 et 10. Dans une étude sur la prescription chez les personnes

> âgées, le nombre de médicaments par ordonnance, une variable discrète, a été rapporté sous forme de moyenne ± écart-type, 4,7 ± 1,51 [38] ; pour un dénombrement comportant de nombreuses valeurs possibles et une distribution relativement symétrique, la moyenne est acceptable et facile à interpréter. Dans une étude sur l'acceptabilité de la technologie mobile chez 103 patients diabétiques, l'âge a été rapporté sous forme de médiane avec l'étendue : 37 ans, avec des extrêmes de 18 et 65 ans [54]. L'étendue montre l'éventail d'âge des participants, mais elle ne dépend que de deux individus extrêmes ; l'écart interquartile aurait mieux décrit où se situe la majorité des patients. Aucun de ces choix n'est "faux", mais chacun transmet une information différente, et le lecteur doit savoir laquelle.

2.2 Les tableaux de fréquences

Le premier résumé d'une variable est le tableau de fréquences : la liste des valeurs (ou des classes de valeurs) accompagnée du nombre d'occurrences de chacune. L'outil paraît modeste, mais c'est le moyen le plus sûr de voir à quoi ressemblent les données, de découvrir les erreurs de codage et de préparer les graphiques.

2.2.1 Fréquences absolues, relatives et cumulées

La fréquence absolue (ni) d'une valeur est le nombre d'observations égales à cette valeur. La somme des fréquences absolues est l'effectif de l'échantillon, n. La fréquence relative (fi) est la fréquence absolue rapportée à n, exprimée sous forme de fraction ou de pourcentage :

$$f_i = \frac{n_i}{n} \text{ (ou } f_i\% = \frac{n_i}{n} \times 100)$$

La fréquence cumulée (absolue ou relative) d'une valeur est la somme des fréquences de toutes les valeurs inférieures ou égales à celle-ci. Elle n'a de sens que si les valeurs sont ordonnées, donc pour les variables ordinales et numériques, et non pour les variables nominales. La fréquence relative cumulée de la dernière classe est toujours de 100% (ou de 99,9-100,1%, en raison des arrondis).

Le tableau 2.4 présente la satisfaction à l'égard du programme à 6 mois, une variable ordinale mesurée sur une échelle de Likert allant de 1 (très insatisfait) à 5 (très satisfait).

Tableau 2.4. Satisfaction à l'égard du programme à 6 mois, jeu de données mg2026 (n = 160).

Satisfaction	Fréquence absolue	Fréquence relative (%)	Fréquence absolue cumulée	Fréquence relative cumulée (%)
1 très insatisfait	4	2,5	4	2,5
2 insatisfait	19	11,9	23	14,4
3 neutre	54	33,8	77	48,1
4 satisfait	55	34,4	132	82,5
5 très satisfait	28	17,5	160	100,0
Total	160	100,0		

Ce tableau se lit de plusieurs façons. La réponse la plus fréquente est 4 (55 patients, 34,4%), presque à égalité avec 3 (54). Seuls 14,4% des patients ont été insatisfaits ou très insatisfaits (fréquence cumulée à la valeur 2), et 51,9% ont été satisfaits ou très satisfaits (34,4 + 17,5). La fréquence cumulée dépasse 50% à la valeur 4, donc la médiane de la satisfaction est 4 : au moins la moitié des patients ont donné la note 4 ou 5.

Pour les variables nominales, le tableau se limite aux fréquences absolues et relatives. Dans mg2026, le statut tabagique se répartit ainsi : non-fumeurs 74 (46,3%), anciens fumeurs 33 (20,6%), fumeurs 53 (33,1%). Le milieu de résidence : urbain 95 (59,4%), rural 65 (40,6%). Le sexe : femmes 86 (53,8%), hommes 74 (46,3%) ; la somme des pourcentages arrondis est de 100,1%, ce qui est normal et peut être signalé dans une note sous le tableau.

2.2.2 Le regroupement des données continues en classes

Une variable continue comme la pression systolique présente, dans mg2026, des dizaines de valeurs distinctes, de 102 à 169 mmHg. Un tableau reprenant chaque valeur compterait plus de 50 lignes, la plupart avec une fréquence de 1 ou 2. C'est pourquoi les valeurs sont regroupées en classes (intervalles) de même largeur.

Le nombre de classes est choisi de façon que le tableau soit lisible et conserve la forme de la distribution. Une règle indicative est la règle de Sturges :

$$k = 1 + 3{,}322 \times \log_{10}(n)$$

Pour n = 160, log10(160) = 2,204, donc k = 1 + 3,322 x 2,204 = 8,32, soit environ 8 classes. La largeur de classe (le pas) s'obtient en divisant l'étendue par le nombre de classes :

$$h = \frac{x_{\max} - x_{\min}}{k}$$

Pour la PAS, l'étendue est de 169 - 102 = 67 mmHg, et $h = \frac{67}{8} = 8{,}4$ mmHg. En pratique, la largeur est arrondie à une valeur commode pour le lecteur, ici 10 mmHg, et les bornes des classes sont choisies à des valeurs rondes. Avec des classes de 10 mmHg commençant à 100, on obtient 7 classes, ce qui est parfaitement acceptable : la règle de Sturges est un point de départ, non une loi.

Les bornes des classes doivent être sans ambiguïté : chaque valeur doit appartenir à une seule classe. Pour des valeurs entières, les classes s'écrivent 100-109, 110-119, et ainsi de suite. Pour des valeurs décimales, on utilise des intervalles de la forme [100 ; 110), où la parenthèse fermante indique que 110 n'appartient pas à la classe, ou des bornes comportant une décimale de plus que les données (99,5-109,5), comme dans les tableaux des recueils classiques de problèmes. Le milieu (centre) de la classe est la moyenne des bornes réelles : pour la classe 100-109, dont les bornes réelles sont 99,5 et 109,5, le centre est 104,5.

Tableau 2.5. Pression artérielle systolique à l'inclusion, regroupée en classes de 10 mmHg, jeu de données mg2026 (n = 160).

PAS (mmHg)	Centre de classe	ni	fi (%)	Fréquence cumulée	Fréquence cumulée (%)
100-109	104,5	11	6,9	11	6,9
110-119	114,5	21	13,1	32	20,0
120-129	124,5	35	21,9	67	41,9
130-139	134,5	52	32,5	119	74,4
140-149	144,5	25	15,6	144	90,0
150-159	154,5	13	8,1	157	98,1
160-169	164,5	3	1,9	160	100,0
Total		160	100,0		

Le tableau permet de lire rapidement des informations utiles en clinique. La classe la plus fréquente (la classe modale) est 130-139 mmHg. La proportion de patients dont la PAS est inférieure à 140 mmHg est la fréquence cumulée de la classe 130-139, soit 74,4% ; par conséquent, 25,6% ont une PAS ≥ 140 mmHg, exactement la proportion calculée à partir des données individuelles (41 patients). La proportion de patients dont la PAS est d'au moins 150 mmHg est de 8,1 + 1,9 = 10,0%, soit 16 patients.

Le regroupement a un prix : on perd la valeur exacte de chaque observation. Si nous ne disposons que du tableau, la moyenne peut être estimée en supposant que toutes les valeurs d'une classe sont égales à son centre :

$$m \approx \frac{\sum(n_i \times c_i)}{n}$$

où ci est le centre de la classe i. Pour le tableau 2.5 :

$\frac{11\times104{,}5+21\times114{,}5+35\times124{,}5+52\times134{,}5+25\times144{,}5+13\times154{,}5+3\times164{,}5}{160} = \frac{21\,020}{160} = 131{,}38$ mmHg. La moyenne calculée à partir des données individuelles est de 131,13 mmHg ; l'erreur due au regroupement est faible, d'un quart de mmHg.

2.2.3 Les tableaux de fréquences comportant des valeurs manquantes

Un type classique de problème, présent dans tous les recueils de la discipline et à l'examen, fournit un tableau de fréquences dont certaines cellules manquent et demande de les reconstituer. Aucune formule nouvelle n'est nécessaire, seulement les trois relations ci-dessus, utilisées dans le bon ordre :

- fréquence absolue = fréquence relative x n (arrondie à l'entier, puisque nous comptons des patients) ;
- fréquence relative = fréquence absolue / n x 100 ;
- fréquence cumulée d'une classe = fréquence cumulée de la classe précédente + fréquence de la classe ; inversement, fréquence d'une classe = différence entre deux fréquences cumulées consécutives ;
- somme des fréquences absolues = n ; somme des fréquences relatives = 100%.

La stratégie consiste à commencer par les cellules que l'on peut calculer directement, puis à s'en servir pour les autres, et à vérifier à la fin que les totaux tombent juste. Les problèmes résolus 2.1 et 2.2 présentent deux tableaux de ce type entièrement résolus, et la séance de travaux pratiques 2 en contient deux autres.

2.3 Les indicateurs de tendance centrale

Les tableaux de fréquences montrent l'ensemble de la distribution. Nous avons pourtant souvent besoin d'une seule valeur qui résume le "centre" des données, la valeur typique. Il existe plusieurs indicateurs de ce genre, et le choix entre eux dépend du type de la variable et de la forme de la distribution.

2.3.1 La moyenne arithmétique

La moyenne arithmétique est la somme des valeurs divisée par leur nombre :

$$m = \frac{x_1 + x_2 + \cdots + \text{xn}}{n} = \frac{\sum x_i}{n}$$

C'est l'indicateur le plus connu et le plus utilisé ; il possède d'excellentes propriétés mathématiques et sert de base à la plupart des tests statistiques. Pour la pression systolique de mg2026, m = 131,13 mmHg. La moyenne a une interprétation physique simple : c'est le "centre de gravité" des données, le point où la somme des écarts positifs est égale à la somme des écarts négatifs, $\sum(x_i - m) = 0$.

La moyenne a aussi un défaut important : elle est sensible aux valeurs extrêmes. Une seule valeur très élevée la tire vers le haut, une seule valeur très basse la tire vers le bas. C'est pourquoi la moyenne est l'indicateur adapté aux variables numériques de distribution approximativement symétrique, mais elle peut être trompeuse pour les distributions asymétriques.

La moyenne pondérée s'utilise lorsque les valeurs ont des importances (des poids) différentes, par exemple lorsque nous combinons des moyennes calculées sur des groupes de tailles différentes :

$$\text{mp} = \frac{\sum(w_i \times x_i)}{\sum w_i}$$

La moyenne de la PAS dans les trois bras de l'étude mg2026 est de 131,53 mmHg (n = 53), 132,79 mmHg (n = 53) et 129,11 mmHg (n = 54). La moyenne globale s'obtient en pondérant par la taille des groupes : $\frac{53\times131{,}53+53\times132{,}79+54\times129{,}11}{160} = \frac{6\,971{,}09+7\,037{,}87+6\,971{,}94}{160} = \frac{20\,980{,}90}{160} =$ 131,13 mmHg, c'est-à-dire exactement la moyenne des 160 patients. Ici, les groupes ont des tailles presque égales, et la moyenne simple des trois moyennes (131,14) en diffère très peu. La différence devient importante lorsque les groupes sont très inégaux, comme dans le cas des trois cabinets du problème résolu 1.5. L'estimation de la moyenne à partir du tableau de fréquences, présentée à la section 2.2.2, est elle aussi une moyenne pondérée, avec les poids ni.

2.3.2 La médiane

La médiane est la valeur qui partage la série ordonnée des observations en deux moitiés d'effectif égal : la moitié des observations lui sont inférieures ou égales, l'autre moitié supérieures ou égales. Pour la calculer, on range les valeurs par ordre croissant. Si n est impair, la médiane est la valeur de rang (n + 1)/2. Si n est pair, la médiane est la moyenne des deux valeurs centrales, de rangs n/2 et n/2 + 1.

Pour les âges des 20 premiers patients de mg2026, rangés par ordre croissant (37, 40, 45, 47, 50, 52, 53, 54, 55, 56, 56, 57, 57, 60, 62, 62, 65, 66, 69, 70), n = 20 est pair, et les valeurs de rangs 10 et 11 sont 56 et 56 ; la médiane est de 56 ans. Pour l'ensemble du jeu de données, la médiane de l'âge est également de 56 ans, et la médiane de la PAS de 131 mmHg.

La médiane est robuste : elle n'est pas influencée par les valeurs extrêmes, mais seulement par leur position. Si le plus âgé des 20 patients avait 99 ans au lieu de 70, la médiane resterait de 56. C'est pourquoi la médiane est l'indicateur privilégié pour les variables asymétriques (triglycérides, durée d'hospitalisation, valeurs de certains marqueurs biologiques, délais d'attente) et pour les variables ordinales, pour lesquelles la moyenne n'a pas d'interprétation correcte.

2.3.3 Le mode

Le mode (la valeur modale), que l'on rencontre aussi sous le nom de valeur dominante, est la valeur dont la fréquence est la plus élevée. C'est le seul indicateur de tendance centrale utilisable pour les variables nominales : le mode du statut tabagique dans mg2026 est "non-fumeur" (74 patients). Pour les variables numériques, le mode est moins utile, car il dépend beaucoup du hasard et de l'arrondi. Dans mg2026, la valeur de PAS la plus fréquente est 138 mmHg (10 patients), l'âge le plus fréquent est 51 ans (10 patients) et le score FINDRISC le plus fréquent est 10 (23 patients).

Une distribution peut avoir plusieurs modes. Une distribution bimodale, à deux pics, suggère souvent que l'échantillon mélange deux populations différentes, par exemple la taille dans un groupe composé de femmes et d'hommes (dans mg2026, la taille moyenne est de 162,6 cm chez les femmes et de 175,97 cm chez les hommes). Dans un tel cas, la moyenne globale décrit un "patient moyen" qui n'existe pas, et une description séparée par sous-groupes est plus informative.

2.3.4 La moyenne géométrique

La moyenne géométrique de n valeurs positives est la racine n-ième de leur produit. Elle se calcule plus commodément en passant aux logarithmes : on calcule la moyenne des logarithmes, puis on revient à l'échelle initiale au moyen de la fonction exponentielle :

$$\mathrm{mg} = \exp\left[\frac{\ln x_1 + \ln x_2 + \cdots + \ln \mathrm{xn}}{n}\right]$$

La moyenne géométrique convient aux variables positives à asymétrie droite, dont la distribution devient approximativement symétrique après transformation logarithmique : les concentrations de certains marqueurs biologiques, les titres d'anticorps, la charge virale, les

triglycérides. Dans mg2026, la moyenne arithmétique des triglycérides est de 161,18 mg/dL, la médiane de 151,5 mg/dL et la moyenne des logarithmes naturels de 4,97 ; la moyenne géométrique est exp (4,97) ≈ 144,0 mg/dL. La moyenne géométrique est inférieure à la moyenne arithmétique (c'est toujours le cas, sauf lorsque toutes les valeurs sont égales) et proche de la médiane, car elle n'est pas tirée vers le haut par les quelques valeurs très élevées (le maximum est de 442 mg/dL). Le chapitre 3 reviendra sur la transformation logarithmique, et le chapitre 4 montrera comment on calcule un intervalle de confiance pour la moyenne géométrique.

Un exemple simple montre la logique. Pour trois titres d'anticorps de 1/10, 1/100 et 1/1 000, la moyenne arithmétique des dénominateurs est de 370, une valeur qui ne ressemble à aucune des trois. La moyenne géométrique est la racine cubique de 10 x 100 x 1 000 = 1 000 000, soit 100, la valeur centrale sur l'échelle logarithmique selon laquelle raisonne en réalité le laboratoire.

2.3.5 Le choix de l'indicateur et l'effet des valeurs extrêmes

La relation entre la moyenne et la médiane renseigne sur la forme de la distribution. Dans une distribution symétrique, la moyenne et la médiane sont pratiquement égales : pour la PAS, 131,13 et 131 mmHg ; pour l'âge, 56,29 et 56 ans. Dans une distribution à asymétrie positive (longue queue à droite), la moyenne est supérieure à la médiane. Pour l'IMC, la moyenne est de 29,04 kg/m² et la médiane de 28,25 kg/m². La figure 2.2 montre pourquoi : quelques patients atteints d'obésité sévère, jusqu'à 48,3 kg/m², tirent la moyenne vers la droite, mais ne déplacent pas la médiane.

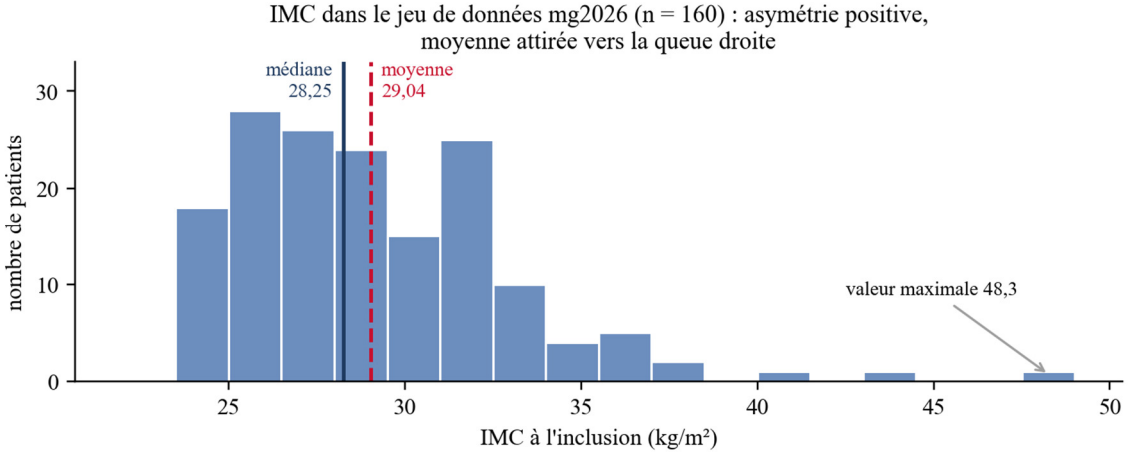


Figure 2.2. Distribution de l'IMC à l'inclusion dans le jeu de données mg2026 (n = 160). L'histogramme présente une queue à droite, et la moyenne (29,04 kg/m², ligne rouge discontinue) est supérieure à la médiane (28,25 kg/m², ligne bleue). La valeur maximale, 48,3 kg/m², influence la moyenne, mais pas la médiane. Données synthétiques (jeu de données mg2026).

L'expérience est facile à faire. Si nous retirons le patient dont l'IMC est de 48,3 kg/m², la moyenne des 159 patients restants passe de 29,04 à 28,92 kg/m², et l'écart-type de 3,91 à 3,61 kg/m² ; la médiane passe de 28,25 à 28,2 kg/m², une différence négligeable. Une seule observation sur 160 a modifié la moyenne de 0,12 kg/m² et l'écart-type de près de 8%. Sur un échantillon de 20 patients, ce même patient aurait modifié la moyenne de plus de 1 kg/m².

Cela ne signifie pas que le patient dont l'IMC est de 48,3 kg/m² doit être exclu. La valeur est biologiquement plausible (obésité de grade III), elle a été vérifiée, et le patient fait partie de la population étudiée. Les valeurs extrêmes ne sont exclues que s'il s'agit d'erreurs démontrées ; sinon, on choisit des méthodes qui les tolèrent (la médiane, l'écart interquartile, les tests non paramétriques) ou l'on rapporte l'analyse avec et sans elles.

Tableau 2.6. Le choix de l'indicateur de tendance centrale.

Type de données	Indicateur recommandé	Remarques
Nominales	mode ; proportions de chaque catégorie	la moyenne des codes n'a pas de sens
Ordinales	médiane (et mode)	la moyenne est parfois rapportée pour des scores comportant de nombreuses valeurs, avec prudence
Numériques, approximativement symétriques	moyenne (et médiane)	PAS, âge, cholestérol, taille dans mg2026
Numériques, asymétriques	médiane ; moyenne géométrique pour des données log-normales	triglycérides, IMC, durée d'hospitalisation, marqueurs inflammatoires
Données comportant des valeurs extrêmes plausibles	médiane	on rapporte aussi leur influence sur la moyenne

2.4 Les indicateurs de dispersion

Deux groupes de patients peuvent avoir la même moyenne de pression systolique, 131 mmHg, et être pourtant très différents : dans l'un, tous les patients se situent entre 125 et 137 mmHg, dans l'autre, les valeurs s'étendent de 100 à 165 mmHg. La tendance centrale ne dit rien de cet étalement. Pour le décrire, nous avons besoin des indicateurs de dispersion. Pour les calculer à la main, nous utiliserons un petit exemple.

Exemple de travail. Lors d'un contrôle, huit patients ont les glycémies à jeun suivantes (mg/dL) : 88, 94, 97, 102, 105, 110, 116, 128. Leur somme est de 840, donc la moyenne est $m = \frac{840}{8} = 105$ mg/dL. Les valeurs sont déjà ordonnées ; la médiane est la moyenne des valeurs de rangs 4 et 5, (102 + 105)/2 = 103,5 mg/dL.

2.4.1 L'étendue

L'étendue (angl. range) est la différence entre la valeur maximale et la valeur minimale : 128 - 88 = 40 mg/dL dans notre exemple. Pour la PAS dans mg2026, l'étendue est de 169 - 102 = 67 mmHg. Elle est simple et intuitive, mais ne dépend que de deux observations, les valeurs extrêmes, et augmente presque toujours avec l'effectif de l'échantillon : plus nous étudions de patients, plus il est probable de rencontrer des valeurs inhabituelles. Dans les articles, elle est habituellement rapportée sous la forme "minimum-maximum", et non comme une différence.

2.4.2 L'écart moyen absolu

L'écart d'une observation à la moyenne est xi - m. Dans notre exemple, les écarts sont -17, -11, -8, -3, 0, 5, 11, 23. Leur somme est nulle, comme toujours, de sorte que la moyenne des écarts ne peut pas mesurer la dispersion. Une solution consiste à prendre les valeurs absolues. L'écart moyen absolu est la moyenne des distances à la moyenne :

$$\text{EMA} = \frac{\sum|x_i - m|}{n} = \frac{17 + 11 + 8 + 3 + 0 + 5 + 11 + 23}{8} = \frac{78}{8} = 9{,}75 \text{ mg/dL}$$

Les patients se situent, en moyenne, à près de 10 mg/dL de la moyenne du groupe. Cet indicateur est facile à comprendre, mais ses propriétés mathématiques sont peu commodes, si bien qu'en statistique inférentielle on l'utilise beaucoup plus rarement que l'écart-type. Pour la PAS dans mg2026, l'écart moyen absolu est de 10,91 mmHg.

2.4.3 La variance et l'écart-type

La seconde solution au problème des signes consiste à élever les écarts au carré. La variance de l'échantillon est la somme des carrés des écarts divisée par n - 1 :

$$s^2 = \frac{\sum(x_i - m)^2}{n - 1}$$

L'écart-type (angl. standard deviation, noté s ou ÉT) est la racine carrée de la variance et s'exprime dans la même unité de mesure que les données :

$$s = \sqrt{s^2}$$

Le tableau 2.7 présente le calcul complet pour les huit glycémies.

Tableau 2.7. Calcul à la main de la variance et de l'écart-type pour huit glycémies.

x_i (mg/dL)	**xi - m**	$(x_i - m)^2$
88	-17	289
94	-11	121
97	-8	64
102	-3	9
105	0	0
110	5	25
116	11	121
128	23	529
Σ = 840	Σ = 0	Σ = 1 158

$$s^2 = \frac{1\,158}{8 - 1} = \frac{1\,158}{7} = 165{,}43 \text{ (mg/dL)}^2$$

$$s = \sqrt{165{,}43} = 12{,}86 \text{ mg/dL}$$

Pourquoi n - 1 et non **n** ? Les écarts sont calculés par rapport à la moyenne de l'échantillon, et non par rapport à la moyenne inconnue de la population, et la moyenne de l'échantillon est, par construction, le point par rapport auquel la somme des carrés est la plus petite. Diviser par n sous-estimerait systématiquement la variance de la population. Diviser par n - 1 corrige ce biais ; le nombre n - 1 s'appelle le nombre de degrés de liberté, car, une fois la moyenne connue, seuls n - 1 écarts peuvent varier librement (le dernier est déterminé par la condition que la somme soit nulle). Dans Google Sheets, la fonction STDEV (ou STDEV.S) divise par n - 1, et STDEVP (STDEV.P) divise par n ; pour les données d'un échantillon, on utilise toujours la première. Pour n grand, la différence est négligeable.

L'écart-type est l'indicateur de dispersion le plus utilisé. Il a une interprétation très utile pour les distributions approximativement normales, que nous détaillerons au chapitre 3 : environ 68% des valeurs se trouvent dans l'intervalle $m \pm s$ et environ 95% dans l'intervalle $m \pm 2s$. Pour la PAS de mg2026, s = 13,88 mmHg, et l'intervalle $m \pm s$ est de 117,3-145,0 mmHg. La variance, en revanche, s'exprime dans l'unité de mesure au carré (mmHg², (mg/dL)²), difficile à interpréter cliniquement ; pour la PAS, $s^2 = 192{,}66$ mmHg^2. Elle est cependant indispensable dans les calculs, par exemple dans l'analyse de la variance du chapitre 6.

2.4.4 Le coefficient de variation

L'écart-type dépend de l'unité de mesure et de l'ordre de grandeur des données. Un écart-type de 13,88 mmHg pour la PAS et un écart-type de 0,45% pour l'HbA1c ne peuvent pas être comparés directement. Le coefficient de variation exprime l'écart-type en pourcentage de la moyenne :

$$\text{CV} = \frac{s}{m} \times 100$$

Pour les huit glycémies, $\text{CV} = \frac{12{,}86}{105} \times 100 = 12{,}2\%$. Dans mg2026, le CV est de 10,6% pour la PAS, de 15,6% pour l'âge, de 13,5% pour l'IMC, de 8,0% pour l'HbA1c, de 17,6% pour le cholestérol et de 47,3% pour les triglycérides. Rapportés à leur moyenne, les triglycérides sont

près de cinq fois plus variables que la pression systolique. Un CV élevé (au-delà de 30-50%) pour une variable qui ne peut pas être négative est aussi un signal d'asymétrie : si l'écart-type était presque égal à la moyenne, l'intervalle m - 2s contiendrait des valeurs négatives, biologiquement impossibles.

Le coefficient de variation a deux utilisations fréquentes en médecine : comparer la variabilité de grandeurs différentes et évaluer la précision des méthodes de laboratoire (un faible CV analytique indique une méthode reproductible). Il n'a de sens que pour les variables mesurées sur une échelle de rapport, avec un zéro absolu et une moyenne positive. Pour la température en degrés Celsius, par exemple, le CV changerait si l'on passait aux degrés Fahrenheit ; il n'a donc pas d'interprétation.

2.4.5 L'écart-type et l'erreur standard : une confusion fréquente

L'erreur standard de la moyenne (angl. standard error of the mean, ES ou SEM) se calcule à partir de l'écart-type et de l'effectif de l'échantillon :

$$\mathrm{ES} = \frac{s}{\sqrt{n}}$$

Pour les huit glycémies, $\mathrm{ES} = \frac{12{,}86}{\sqrt{8}} = \frac{12{,}86}{2{,}83} = 4{,}55$ mg/dL. Pour la PAS dans mg2026, $\mathrm{ES} = \frac{13{,}88}{\sqrt{160}} = \frac{13{,}88}{12{,}65} = 1{,}097$ mmHg.

Ces deux grandeurs répondent à des questions différentes. L'écart-type décrit à quel point les patients varient les uns par rapport aux autres ; c'est une caractéristique de la population, et il ne diminue pas lorsque nous étudions davantage de patients. L'erreur standard décrit avec quelle précision la moyenne de l'échantillon estime la moyenne de la population ; elle diminue à mesure que n augmente. Si nous étudiions 640 patients au lieu de 160, l'écart-type de la PAS resterait d'environ 14 mmHg, mais l'erreur standard serait divisée par deux, à environ 0,55 mmHg.

Cette confusion a des conséquences pratiques. Certains articles rapportent "moyenne ± ES" dans les tableaux descriptifs, sans le préciser. Pour la PAS, "131,13 ± 13,88" et "131,13 ± 1,10" décrivent les mêmes données, mais la seconde forme suggère à un lecteur inattentif une population extrêmement homogène. La règle de présentation est simple : pour décrire les patients, on utilise l'écart-type ; pour décrire la précision d'une estimation, on utilise l'intervalle de confiance (chapitre 4), qui se construit à partir de l'erreur standard. Tout tableau doit préciser ce que représente le chiffre placé après le signe ±.

2.5 Quantiles, percentiles, quartiles et résumé en cinq valeurs

La médiane est le cas particulier d'une famille plus large d'indicateurs de position, appelés quantiles. Un quantile d'ordre q (compris entre 0 et 1) est la valeur au-dessous de laquelle se trouve une fraction q des observations. Les percentiles partagent la série ordonnée en 100 parties : le 10e percentile (P10) est la valeur au-dessous de laquelle se trouvent 10% des observations, P90 la valeur au-dessous de laquelle s'en trouvent 90%. Les quartiles partagent la série en quatre parties égales : le premier quartile Q1 = P25, le deuxième Q2 = P50 = la médiane, le troisième Q3 = P75.

Les percentiles sont familiers à tout pédiatre, grâce aux courbes de croissance : un enfant "au 90e percentile pour le poids" pèse plus que 90% des enfants de même âge et de même sexe de la population de référence. Dans mg2026, P10 et P90 valent 112 et 148,2 mmHg pour la PAS, et 86 et 126 mg/dL pour la glycémie à jeun. Autrement dit, 10% des patients ont une glycémie d'au moins 126 mg/dL, le seuil diagnostique du diabète, ce qui concorde avec le comptage direct : 18 patients sur 160 (11,3%).

2.5.1 Le calcul des quartiles

Pour de petites séries, les quartiles peuvent se calculer à la main par la méthode des "médianes des moitiés" : on trouve la médiane, puis Q1 est la médiane de la moitié inférieure et Q3 la médiane de la moitié supérieure. Pour les huit glycémies (88, 94, 97, 102 | 105, 110, 116, 128), la moitié inférieure a pour médiane (94 + 97)/2 = 95,5, et la moitié supérieure (110 + 116)/2 = 113. Donc $Q_1 = 95{,}5$ mg/dL et $Q_3 = 113$ mg/dL.

Les logiciels utilisent des formules d'interpolation légèrement différentes. Google Sheets (fonction QUARTILE) et JASP utilisent par défaut une interpolation linéaire qui, pour ces mêmes huit valeurs, donne Q1 = 96,25 et $Q_3 = 111{,}5$ mg/dL. Sur de grands échantillons, les différences deviennent négligeables, et pour les petites séries de l'examen, toute méthode est acceptée, à condition d'être appliquée correctement. Ne vous inquiétez pas si MedCalc et JASP donnent des quartiles légèrement différents pour les mêmes données.

2.5.2 L'écart interquartile et le résumé en cinq valeurs

L'écart interquartile (EIQ, angl. interquartile range, IQR) est la différence Q3 - Q1 et englobe la moitié centrale des observations. C'est l'indicateur de dispersion qui accompagne naturellement la médiane et, comme elle, il est robuste aux valeurs extrêmes. Dans les articles, il est habituellement rapporté sous forme d'intervalle, "médiane (Q1-Q3)", par exemple IMC 28,25 (26,2-31,4) kg/m², et non comme une simple différence.

Le résumé en cinq valeurs (minimum, Q1, médiane, Q3, maximum) décrit une distribution de façon compacte et sert de base à la boîte à moustaches (box-plot). Pour les principales variables de mg2026 :

Tableau 2.8. Résumé en cinq valeurs de quelques variables de mg2026 (n = 160).

Variable	Minimum	Q1	Médiane	Q3	Maximum	EIQ
Âge (ans)	35	50,75	56	62	79	11,25
IMC $\left(\frac{\text{kg}}{m^2}\right)$	23,5	26,2	28,25	31,4	48,3	5,2
PAS (mmHg)	102	122	131	140	169	18
Glycémie à jeun (mg/dL)	75	93	104,5	116	154	23
Triglycérides (mg/dL)	44	107	151,5	203,25	442	96,25

Un résumé en cinq valeurs se lit comme une carte. Pour la PAS, les distances de la médiane à Q1 (9 mmHg) et à Q3 (9 mmHg) sont égales, signe de symétrie. Pour les triglycérides, la distance de la médiane à Q3 (51,75) est supérieure à celle qui la sépare de Q1 (44,5), et le maximum (442) est très éloigné de Q3 : la distribution présente une longue queue à droite.

2.5.3 Les valeurs extrêmes : la règle 1,5 x EIQ

Une règle pratique, proposée par Tukey, considère comme "valeurs extrêmes" (angl. outliers) les observations situées à plus de 1,5 x EIQ au-dessous de Q1 ou au-dessus de Q3. Les limites (parfois appelées "barrières") sont :

limite inférieure = Q1 - 1,5 x EIQ ; limite supérieure = Q3 + 1,5 x EIQ

Pour la PAS : EIQ = 18, donc les limites sont 122 - 27 = 95 et 140 + 27 = 167 mmHg. Deux valeurs (168 et 169 mmHg) dépassent la limite supérieure. Pour l'IMC : EIQ = 5,2, les limites sont 26,2 - 7,8 = 18,4 et 31,4 + 7,8 = 39,2 kg/m² ; trois valeurs dépassent la limite supérieure (40,0 ; 43,4 et 48,3). Pour les triglycérides : la limite supérieure est $203{,}25 + 1{,}5 \times 96{,}25 = 347{,}6$ mg/dL, dépassée par cinq valeurs (364, 364, 373, 405 et 442 mg/dL).

La règle identifie des valeurs qui méritent d'être examinées, et non des valeurs à supprimer. Une pression de 169 mmHg ou une triglycéridémie de 442 mg/dL se rencontrent tous les jours au cabinet. La bonne question est toujours "La valeur est-elle possible et a-t-elle été vérifiée ?", et

non "Fausse-t-elle la moyenne ?". En revanche, une pression de 1540 mmHg, celle du fichier brut, n'est pas une valeur extrême, mais une erreur : elle est biologiquement impossible.

2.6 La forme de la distribution : un premier aperçu de l'asymétrie et de l'aplatissement

Deux indicateurs numériques décrivent la forme d'une distribution et seront étudiés en détail au chapitre 3. Le coefficient d'asymétrie (angl. skewness) est approximativement nul pour une distribution symétrique, positif pour une distribution dont la queue s'étire vers la droite (valeurs élevées rares) et négatif pour une queue vers la gauche. Le coefficient d'aplatissement (excès de kurtosis, angl. excess kurtosis) compare le "poids des queues" à celui de la loi normale : il est nul pour la loi normale, positif pour les distributions à queues lourdes et à pic pointu, négatif pour les distributions plus aplaties.

Dans mg2026, la PAS a une asymétrie de 0,01 et un aplatissement de -0,10, pratiquement les valeurs d'une distribution normale. L'IMC a une asymétrie de 1,42 et un aplatissement de 3,68, donc une queue droite prononcée et quelques valeurs très élevées. Les triglycérides ont une asymétrie de 0,95 et un aplatissement de 1,12 ; après transformation logarithmique, l'asymétrie devient -0,28, proche de la symétrie. La glycémie à jeun est légèrement asymétrique (0,37). En pratique, une asymétrie comprise entre -0,5 et 0,5 est considérée comme faible, et une asymétrie supérieure à 1 en valeur absolue comme importante. Pour les variables biologiques qui ne peuvent pas être négatives et dont la moyenne est faible par rapport à l'écart-type, l'asymétrie positive est la règle, et non l'exception.

Un indice simple, sans aucun calcul : si la moyenne est nettement supérieure à la médiane et que l'écart-type est comparable à la moyenne, la distribution est presque certainement asymétrique à droite.

2.7 Les représentations graphiques

Un bon graphique montre en quelques secondes ce qu'un tableau montre en quelques minutes. Le choix du graphique dépend, comme celui des indicateurs, du type de la variable. La figure 2.3 présente quatre des graphiques les plus utilisés, construits à partir du jeu de données mg2026.

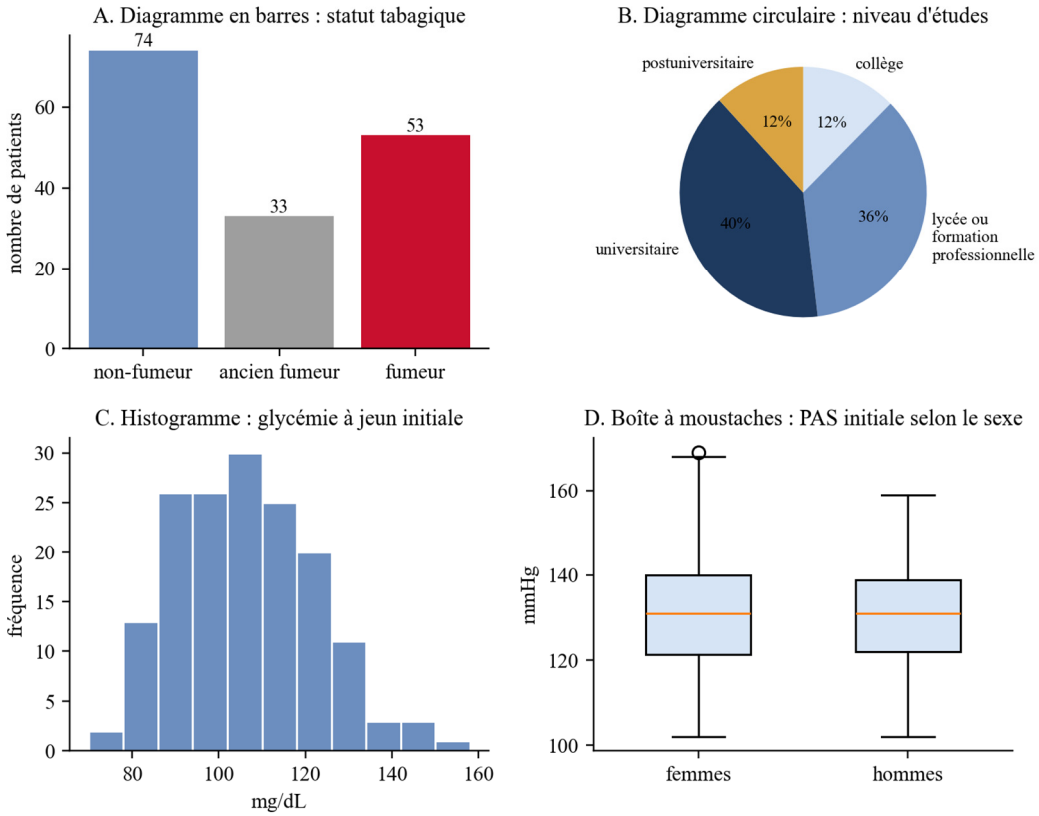


Figure 2.3. Graphiques usuels pour les données biomédicales, jeu de données mg2026 (n = 160). A : diagramme en barres du statut tabagique (variable nominale). B : diagramme circulaire du niveau d'instruction (variable ordinale à quatre catégories). C : histogramme de la glycémie à jeun à l'inclusion (variable continue). D : boîtes à moustaches de la pression systolique chez les femmes et chez les hommes. Données synthétiques.

2.7.1 Graphiques pour les variables qualitatives

Le diagramme en barres (en colonnes) représente chaque catégorie par une barre dont la hauteur est proportionnelle à la fréquence (absolue ou relative). Les barres sont séparées les unes des autres, car les catégories sont distinctes. Pour les variables nominales, l'ordre des barres peut être choisi librement (par exemple, par fréquence décroissante) ; pour les variables ordinales, on conserve l'ordre naturel des catégories. La figure 2.3A montre que les non-fumeurs constituent le groupe le plus nombreux (74), suivis des fumeurs (53) et des anciens fumeurs (33). Les diagrammes en barres groupées ou empilées permettent de comparer une variable entre groupes, par exemple le statut tabagique selon le sexe.

Le diagramme circulaire (angl. pie chart), ou diagramme en secteurs, divise un cercle en secteurs proportionnels aux fréquences relatives ; l'angle de chaque secteur est fi x 360°. Il montre bien la structure d'un tout comportant peu de catégories (de deux à cinq). Pour le niveau d'instruction dans mg2026 (figure 2.3B), les secteurs correspondent aux proportions de 12,5% (collège), 35,6% (lycée ou école professionnelle), 40,0% (études universitaires) et 11,9% (études postuniversitaires) ; le secteur des études universitaires a un angle de 0,40 x 360 = 144°. L'œil humain compare toutefois difficilement des angles voisins, de sorte que, pour plus de cinq catégories ou pour des comparaisons entre groupes, le diagramme en barres est presque toujours préférable.

2.7.2 Graphiques pour les variables quantitatives

L'histogramme est la représentation graphique du tableau de fréquences des données regroupées en classes. Les classes figurent sur l'axe horizontal, les fréquences sur l'axe vertical, et les barres sont accolées les unes aux autres, car la variable est continue. L'histogramme montre la forme de

la distribution : symétrique ou asymétrique, avec un ou deux pics, avec des valeurs isolées. La figure 2.3C montre une distribution de la glycémie présentant un pic autour de 95-115 mg/dL et une légère queue à droite. L'aspect de l'histogramme dépend de la largeur des classes : des classes trop larges masquent la forme de la distribution, des classes trop étroites produisent un graphique en dents de scie. Il vaut la peine d'essayer deux ou trois variantes.

La boîte à moustaches (box-plot) représente le résumé en cinq valeurs. La boîte s'étend de Q1 à Q3, une ligne à l'intérieur marque la médiane, et les moustaches se prolongent jusqu'à la plus petite et la plus grande valeur situées dans les limites de 1,5 x EIQ ; les valeurs situées en dehors de ces limites apparaissent sous forme de points isolés. La boîte à moustaches est idéale pour comparer plusieurs groupes. Dans la figure 2.3D, la médiane de la PAS est de 131 mmHg aussi bien chez les femmes que chez les hommes, et les boîtes se superposent presque complètement (femmes : Q1 121,25, Q3 140 ; hommes : Q1 122, Q3 138,75). Le graphique laisse prévoir le résultat du test du chapitre 6 : il n'existe pas de différence de PAS entre les sexes dans cet échantillon.

Le diagramme tige-feuilles (angl. stem-and-leaf) est un histogramme "écrit", qui conserve les valeurs individuelles. Chaque valeur est séparée en une tige (les premiers chiffres) et une feuille (le dernier chiffre). Pour les âges des 20 premiers patients de mg2026 :

Tableau 2.9. Diagramme tige-feuilles des âges des 20 premiers patients (tige = dizaines, feuille = unités).

Tige	Feuilles
3	7
4	0 5 7
5	0 2 3 4 5 6 6 7 7
6	0 2 2 5 6 9
7	0

Le diagramme montre immédiatement que la majorité des patients ont entre 50 et 59 ans et que la distribution est approximativement symétrique. On peut y lire directement le minimum (37), le maximum (70) et la médiane (moyenne des valeurs de rangs 10 et 11, toutes deux égales à 56). Il est utile pour de petits ensembles de données, écrits à la main.

Le diagramme en points (angl. dot plot) représente chaque observation par un point sur un axe ; les valeurs égales s'empilent les unes sur les autres. Pour de petits échantillons (moins de 30-40 observations), il est plus honnête qu'un diagramme en barres qui ne montre que la moyenne, car il permet au lecteur de voir chaque patient. De nombreuses revues recommandent aujourd'hui, pour les petits groupes, des graphiques qui superposent les points individuels à la boîte à moustaches.

Le diagramme de dispersion (nuage de points, angl. scatter plot) représente la relation entre deux variables numériques : chaque patient est un point de coordonnées (x, y). Nous l'utiliserons largement au chapitre 9 pour la corrélation et la régression ; par exemple, entre l'IMC et la PAS dans mg2026, le nuage de points monte modérément, avec un coefficient de corrélation r = 0,405. Le diagramme de dispersion aide aussi à juger certains résultats publiés. Dans une étude portant sur 867 patients atteints d'un cancer du côlon, la corrélation entre l'âge et la durée de l'intervention était de r = -0,088, statistiquement significative (p = 0,01) [55]. Un nuage de points avec r = -0,088 ressemble pratiquement à une tache sans aucune direction ; la significativité statistique vient du grand effectif de l'échantillon, et non d'une relation cliniquement importante.

Le graphique linéaire (angl. line plot) relie par des lignes des valeurs ordonnées dans le temps : la moyenne de la glycémie à 0, 3 et 6 mois, l'incidence mensuelle d'une maladie, l'évolution de la température d'un patient sur la feuille de surveillance. Les lignes n'ont de sens que si l'axe horizontal possède un ordre naturel, en général le temps.

Tableau 2.10. Le choix du graphique en fonction du type de variable.

Variable(s)	Graphique recommandé	À éviter
Une variable nominale	barres ; diagramme circulaire pour 2-5 catégories	diagramme circulaire avec de nombreuses catégories ou en 3D
Une variable ordinale	barres dans l'ordre des catégories ; barres empilées pour les échelles de Likert	diagramme circulaire qui brouille l'ordre
Une variable numérique	histogramme ; boîte à moustaches ; tige-feuilles ou points pour n petit	barres montrant la seule moyenne
Variable numérique par groupes	boîtes à moustaches juxtaposées ; points individuels pour n petit	barres avec "moyenne ± ES" sans précision
Deux variables numériques	diagramme de dispersion	lignes reliant des points sans ordre temporel
Évolution dans le temps	graphique linéaire	barres en 3D

2.7.3 Règles pour un graphique correct

Un graphique peut être exact sur le plan numérique et néanmoins trompeur. La figure 2.4 montre le piège le plus fréquent : l'axe tronqué. Dans mg2026, la glycémie moyenne est passée de 105,4 mg/dL à l'inclusion à 99,4 mg/dL à 3 mois et à 95,2 mg/dL à 6 mois, soit une baisse d'environ 10%. Si l'axe vertical commence à 94 mg/dL (panneau A), la barre de 6 mois paraît presque dix fois plus petite que celle de l'inclusion, comme si la glycémie s'était effondrée. Avec un axe partant de zéro (panneau B), les proportions entre les barres sont les proportions réelles.

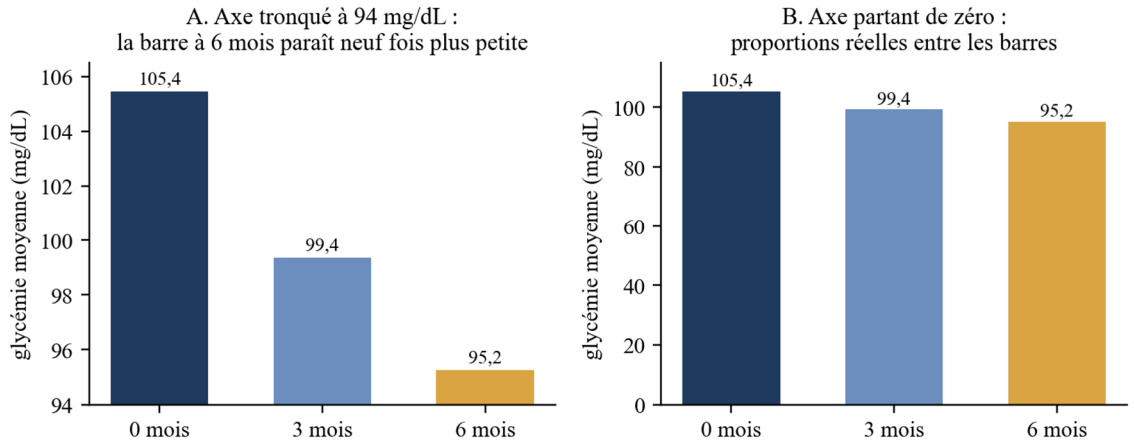


Figure 2.4. Le même résultat, deux impressions différentes. Glycémie moyenne à 0, 3 et 6 mois dans le jeu de données mg2026 (105,4 ; 99,4 ; 95,2 mg/dL). A : l'axe vertical tronqué à 94 mg/dL exagère visuellement la baisse. B : l'axe partant de zéro conserve les proportions réelles entre les barres. Données synthétiques (jeu de données mg2026).

La règle est simple : dans les diagrammes en barres, l'axe des valeurs commence à zéro, car c'est la longueur de la barre qui transmet la grandeur. Dans les graphiques linéaires et les diagrammes de dispersion, l'axe peut commencer à une valeur différente de zéro, car c'est la position des points qui compte, et non la longueur des barres ; dans ce cas, la coupure de l'axe doit être signalée de façon visible.

D'autres règles, tirées de l'expérience de la correction des travaux :

- Pas d'effets tridimensionnels. Un diagramme circulaire en 3D incliné fait paraître les secteurs du premier plan plus grands qu'ils ne le sont.
- Pas de deux axes verticaux à échelles différentes sur le même graphique, sauf nécessité absolue ; le choix des échelles peut faire paraître deux courbes parallèles ou divergentes, au gré de l'auteur.
- Les axes ont un titre et une unité de mesure (PAS, mmHg ; glycémie, mg/dL).

- La légende ou le titre de la figure précise n, la période et la source des données ; la figure doit pouvoir être comprise sans le texte.
- Les barres d'erreur sont expliquées : écart-type, erreur standard ou intervalle de confiance à 95%. Sans cette précision, les barres ne peuvent pas être interprétées.
- Les couleurs sont choisies de manière que le graphique puisse être lu et imprimé en noir et blanc, et par les personnes atteintes d'un déficit de la vision des couleurs.
- Un graphique, un message. Si un graphique a besoin d'un paragraphe d'explications pour être compris, il faut probablement le scinder en deux.

2.8 Le tableau descriptif d'un article : le "Tableau 1"

Presque tout article de recherche clinique commence la section des résultats par un tableau qui décrit les participants, appelé de manière informelle le "Tableau 1". Il présente les caractéristiques démographiques et cliniques de l'échantillon, souvent séparément pour les groupes comparés. Les conventions sont les mêmes dans toutes les revues : les variables numériques symétriques sont rapportées sous forme de moyenne ± écart-type ; les variables numériques asymétriques ou ordinales, sous forme de médiane (Q1-Q3) ; les variables qualitatives, sous forme de nombre de patients et de pourcentage, n (%). Les recommandations pour la publication exigent que le tableau précise, pour chaque variable, combien de valeurs manquent [25]. Le tableau 2.11 est le "Tableau 1" de l'étude simulée mg2026.

Tableau 2.11. Caractéristiques des patients à l'inclusion, par bras de traitement, jeu de données mg2026.

Caractéristique	Bras A (n = 53)	Bras B (n = 53)	Bras C (n = 54)	Total (n = 160)
Âge, ans, moyenne ± ÉT	57,53 ± 9,19	56,87 ± 9,42	54,52 ± 7,47	56,29 ± 8,77
Sexe féminin, n (%)	32 (60,4)	26 (49,1)	28 (51,9)	86 (53,8)
Milieu urbain, n (%)	25 (47,2)	36 (67,9)	34 (63,0)	95 (59,4)
Fumeurs actifs, n (%)	18 (34,0)	18 (34,0)	17 (31,5)	53 (33,1)
IMC, kg/m², médiane (Q1-Q3)	29,2 (26,6-31,1)	27,9 (26,1-31,4)	27,7 (25,95-31,4)	28,25 (26,2-31,4)
PAS, mmHg, moyenne ± ÉT	131,53 ± 13,84	132,79 ± 13,79	129,11 ± 14,01	131,13 ± 13,88
Glycémie à jeun, mg/dL, moyenne ± ÉT	107,79 ± 15,65	106,28 ± 14,70	102,30 ± 16,64	105,44 ± 15,76
Triglycérides, mg/dL, médiane (Q1-Q3)	143 (95-204)	144 (112-185)	166,5 (107-213)	151,5 (107-203,25)
Score FINDRISC, médiane (Q1-Q3)	10	10	10	10 (8-13)
Diabète confirmé par HGPO, n (%)	10 (18,9)	9 (17,0)	10 (18,5)	29 (18,1)

Note : ÉT, écart-type ; Q1, Q3, premier et troisième quartiles. Il n'y a pas de valeurs manquantes. Les pourcentages sont calculés à partir du nombre de patients de chaque bras.

Quelques remarques sur la lecture d'un tel tableau. La première chose à vérifier est que les nombres tombent juste : 32 + 26 + 28 = 86 femmes ; 53 + 53 + 54 = 160. La deuxième est que le format de chaque variable correspond à sa distribution : l'IMC et les triglycérides, asymétriques, sont rapportés sous forme de médiane (Q1-Q3), et la PAS et l'âge, symétriques, sous forme de moyenne ± ÉT. La troisième, propre aux essais randomisés, est la comparaison des bras. Le bras A compte davantage de femmes (60,4%) et moins de patients du milieu urbain (47,2%) que les

autres. Ces différences sont l'effet du hasard, puisque la répartition a été aléatoire ; c'est pourquoi les recommandations pour les essais randomisés préconisent de ne pas calculer de valeurs p pour les comparaisons du Tableau 1, mais de juger si les déséquilibres sont assez importants pour compter sur le plan clinique [56].

Les articles des collègues de la discipline offrent des exemples variés de présentation descriptive. Une étude d'exactitude diagnostique de l'échographie de contraste dans la cirrhose a rapporté l'âge sous forme de moyenne ± ÉT, 58,98 ± 8,90 ans, et le sexe en pourcentage, 56,3% d'hommes, pour 64 participants (37 atteints de cirrhose, 27 volontaires sains) ; le paramètre échographique était de 0,83 ± 0,09 chez les patients cirrhotiques contre 0,49 ± 0,11 chez les volontaires [2]. Une étude sur le diagnostic de l'obstruction biliaire maligne, portant sur 160 patients, a rapporté 85 hommes (53,1%) et un âge moyen de 69,31 ± 10,96 ans [57]. Dans l'étude sur l'hémodialyse du chapitre 1, les âges des deux cohortes ont été rapportés comme 59,7 ± 12,92 et 61 ± 12,93 ans [39]. Toutes ces présentations respectent la convention "moyenne ± ÉT" pour un âge approximativement symétrique.

Il vaut la peine de toujours vérifier aussi la cohérence interne des chiffres. Dans une étude sur la quantification échographique de la stéatose hépatique, les grades de stéatose ont été rapportés comme S0 48/176 (27,2%), S1 56/176 (31,6%), S2 14/176 (7,3%) et S3 59/176 (33,9%), avec des valeurs moyennes du paramètre UGAP croissant de $198{,}3 \pm \frac{25{,}7 d_B}{m}$ en S0 à $270{,}8 \pm \frac{31{,}62 d_B}{m}$ en S3 [58]. Si vous additionnez les numérateurs, vous obtenez 177, et non 176, et 14/176 correspond à 8,0%, et non à 7,3%. Il s'agit très probablement d'une erreur de rédaction du résumé, sans conséquence sur les conclusions, mais l'exercice montre pourquoi un lecteur attentif refait toujours les sommes. Remarquez aussi le nombre inégal de décimales (198,3 contre 216,86) : une bonne règle consiste à rapporter toutes les valeurs d'une même variable avec le même nombre de décimales, sans dépasser la précision de la mesure.

Pour les variables qualitatives issues de questionnaires, la présentation consiste presque exclusivement en pourcentages. Dans le sondage sur l'ostéoporose mené en pharmacie, sur 189 répondants, 78,8% étaient des femmes, les classes d'âge les plus nombreuses étaient 60-69 ans (31,7%) et 50-59 ans (30,7%), et 69,3% vivaient en milieu urbain [27]. Un tableau de fréquences complet ajouterait à chaque pourcentage l'effectif absolu, afin que le lecteur puisse refaire les calculs.

Étude de cas : moyennes, médianes et scores avec effet plancher

Une étude rétrospective portant sur 200 patients atteints de diabète de type 2 traités par inhibiteurs du SGLT-2 a rapporté les valeurs avant et après un an de traitement [59]. La glycémie est passée de 180,00 mg% (EIQ 154,50-207,00) à 130,00 mg% (EIQ 117,50-150,00), l'HbA1c de 8,40% (EIQ 7,98-9,15%) à 7,30% (EIQ 6,90-7,95%), le poids de 90,50 kg (82,00-106,50) à 89,00 kg (77,50-100,00), la pression systolique de 140,00 mmHg (130,00-160,00) à 120,00 mmHg (115,50-130,00), toutes avec $p < 0{,}001$, à l'exception du poids ($p = 0{,}018$). Le tour abdominal a été rapporté différemment, comme 107,05 (± 16,39) et $102{,}50 (\pm 15{,}11)$ ($p = 0{,}042$). La présence de l'EIQ montre que les premières valeurs sont des médianes, bien que le résumé les appelle "moyennes" ; c'est une incohérence terminologique fréquente, que le lecteur doit reconnaître d'après la forme de la présentation. Le fait que certaines variables soient rapportées sous forme de médiane (EIQ) et d'autres sous forme de moyenne ± ÉT est correct si le choix a suivi la distribution de chaque variable, mais les méthodes devraient l'indiquer explicitement, ainsi que le test apparié utilisé. Un second exemple : dans une étude transversale portant sur 304 adultes diabétiques, un score d'implication numérique a été rapporté sous forme de médiane [Q1-Q3], 1,71 [1,71-2,39] contre 3,86 [3,86-4,43] entre les groupes, et 35,5% des participants avaient une implication numérique élevée [7]. Remarquez que, dans les deux groupes, la médiane coïncide avec le

> premier quartile. Cela se produit lorsqu'un score a peu de valeurs possibles et que de nombreuses observations sont égales à la plus petite valeur d'un groupe : au moins un quart des patients, voire jusqu'à la moitié, ont exactement la même valeur. Ce phénomène s'apparente à l'effet plancher, décrit explicitement dans une étude de validation d'un questionnaire sur le fardeau de la ménopause, portant sur 308 femmes, où 37,0% des répondantes avaient le score minimal à une sous-échelle et 23,4% à une autre [60]. Lorsqu'une forte proportion de patients a le score minimal, la moyenne et l'écart-type décrivent mal la distribution, et l'instrument ne peut pas distinguer entre les patients situés dans le bas de l'échelle. Ce qu'il faudrait rapporter en plus : la proportion d'observations à la valeur minimale et à la valeur maximale et, pour les scores bornés, un histogramme.

2.9 L'analyse descriptive complète du jeu de données mg2026

Nous terminons la partie théorique par la description complète du jeu de données que vous utiliserez tout au long du semestre. Les valeurs des tableaux 2.12 et 2.13 sont celles que vous devez retrouver dans JASP et dans Google Sheets lors des séances de travaux pratiques 1 et 2 ; si votre résultat diffère, cherchez la cause au niveau de l'importation ou du nettoyage.

Tableau 2.12. Statistiques descriptives des variables numériques à l'inclusion, jeu de données mg2026 (n = 160).

Variable	Moyenne	ÉT	Médiane	Q1-Q3	Min-max	CV (%)	Asymétrie
Âge (ans)	56,29	8,77	56	50,75-62	35-79	15,6	0,01
Taille (cm)	168,79	9,05	168	162-177	145-192	5,4	0,10
Poids (kg)	83,05	15,09	80,2	72,05-90,93	58,1-159,9	18,2	1,29
$\text{IMC}\left(\frac{\text{kg}}{m^2}\right)$	29,04	3,91	28,25	26,2-31,4	23,5-48,3	13,5	1,42
Tour de taille (cm)	88,54	12,11	86	80,75-95	65-140	13,7	0,86
PAS (mmHg)	131,13	13,88	131	122-140	102-169	10,6	0,01
PAD (mmHg)	79,23	8,63	80	74-85	58-103	10,9	-0,09
Fréquence cardiaque (battements/min)	74,71	9,30	74,5	69-80	51-102	12,4	0,05
Cholestérol total (mg/dL)	209,39	36,91	207	183-235	119-306	17,6	0,02
HDL-cholestérol (mg/dL)	49,52	11,78	49	41-59	20-78	23,8	0,04
Triglycérides (mg/dL)	161,18	76,30	151,5	107-203,25	44-442	47,3	0,95
Glycémie à jeun (mg/dL)	105,44	15,76	104,5	93-116	75-154	15,0	0,37
HbA1c (%)	5,58	0,45	5,6	5,3-5,9	4,5-7,0	8,0	0,14
Score FINDRISC	10,37	4,27	10	8-13	0-25	41,2	0,39

Tableau 2.13. Distribution des variables qualitatives, jeu de données mg2026 (n = 160).

Variable	Catégorie	n	%
Sexe	femmes / hommes	86 / 74	53,8 / 46,3
Milieu	urbain / rural	95 / 65	59,4 / 40,6
Tabagisme	non-fumeur / ancien fumeur / fumeur	74 / 33 / 53	46,3 / 20,6 / 33,1
Niveau d'instruction	collège / lycée ou école professionnelle / études universitaires / études postuniversitaires	20 / 57 / 64 / 19	12,5 / 35,6 / 40,0 / 11,9
Activité physique (1-5)	1 / 2 / 3 / 4 / 5	21 / 42 / 60 / 34 / 3	13,1 / 26,3 / 37,5 / 21,3 / 1,9
Satisfaction à 6 mois (1-5)	1 / 2 / 3 / 4 / 5	4 / 19 / 54 / 55 / 28	2,5 / 11,9 / 33,8 / 34,4 / 17,5
Bras de traitement	$\frac{A}{B}$ / C	53 / 53 / 54	33,1 / 33,1 / 33,8
Diabète confirmé par HGPO	oui / non	29 / 131	18,1 / 81,9

Note : la somme des pourcentages peut être de 100,1 en raison des arrondis.

Comment rédiger l'interprétation de ces tableaux ? Voici un modèle, dans le style de la section des résultats d'un article. "Ont été inclus 160 patients (86 femmes, 53,8%), d'un âge moyen de 56,29 ± 8,77 ans (extrêmes 35-79). La majorité vivaient en milieu urbain (59,4%), et un tiers étaient fumeurs actifs (33,1%). La médiane de l'IMC était de 28,25 kg/m² (Q1-Q3 26,2-31,4), avec 54 patients (33,8%) obèses. La pression systolique moyenne était de 131,13 ± 13,88 mmHg, et 41 patients (25,6%) avaient une PAS ≥ 140 mmHg à l'inclusion. La glycémie à jeun moyenne était de 105,44 ± 15,76 mg/dL, l'HbA1c moyenne de 5,58 ± 0,45%, et le diabète a été confirmé par HGPO chez 29 patients (18,1%). Les triglycérides avaient une distribution asymétrique, avec une médiane de 151,5 mg/dL (Q1-Q3 107-203,25)."

Quelques remarques sur les choix faits dans ce texte. L'âge, la PAS, la glycémie, l'HbA1c, le cholestérol et le HDL ont une faible asymétrie et des moyennes pratiquement égales à leurs médianes ; on les rapporte donc sous forme de moyenne ± ÉT. L'IMC, le poids, le tour de taille et les triglycérides présentent une asymétrie positive importante (supérieure à 0,8) et sont plus correctement rapportés sous forme de médiane (Q1-Q3). Le score FINDRISC est une variable discrète comportant de nombreuses valeurs possibles ; on peut le rapporter des deux façons, mais la médiane avec l'EIQ est plus prudente. Le niveau d'instruction, l'activité physique et la satisfaction sont ordinaux : on les rapporte par des fréquences et, éventuellement, par la médiane (respectivement 3, 3 et 4), jamais par la moyenne ± ÉT. Pour les variables dichotomiques, le pourcentage décrit entièrement la variable ; on ne rapporte pas d'écart-type.

Étude de cas : quand l'écart-type est presque égal à la moyenne

L'écart-type rapporté dans un article peut servir de test rapide de la forme de la distribution. Dans une étude rétrospective portant sur 486 patients atteints de COVID-19 (200 sans diabète, 62 atteints de diabète de type 1, 224 de diabète de type 2), la durée d'hospitalisation était de 11,6 ± 7,0 jours dans le diabète de type 2 et de 9,1 ± 5,8 jours dans le diabète de type 1, et la glycémie à l'admission de 178,3 ± 34,7 mg/dL contre 164,8 ± 39,6 mg/dL [8]. Pour la glycémie, le coefficient de variation est d'environ 20%, ce qui est compatible avec une distribution raisonnablement symétrique. Pour la durée d'hospitalisation, le CV est d'environ 60% (7,0/11,6), et m - 2s = 11,6 - 14,0 = -2,4 jours, une valeur impossible.

> La distribution est presque certainement asymétrique à droite, avec quelques patients hospitalisés très longtemps, et la médiane avec l'EIQ l'aurait mieux décrite. La situation est encore plus nette dans une étude sur le diagnostic de l'accident vasculaire cérébral aux urgences, portant sur 85 patients : la protéine GFAP était de $142{,}91 \pm 102{,}19$ pg/mL chez les patients ayant un accident vasculaire et de $37{,}76 \pm 19{,}92$ pg/mL chez ceux qui n'en avaient pas, et l'UCH-L1 de 1307,68 ± 967,54 contre $189{,}81 \pm 92{,}69$ pg/mL[61]. Un écart-type supérieur à 70% de la moyenne signale une forte asymétrie ; pour une telle variable, la comparaison devrait se faire avec un test non paramétrique (chapitre 7) ou après transformation logarithmique (chapitre 3). En revanche, une autre étude du même service d'urgences, consacrée à la thrombolyse dans l'accident vasculaire cérébral ischémique, a correctement rapporté les délais, variables typiquement asymétriques, au moyen de médianes : une médiane de 100 minutes entre le début des symptômes et l'arrivée aux urgences et de 15,5 minutes entre l'arrivée et la tomodensitométrie, chez 258 patients éligibles parmi les 355 évalués initialement, dont 46 ont reçu une thrombolyse (17,83%) [62]. Conclusion pour le lecteur : avant d'accepter une comparaison de moyennes, comparez l'écart-type à la moyenne ; si s dépasse environ la moitié de m pour une variable qui ne peut pas être négative, la distribution est asymétrique et la moyenne n'est pas la valeur typique.

2.10 Problèmes résolus

Problème résolu 2.1. Tableau de fréquences comportant des valeurs manquantes : la pression systolique dans une consultation externe de cardiologie

Énoncé. Dans une consultation externe de cardiologie, on a mesuré la pression artérielle systolique de 50 patients hypertendus lors de leur première visite. Les résultats ont été regroupés dans le tableau ci-dessous, dont certaines valeurs manquent. (a) Complétez les valeurs A, B, C et D. (b) Quelle est la probabilité qu'un patient choisi au hasard dans ce groupe ait une PAS d'au moins 160 mmHg ? (c) Dans quelle classe se trouve la médiane ? Estimez la médiane et la moyenne à partir du tableau.

Tableau 2.14. PAS chez 50 patients hypertendus (données groupées).

PAS (mmHg)	Fréquence absolue	Fréquence relative (%)	Fréquence cumulée (%)
100-119	4	8,0	8,0
120-139	11	22,0	30,0
140-159	A	34,0	C
160-179	12	B	88,0
180-199	D	8,0	96,0
200-219	2	4,0	100,0
Total	50	100,0	

Solution. (a) A s'obtient à partir de la fréquence relative : A = 34,0% x 50 = 0,34 x 50 = 17 patients. B s'obtient à partir de la fréquence absolue : $B = \frac{12}{50} \times 100 = 24{,}0\%$. C est la fréquence cumulée de la troisième classe : C = 30,0 + 34,0 = 64,0%. Vérification : 64,0 + 24,0 = 88,0, exactement la valeur de la ligne suivante. D s'obtient à partir de la fréquence relative : D = 0,08 x 50 = 4 patients. Vérification finale : 4 + 11 + 17 + 12 + 4 + 2 = 50, et 8 + 22 + 34 + 24 + 8 + 4 = 100%. Le tableau tombe juste.

(b) Une PAS ≥ 160 mmHg correspond aux trois dernières classes : 24,0 + 8,0 + 4,0 = 36,0%. De façon équivalente, 100 - 64,0 = 36,0% (le complément de la fréquence cumulée jusqu'à 159 mmHg). La probabilité est de 0,36, soit 18 patients sur 50.

(c) La médiane est la valeur qui partage les 50 patients en deux moitiés ; elle se situe donc entre la 25e et la 26e valeur ordonnée. La fréquence cumulée est de 30,0% (15 patients) à

la fin de la classe 120-139 et de 64,0% (32 patients) à la fin de la classe 140-159 ; la 25e valeur se trouve donc dans la classe 140-159, qui est la classe médiane. Si nous supposons que les 17 patients de cette classe sont répartis uniformément, la médiane s'estime par interpolation, à l'aide des bornes réelles de la classe (139,5-159,5, largeur h = 20) :

$$\text{Médiane} \approx 139{,}5 + \frac{25-15}{17} \times 20 = 139{,}5 + 11{,}8 = 151{,}3 \text{ mmHg}$$

La moyenne s'estime à l'aide des centres de classe (109,5 ; 129,5 ; 149,5 ; 169,5 ; 189,5 ; 209,5) :

$$m \approx \frac{4 \times 109{,}5 + 11 \times 129{,}5 + 17 \times 149{,}5 + 12 \times 169{,}5 + 4 \times 189{,}5 + 2 \times 209{,}5}{50}$$

$$= \frac{438 + 1\,424{,}5 + 2\,541{,}5 + 2\,034 + 758 + 419}{50} = \frac{7\,615}{50} = 152{,}3 \text{ mmHg}$$

Interprétation. Plus d'un tiers des patients hypertendus (36%) se présentent avec des valeurs d'au moins 160 mmHg, c'est-à-dire dans la zone de l'hypertension de grade 2 ou plus. La moyenne et la médiane estimées sont proches (152,3 contre 151,3 mmHg), la distribution est donc approximativement symétrique, avec une légère queue vers les valeurs élevées. Attention : ces deux valeurs sont des estimations obtenues à partir de données groupées ; avec les données individuelles, nous obtiendrions des valeurs légèrement différentes.

Problème résolu 2.2. Tableau de fréquences d'une variable discrète : les consultations annuelles

Énoncé. Un médecin de famille a compté les consultations de la dernière année pour 80 patients atteints de diabète de type 2 choisis au hasard sur sa liste. (a) Complétez les valeurs A, B, C et D. (b) Calculez le mode, la médiane et la moyenne du nombre de consultations. (c) Quelle proportion des patients a eu au moins 4 consultations ?

Tableau 2.15. Nombre de consultations par an chez 80 patients atteints de diabète de type 2.

Consultations par an	Fréquence absolue	Fréquence relative (%)	Fréquence cumulée (%)
0	6	7,5	7,5
1	14	17,5	25,0
2	A	25,0	50,0
3	18	B	72,5
4	C	12,5	D
5	8	10,0	95,0
6	4	5,0	100,0
Total	80	100,0	

Solution. (a) A = 0,25 x 80 = 20 patients (ou à partir des fréquences cumulées : 50,0 - 25,0 = 25,0%, donc également 20). $B = \frac{18}{80} \times 100 = 22{,}5\%$ (vérification : 50,0 + 22,5 = 72,5). C = 0,125 x 80 = 10 patients. D = 72,5 + 12,5 = 85,0% (vérification : 85,0 + 10,0 = 95,0). Somme des fréquences absolues : 6 + 14 + 20 + 18 + 10 + 8 + 4 = 80.

(b) Le mode est la valeur dont la fréquence est la plus élevée : 2 consultations (20 patients). Pour la médiane, n = 80 est pair, donc la médiane est la moyenne des valeurs de rangs 40 et 41. La fréquence cumulée est exactement de 50,0% (40 patients) à la valeur 2 : la 40e valeur est 2, et la 41e est la première valeur égale à 3. La médiane est de (2 + 3)/2 = 2,5 consultations. La moyenne est une moyenne pondérée par les fréquences :

$$m = \frac{0\times6+1\times14+2\times20+3\times18+4\times10+5\times8+6\times4}{80} = \frac{0+14+40+54+40+40+24}{80} = \frac{212}{80} = 2{,}65 \text{ consultations}$$

(c) Au moins 4 consultations : 12,5 + 10,0 + 5,0 = 27,5%, soit 22 patients. De façon équivalente, 100 - 72,5 = 27,5%.

Interprétation. Le patient "typique" a deux ou trois consultations par an, et un quart des patients (25,0%, la fréquence cumulée à la valeur 1) ont au plus une consultation. Le piège du problème est la médiane : lorsque la fréquence cumulée atteint exactement 50% à une valeur, la médiane n'est pas cette valeur, mais la moyenne entre celle-ci et la valeur suivante. À l'examen, de nombreux étudiants répondent "2". Sur le plan clinique, les 6 patients (7,5%) qui n'ont eu aucune consultation en un an sont ceux qu'il faudrait convoquer activement pour un contrôle.

Problème résolu 2.3. Classement des variables d'une fiche d'observation

Énoncé. La fiche d'une étude sur des patients hospitalisés pour pneumonie comporte les champs suivants : (1) code du patient ; (2) âge (ans) ; (3) sexe ; (4) score CURB-65 (0-5) ; (5) saturation en oxygène (%) ; (6) nombre de jours d'hospitalisation ; (7) intensité perçue de la dyspnée (absente, légère, modérée, sévère) ; (8) germe isolé (pneumocoque, Haemophilus, Mycoplasma, autre, non isolé) ; (9) décès à l'hôpital (oui/non) ; (10) température à l'admission (°C). Pour chaque variable, précisez le type, l'indicateur descriptif adapté et un graphique adapté. Solution. Le tableau 2.16 résume la réponse.

Tableau 2.16. Classement des variables de la fiche de l'étude sur la pneumonie.

Variable	Type	Indicateur descriptif	Graphique
Code du patient	identifiant, non analysé	-	-
Âge	numérique continue, échelle de rapport	moyenne ± ÉT (ou médiane en cas d'asymétrie)	histogramme, boîte à moustaches
Sexe	nominale dichotomique	n (%)	barres
Score CURB-65	ordinale (score à 6 valeurs)	médiane (Q1-Q3), fréquences	barres dans l'ordre du score
Saturation en O2	numérique continue, bornée supérieurement à 100%	médiane (Q1-Q3), car elle est habituellement asymétrique à gauche	histogramme, boîte à moustaches
Jours d'hospitalisation	numérique discrète, asymétrique	médiane (Q1-Q3)	histogramme, boîte à moustaches
Intensité de la dyspnée	ordinale	fréquences, médiane	barres dans l'ordre des catégories
Germe isolé	nominale à 5 catégories	n (%)	barres (par ordre décroissant)
Décès à l'hôpital	nominale dichotomique	n (%)	barres
Température	numérique continue, échelle d'intervalle	moyenne ± ÉT	histogramme, boîte à moustaches

Interprétation. Deux détails méritent attention. Le score CURB-65 est numérique par sa forme, mais ordinal par sa nature : la différence entre 0 et 1 point n'a pas la même signification clinique que celle entre 3 et 4. La saturation en oxygène ne peut pas dépasser 100%, et chez les patients atteints de pneumonie les valeurs se concentrent vers le haut de l'échelle, avec une queue vers les petites valeurs ; la médiane convient mieux que la moyenne. La température est l'exemple classique d'échelle d'intervalle : cela n'a pas de sens de dire que 40 °C est le double de 20 °C, et le coefficient de variation n'a donc pas de sens non plus.

Problème résolu 2.4. Description complète d'une petite série : l'HbA1c chez dix patients

Énoncé. Les valeurs d'HbA1c (%) des 10 patients atteints de diabète de type 2 vus en consultation au cours d'une journée sont : 6,2 ; 6,8 ; 7,1 ; 7,4 ; 7,5 ; 7,9 ; 8,3 ; 8,8 ; 9,6 ; 11,4. Calculez la moyenne, la médiane, l'écart-type, le coefficient de variation, l'erreur standard, les quartiles et l'écart interquartile, et vérifiez s'il existe des valeurs extrêmes selon la règle 1,5 x EIQ. Quelle proportion des patients a une HbA1c $\geq 7\%$?

Solution. Les valeurs sont déjà ordonnées, n = 10.

Moyenne : la somme des valeurs est de 81,0, donc $m = \frac{81,0}{10} = 8,1\%$.

Médiane : n est pair, les valeurs de rangs 5 et 6 sont 7,5 et 7,9 ; médiane = (7,5 + 7,9)/2 = 7,7%.

Écart-type : les écarts à la moyenne sont -1,9 ; -1,3 ; -1,0 ; -0,7 ; -0,6 ; -0,2 ; 0,2 ; 0,7 ; 1,5 ; 3,3. Leurs carrés sont 3,61 ; 1,69 ; 1,00 ; 0,49 ; 0,36 ; 0,04 ; 0,04 ; 0,49 ; 2,25 ; 10,89, de somme 20,86.

$s^2 = \frac{20,86}{9} = 2,318$; $s = \sqrt{2,318} = 1,52\%$

Coefficient de variation : $\text{CV} = \frac{1,52}{8,1} \times 100 = 18,8\%$. Erreur standard : $\text{ES} = \frac{1,52}{\sqrt{10}} = \frac{1,52}{3,162} = 0,48\%$.

Quartiles (méthode des médianes des moitiés) : la moitié inférieure 6,2 ; 6,8 ; 7,1 ; 7,4 ; 7,5 a pour médiane Q1 = 7,1% ; la moitié supérieure 7,9 ; 8,3 ; 8,8 ; 9,6 ; 11,4 a pour médiane Q3 = 8,8%. EIQ = 8,8 - 7,1 = 1,7%. (Google Sheets et JASP donnent, par interpolation, Q1 = 7,175 et Q3 = 8,675.)

Valeurs extrêmes : les limites sont 7,1 - 1,5 x 1,7 = 7,1 - 2,55 = 4,55% et 8,8 + 2,55 = 11,35%. La valeur de 11,4% dépasse légèrement la limite supérieure et apparaît comme une valeur extrême.

Proportion avec HbA1c ≥ 7% : 8 patients sur 10, soit 80%.

Interprétation. La moyenne (8,1%) est supérieure à la médiane (7,7%), signe qu'une valeur élevée, 11,4%, tire la moyenne vers le haut. Sans cette valeur, la moyenne des neuf patients $\frac{\text{ar fi}\quad 69,6}{9} = 7,73\%$, pratiquement égale à la médiane. Le patient à 11,4% n'est pas une erreur, mais un patient dont le diabète est très mal équilibré et qui a besoin d'une intensification urgente du traitement. Pour une série aussi petite, la présentation correcte est la médiane de 7,7% (Q1-Q3 7,1-8,8), éventuellement complétée par la moyenne ± ÉT, 8,1 ± 1,52%. Sur le plan clinique, 80% des patients n'atteignent pas la cible d'HbA1c inférieure à 7%, et c'est en fait le message principal de la série.

Problème résolu 2.5. Écart-type ou erreur standard ?

Énoncé.

(a) Un article rapporte pour 100 patients hypertendus "PAS $158 \pm 2,1$ mmHg", sans préciser ce que représente 2,1. Qu'est-ce qui est le plus plausible : un écart-type ou une erreur standard ? (b) Un autre article rapporte la durée d'hospitalisation de 81 patients comme 6,5 ± 5,9 jours (moyenne ± ÉT). Que pouvez-vous dire de la forme de la distribution ? Calculez l'erreur standard.

Solution.

(a) Si 2,1 était l'écart-type, environ 95% des patients auraient une PAS comprise entre 158 - 4,2 = 153,8 et 158 + 4,2 = 162,2 mmHg, c'est-à-dire que tous les patients hypertendus auraient des pressions presque identiques. C'est biologiquement invraisemblable : dans mg2026, l'écart-type de la PAS est de 13,88 mmHg. Si 2,1 est l'erreur standard, l'écart-type ar fi $s = \text{ES} \times \sqrt{n} = 2,1 \times \sqrt{100} = 2,1 \times 10 = 21$ mmHg, une valeur parfaitement plausible pour un groupe d'hypertendus. Conclusion : il s'agit presque certainement de l'erreur standard, et l'auteur aurait dû le préciser.

(b) Le coefficient de variation est $\frac{5,9}{6,5} \times 100 = 91\%$, et m - 2s = 6,5 - 11,8 = -5,3 jours, ce qui est impossible. La distribution est fortement asymétrique à droite : la plupart des patients restent quelques jours, et quelques-uns restent très longtemps. La médiane et l'EIQ auraient été plus informatifs. L'erreur standard est de $\text{ES} = \frac{5,9}{\sqrt{81}} = \frac{5,9}{9} = 0,66$ jours.

Interprétation. La première question à vous poser lorsque vous voyez le signe ± est "que représente-t-il ?". La seconde est "est-ce plausible ?". Un rapport s/m très faible pour une variable biologique suggère une erreur standard ; un rapport s/m proche de 1 pour une variable positive signale une asymétrie. L'erreur standard de 0,66 jour montre que la moyenne de 6,5 jours est estimée assez précisément, bien que les patients pris individuellement varient beaucoup ; ce sont deux choses différentes.

Problème résolu 2.6. Moyenne pondérée et moyenne géométrique

Énoncé. (a) Trois cabinets rapportent la moyenne de l'HbA1c de leurs patients diabétiques : 7,2% (40 patients), 7,8% (25 patients) et 6,9% (85 patients). Quelle est la moyenne de l'HbA1c pour l'ensemble des 150 patients ? (b) Chez un patient, la protéine C réactive (CRP) a pris, lors de trois dosages successifs, les valeurs 2, 8 et 32 mg/L. Calculez la moyenne arithmétique et la moyenne géométrique. Laquelle décrit le mieux la valeur typique ?

Solution. (a) Moyenne pondérée par le nombre de patients :

$$m = \frac{40 \times 7{,}2 + 25 \times 7{,}8 + 85 \times 6{,}9}{40 + 25 + 85} = \frac{288 + 195 + 586{,}5}{150} = \frac{1\,069{,}5}{150} = 7{,}13\%$$

La moyenne simple des trois moyennes $\frac{\text{ar fi} \quad (7{,}2+7{,}8+6{,}9)}{3} = 7{,}3\%$, une valeur erronée, car elle accorde au cabinet de 25 patients la même importance qu'à celui de 85.

(b) Moyenne arithmétique : $\frac{2+8+32}{3} = \frac{42}{3} = \frac{14\text{mg}}{L}$. Moyenne géométrique : racine cubique de 2 x 8 x 32 = 512, soit 8 mg/L. De façon équivalente, avec les logarithmes décimaux : log 2 = 0,301 ; log 8 = 0,903 ; log 32 = 1,505 ; leur moyenne = 2,709/3 = 0,903 ; 10 à la puissance 0,903 = 8.

Interprétation. Les valeurs de la CRP sont multipliées par quatre d'un dosage à l'autre ; sur l'échelle logarithmique, elles sont équidistantes, et la valeur centrale, 8 mg/L, est la valeur typique. La moyenne arithmétique, 14 mg/L, est tirée vers le haut par la dernière valeur. Pour les marqueurs qui varient de façon multiplicative (CRP, titres, charge virale, triglycérides), la moyenne géométrique est le choix naturel.

2.11 Exercices proposés

1. Classez les variables suivantes (nominale, ordinale, numérique discrète, numérique continue) et précisez l'échelle de mesure : le groupe sanguin ; le score d'Apgar à 5 minutes ; le nombre de grossesses antérieures ; la température corporelle ; la classe NYHA ; l'HbA1c ; l'accord avec l'affirmation "Je suis satisfait du traitement" (échelle de Likert à 5 niveaux) ; le code numérique du patient ; la durée d'hospitalisation en jours ; le statut tabagique (oui/non).
2. Complétez le tableau de fréquences de la glycémie à jeun chez 40 patients : classe 70-89 mg/dL, 6 patients, 15,0%, cumulé 15,0% ; classe 90-109, A patients, 35,0%, cumulé B ; classe 110-125, 10 patients, C%, cumulé 75,0% ; classe 126-139, D patients, 12,5%, cumulé 87,5% ; classe ≥ 140, E patients, cumulé 100,0%. Quelle proportion des patients a une glycémie située dans l'intervalle diagnostique du diabète (≥ 126 mg/dL) ?
3. Pour les valeurs de PAS 118, 124, 130, 136, 142, 150 mmHg, calculez à la main la moyenne, la médiane, l'écart moyen absolu, la variance, l'écart-type, le coefficient de variation et l'erreur standard.
4. La durée d'hospitalisation (jours) de 10 patients opérés pour cholécystite a été de : 3, 4, 4, 5, 5, 6, 7, 8, 12, 31. Calculez la moyenne, la médiane et le mode. Déterminez les quartiles par la méthode des médianes des moitiés et vérifiez les valeurs extrêmes à l'aide de la

règle 1,5 x EIQ. Comment la moyenne change-t-elle sans la dernière valeur ? Quels indicateurs rapporteriez-vous ?

5. Une étude compte 250 valeurs de triglycérides, comprises entre 68 et 412 mg/dL. Combien de classes la règle de Sturges recommande-t-elle ? Quelle largeur de classe en résulte et à quelle valeur commode l'arrondiriez-vous ?
6. Un article rapporte, pour 64 patients, une glycémie à jeun de 112 ± 3 mg/dL. Si 3 est l'erreur standard, quel est l'écart-type ? Quelle interprétation (ÉT ou ES) est la plus plausible, et pourquoi ? Calculez l'erreur standard de la PAS pour les femmes de mg2026 (ÉT 14,47 mmHg, n = 86).
7. L'âge moyen des patients de trois services d'un hôpital est de 45,2 ans (120 patients), 61,8 ans (60 patients) et 70,4 ans (20 patients). Calculez l'âge moyen de l'ensemble des 200 patients et comparez-le à la moyenne simple des trois moyennes.
8. À l'aide du tableau 2.11 : (a) vérifiez le pourcentage de femmes dans le bras B ; (b) identifiez les variables rapportées sous forme de médiane (Q1-Q3) et expliquez ce choix ; (c) expliquez pourquoi, dans un essai randomisé, on ne calcule pas de valeurs p pour les comparaisons du Tableau 1.
9. Choisissez le graphique adapté pour : (a) la distribution des groupes sanguins chez 500 donneurs ; (b) la comparaison de la glycémie entre trois bras de traitement ; (c) la relation entre l'âge et la pression systolique ; (d) l'évolution mensuelle des cas de grippe dans un département sur deux ans ; (e) la distribution du poids de naissance de 1 200 nouveau-nés ; (f) les réponses à une question de Likert à 5 niveaux, comparées entre femmes et hommes.
10. La charge virale du VIH chez quatre patients est de 1 000, 10 000, 100 000 et 1 000 000 copies/mL. Calculez la moyenne arithmétique et la moyenne géométrique (utilisez les logarithmes décimaux). Pourquoi les laboratoires et les cliniciens expriment-ils ces valeurs sur une échelle logarithmique ?

Pièges fréquents

- Calculer la moyenne de variables nominales codées par des chiffres (par exemple, "moyenne du sexe = 1,46").
- Rapporter la moyenne ± ÉT pour des variables ordinales (scores de Likert, stades) ou pour des variables très asymétriques (durée d'hospitalisation, marqueurs inflammatoires).
- Confondre l'écart-type et l'erreur standard ; "moyenne ± ES" fait paraître les données beaucoup plus homogènes.
- Diviser par n au lieu de n - 1 dans le calcul de la variance d'un échantillon (STDEVP au lieu de STDEV).
- Prendre comme médiane la valeur à laquelle la fréquence cumulée atteint exactement 50% ; la réponse correcte est la moyenne entre cette valeur et la valeur suivante.
- Exclure des valeurs extrêmes simplement parce qu'elles "faussent la moyenne", sans démontrer qu'il s'agit d'erreurs.
- Diagrammes en barres à axe tronqué, diagrammes circulaires en 3D ou comportant trop de catégories.
- Coefficient de variation calculé pour des variables mesurées sur une échelle d'intervalle (température en °C).

- Pourcentages invérifiables faute d'effectifs absolus.

À retenir

- Le type de la variable (nominale, ordinale, discrète, continue) détermine les indicateurs, les graphiques et les tests autorisés.
- Les données s'organisent à raison d'une ligne par patient et d'une colonne par variable, avec un dictionnaire des données rédigé avant le recueil ; les erreurs se vérifient dans le document source.
- Les fréquences relatives et cumulées peuvent être reconstituées les unes à partir des autres ; la fréquence cumulée n'a de sens que pour des données ordonnées.
- La moyenne convient aux données numériques symétriques ; la médiane aux données asymétriques ou ordinales ; le mode aux données nominales ; la moyenne géométrique aux données positives à variation multiplicative.
- L'écart-type décrit la variabilité des patients ; l'erreur standard $\left(\frac{s}{\sqrt{n}}\right)$ décrit la précision de la moyenne et diminue quand n augmente.
- Le résumé en cinq valeurs et la boîte à moustaches montrent le centre, la dispersion, l'asymétrie et les valeurs extrêmes ; la règle 1,5 x EIQ signale des valeurs à vérifier, et non à supprimer.
- Un s proche de m pour une variable positive signale une asymétrie ; dans ce cas, on rapporte la médiane (Q1-Q3).
- Dans le "Tableau 1" : moyenne ± ÉT pour les variables symétriques, médiane (Q1-Q3) pour les variables asymétriques, n (%) pour les variables qualitatives ; les nombres doivent tomber juste.
- Les diagrammes en barres commencent à zéro, sans effets 3D ; chaque graphique a des axes légendés, des unités et n.

Questions d'autoévaluation

1. (QRU) Le stade de fibrose hépatique (F0, F1, F2, F3, F4) est une variable :
    a) nominale
    b) ordinale
    c) numérique discrète sur une échelle de rapport
    d) numérique continue
    e) dichotomique
2. (QRU) Pour la variable "durée d'hospitalisation", de moyenne 11,6 jours et d'écart-type 7,0 jours, l'indicateur de tendance centrale le plus adapté est :
    a) la moyenne arithmétique
    b) le mode
    c) la médiane
    d) l'étendue
    e) le coefficient de variation
3. (QRU) Dans un tableau de fréquences portant sur 40 patients, une classe a une fréquence relative de 22,5%. Sa fréquence absolue est :
    a) 22
    b) 9
    c) 18
    d) 10
    e) 4

4. (QRU) L'écart-type de la PAS dans un échantillon de 100 patients est de 15 mmHg. L'erreur standard de la moyenne est :
    a) 0,15 mmHg
    b) 1,5 mmHg
    c) 15 mmHg
    d) 150 mmHg
    e) 3,87 mmHg
5. (QRU) Pour la série 2, 3, 3, 4, 5, 6, 7, 8, la médiane est :
    a) 4
    b) 4,5
    c) 4,75
    d) 3
    e) 5
6. (QRM) Au sujet de la médiane, les affirmations correctes sont :
    a) elle est égale au deuxième quartile et au 50e percentile
    b) elle est peu influencée par les valeurs extrêmes
    c) c'est l'indicateur adapté aux variables ordinales
    d) elle est toujours supérieure à la moyenne
    e) elle peut être calculée pour des variables nominales
7. (QRM) Au sujet de l'écart-type, les affirmations correctes sont :
    a) il s'exprime dans la même unité de mesure que les données
    b) c'est la racine carrée de la variance
    c) il diminue à mesure que l'effectif de l'échantillon augmente, tout comme l'erreur standard
    d) pour un échantillon, il se calcule avec n - 1 au dénominateur
    e) il ne peut pas être négatif
8. (QRM) Dans une distribution à asymétrie positive (queue à droite) :
    a) la moyenne est, en règle générale, supérieure à la médiane
    b) le coefficient d'asymétrie est positif
    c) les valeurs élevées rares tirent la moyenne vers le haut
    d) la moyenne et la médiane sont égales
    e) la médiane avec l'EIQ décrit mieux les données que la moyenne ± ÉT
9. (QRM) Les graphiques adaptés à une variable numérique continue sont :
    a) l'histogramme
    b) la boîte à moustaches
    c) le diagramme tige-feuilles
    d) le diagramme circulaire
    e) le diagramme en points (dot plot) pour les petits échantillons
10. (QRM) Dans le jeu de données mg2026, l'IMC a une moyenne de 29,04 kg/m², une médiane de 28,25 kg/m², Q1 = 26,2, Q3 = 31,4 et un maximum de 48,3 kg/m². Les affirmations correctes sont :

a) l'écart interquartile est de 5,2 kg/m²

b) la limite supérieure selon la règle 1,5 x EIQ est de 39,2 kg/m²

c) la valeur de 48,3 kg/m² doit être exclue de l'analyse parce qu'elle est extrême

d) la distribution est asymétrique à droite

e) la présentation correcte dans le Tableau 1 est 28,25 (26,2-31,4) kg/m²

Chapitre 3. Les lois statistiques : loi normale, vérification de la normalité, asymétrie, aplatissement, score z et transformation des données

Objectifs

À la fin de ce chapitre, l'étudiant sera capable de :

- distinguer une variable aléatoire discrète d'une variable aléatoire continue et expliquer ce que décrivent la fonction de masse, la densité et la fonction de répartition ;
- calculer des probabilités simples avec la loi binomiale et avec la loi de Poisson, à partir d'exemples tirés du cabinet de médecine générale ;
- énumérer les propriétés de la loi normale et appliquer la règle 68,3-95,4-99,7 à des données réelles ;
- calculer un score z, trouver le percentile correspondant dans la table de l'annexe E et dans Google Sheets, et interpréter le résultat sur le plan clinique ;
- construire un intervalle de référence de type $m \pm 1{,}96s$ et expliquer pourquoi 5% des personnes en bonne santé se retrouvent "en dehors de la normale" ;
- vérifier la normalité d'une variable à l'aide de graphiques (histogramme, Q-Q, boîte à moustaches), d'indicateurs (asymétrie, aplatissement et leurs erreurs standard) et de tests (Shapiro-Wilk, Kolmogorov-Smirnov avec correction de Lilliefors, Anderson-Darling), en connaissant les limites de chaque méthode ;
- choisir une transformation adaptée (logarithmique, racine carrée, inverse, Box-Cox) et rétrotransformer correctement les résultats, y compris la moyenne géométrique et son intervalle de confiance ;
- repérer les valeurs extrêmes à l'aide du score z et de la règle 1,5 x EIQ, et décider de manière argumentée de ce qu'il faut en faire ;
- lire de manière critique la façon dont les articles publiés rapportent les variables symétriques et les variables asymétriques.

3.1 Variables aléatoires et lois de probabilité

Au chapitre 2, nous avons décrit les données telles qu'elles sont : tableaux de fréquences, moyennes, médianes, histogrammes. Faisons maintenant un pas en arrière et demandons-nous d'où viennent ces formes. Pourquoi l'histogramme de la pression systolique du jeu de données mg2026 ressemble-t-il à une cloche, alors que celui des triglycérides présente une longue "queue" vers la droite ? La réponse tient à la notion de distribution, c'est-à-dire à la loi selon laquelle se répartissent les valeurs d'une variable dans une population. Presque tout ce qui suit dans ce livre, des intervalles de confiance aux tests statistiques, repose sur des hypothèses relatives aux distributions ; il vaut donc la peine de bien les comprendre dès le départ [19,63].

3.1.1 De la fréquence observée à la probabilité

Lorsque nous disons que 29 des 160 patients de mg2026 ont un diabète confirmé par HGPO, nous décrivons l'échantillon : la fréquence relative est de 29/160 = 18,1%. Si nous imaginons choisir au hasard un patient de la même population d'adultes à risque cardiométabolique, la probabilité qu'il ait un diabète est à peu près la même, autour de 0,18. La probabilité est, en un sens pratique, la fréquence relative dont nous nous rapprochons lorsque le nombre d'observations devient très

grand. Une pièce lancée 10 fois peut donner 7 fois face ; lancée 10 000 fois, elle donnera une proportion très proche de 50%.

La probabilité prend des valeurs comprises entre 0 (événement impossible) et 1 (événement certain). Deux règles nous suffisent pour ce chapitre. La probabilité qu'un événement NE se produise PAS est égale à 1 moins la probabilité qu'il se produise. Pour des événements indépendants, la probabilité que les deux se produisent est le produit des probabilités : si deux patients sont choisis indépendamment, la probabilité qu'ils aient tous deux un diabète est de 0,18 x 0,18 = 0,0324.

3.1.2 Variables aléatoires discrètes et continues

Une variable aléatoire est une grandeur dont nous ne connaissons pas le résultat à l'avance, mais dont nous savons quelles valeurs elle peut prendre et avec quelles probabilités. Le nombre de patients diabétiques dans un groupe de 10 est une variable aléatoire discrète : il peut valoir 0, 1, 2, et ainsi de suite jusqu'à 10, mais il ne peut pas valoir 2,7. De même, le nombre d'effets indésirables déclarés en un mois dans un cabinet, le nombre de naissances par jour dans une maternité ou le nombre de colonies sur une boîte de Petri sont des variables discrètes (des dénombrements).

La pression artérielle, la glycémie, le poids ou le délai jusqu'à la thrombolyse sont des variables continues : entre deux valeurs possibles quelconques, il existe encore, en théorie, une infinité de valeurs. Le fait que le tensiomètre affiche des nombres entiers tient à l'instrument, et non au phénomène. Cette distinction compte, car la manière de calculer les probabilités diffère : pour les variables discrètes, nous additionnons les probabilités de valeurs ponctuelles ; pour les variables continues, nous calculons des aires sous une courbe.

3.1.3 Fonction de masse, densité et fonction de répartition

Pour une variable discrète, la loi se décrit par une fonction de masse (angl. probability mass function) : une liste qui indique, pour chaque valeur possible, sa probabilité. La somme de toutes ces probabilités vaut 1. Sa représentation graphique est un diagramme en barres, comme dans la figure 3.1.

Pour une variable continue, la probabilité d'une valeur exacte est nulle. Cela n'a pas de sens de demander quelle est la probabilité qu'un patient ait une PAS d'exactement 131,000... mmHg ; cela a un sens de demander quelle est la probabilité qu'il ait une PAS comprise entre 130 et 140 mmHg. C'est pourquoi nous utilisons la densité de probabilité (angl. probability density function), une courbe f(x) telle que l'aire sous la courbe entre deux valeurs a et b est la probabilité que la variable tombe dans l'intervalle [a, b]. L'aire totale sous la courbe vaut 1. L'histogramme d'un grand échantillon, tracé avec des fréquences relatives et des classes de plus en plus étroites, se rapproche de la densité.

La fonction de répartition (angl. cumulative distribution function), notée F(x), donne la probabilité que la variable soit inférieure ou égale à x :

$$F(x) = P(X \leq x)$$

Elle croît de 0 à 1 et correspond aux fréquences cumulées des tableaux de fréquences du chapitre 2. Lorsque nous lirons la table de la loi normale de l'annexe E, nous lirons en fait des valeurs de la fonction de répartition. C'est aussi de là que vient la notion de percentile : le 90e percentile d'une variable est la valeur x pour laquelle F(x) = 0,90.

3.2 Deux lois discrètes utiles en médecine

3.2.1 La loi binomiale

Supposons que, dans la population d'adultes à risque cardiométabolique dont provient le jeu de données mg2026, la prévalence du diabète non diagnostiqué soit de 18%. Un médecin de famille adresse pour une HGPO 10 patients choisis au hasard dans cette population. Combien auront un diabète ? Nous ne pouvons pas le savoir avec certitude, mais nous pouvons calculer la probabilité de chaque résultat possible.

La situation remplit les conditions de la loi binomiale : un nombre fixe d'épreuves (n = 10 patients), chacune avec deux issues possibles (diabète ou non), la même probabilité de "succès" à chaque épreuve (p = 0,18) et l'indépendance entre les épreuves. La probabilité d'obtenir exactement k cas est :

$$P(X = k) = C(n,\ k) \times p^k \times (1 - p)^{n-k}$$

où $C(n,\ k) = \frac{n!}{k! \times (n-k)!}$ est le nombre de façons de choisir les k patients parmi n. La moyenne et la variance de la loi binomiale ont des formules très simples :

moyenne = n x p variance = n x p x (1 - p)

Dans notre exemple, la moyenne est de 10 x 0,18 = 1,8 patient, la variance de 10 x 0,18 x 0,82 = 1,476 et l'écart-type de $\sqrt{1,476}$ = 1,21 patient. Les probabilités individuelles sont données dans le tableau 3.1.

Tableau 3.1. Loi binomiale pour n = 10 patients et une prévalence p = 0,18.

k (patients diabétiques)	**0**	**1**	**2**	**3**	**4**	**5**	**6 ou plus**
$P(X = k)$	0,1374	0,3017	0,2980	0,1745	0,0670	0,0177	0,0037
P(X ≤ k), cumulée	0,1374	0,4392	0,7372	0,9117	0,9787	0,9963	1,0000

Vérifions la première valeur à la main : $P(X = 0) = 0{,}82^{10} = 0{,}1374$. Autrement dit, dans près d'un groupe de 10 patients sur sept, nous ne trouverons aucun diabétique, bien que la prévalence soit de 18%. La probabilité de trouver au moins un cas est de 1 - 0,1374 = 0,8626, et la probabilité de trouver trois cas ou plus est de 1 - 0,7372 = 0,2628. Ces chiffres expliquent une observation que fait tout jeune médecin : dans les petits groupes, les fréquences observées "sautent" beaucoup autour de la valeur attendue, sans que cela signifie quoi que ce soit de particulier.

Appliquons la même logique à l'ensemble de l'échantillon mg2026. Si la prévalence réelle était de 18%, le nombre attendu de diabétiques parmi 160 patients serait de 160 x 0,18 = 28,8, avec un écart-type de $\sqrt{160 \times 0{,}18 \times 0{,}82} = 4{,}86$. Nous en avons observé 29, soit pratiquement la valeur attendue. La probabilité d'en observer 35 ou plus serait de 0,122 ; une telle valeur n'aurait donc pas été surprenante non plus. Remarquez un point important pour les chapitres suivants : lorsque n est grand, la loi binomiale ressemble de plus en plus à une loi normale de moyenne n x p et d'écart-type $\sqrt{n \times p \times (1 - p)}$. Le calcul approché avec la loi normale (avec une petite correction de continuité) donne 0,120 pour la même question, très proche de la valeur exacte. C'est sur cette approximation que repose l'intervalle de confiance de Wald d'une proportion, que nous discuterons au chapitre 4.

3.2.2 La loi de Poisson

Dans un cabinet de médecine générale, on déclare en moyenne deux effets indésirables médicamenteux par mois. Au cours d'un mois donné, il peut y en avoir zéro, un, trois ou six. Ici,

nous n'avons pas de nombre fixe d'épreuves : le nombre de patients exposés est grand, et la probabilité qu'un patient donné présente un effet indésirable au cours d'un mois donné est très faible. Pour de tels événements rares, comptés dans un intervalle de temps ou d'espace, on utilise la loi de Poisson :

$$P(X = k) = \frac{e^{-\lambda} \times \lambda^{k}}{k!}$$

où λ (lambda) est le nombre moyen d'événements par intervalle, et e = 2,718... la base du logarithme naturel. La loi de Poisson possède une propriété inhabituelle : sa moyenne et sa variance sont égales, toutes deux valant λ. Pour λ = 2, l'écart-type est √2 = 1,41.

Tableau 3.2. Loi de Poisson pour λ = 2 effets indésirables par mois.

k (effets en un mois)	0	1	2	3	4	5 ou plus
$P(X = k)$	0,1353	0,2707	0,2707	0,1804	0,0902	0,0527

La probabilité d'un mois sans aucun effet indésirable est $e^{-2} = 0{,}1353$, soit environ un mois sur sept. La probabilité d'avoir quatre effets ou plus est de 0,0902 + 0,0527 = 0,1429, et cinq ou plus surviennent au cours de 5,3% des mois. Si le médecin tirait la sonnette d'alarme chaque fois qu'il compte quatre effets dans un mois, il déclencherait de fausses alertes près d'un mois sur sept, même si rien n'a changé. Le seuil de cinq effets donne de fausses alertes environ un mois sur vingt. C'est exactement le type de raisonnement qu'utilisent les systèmes de pharmacovigilance et de surveillance épidémiologique : un nombre observé ne devient un "signal" que lorsqu'il est improbable dans le régime habituel.

La loi de Poisson est en fait le cas limite de la loi binomiale lorsque n est très grand et p très petit, avec n x p = λ. Si 500 patients reçoivent un médicament, chacun ayant une probabilité de 0,004 de présenter un effet donné au cours d'un mois, la loi binomiale donne P(X = 0) = 0,1348, et la loi de Poisson de paramètre λ = 2 donne 0,1353. La différence est négligeable. Sur une année entière, avec λ = 24, la probabilité de compter 30 effets ou plus est de 0,132 ; remarquez que, pour λ grand, la loi de Poisson devient elle aussi presque symétrique et se rapproche d'une loi normale. La figure 3.1 compare les deux lois discrètes.

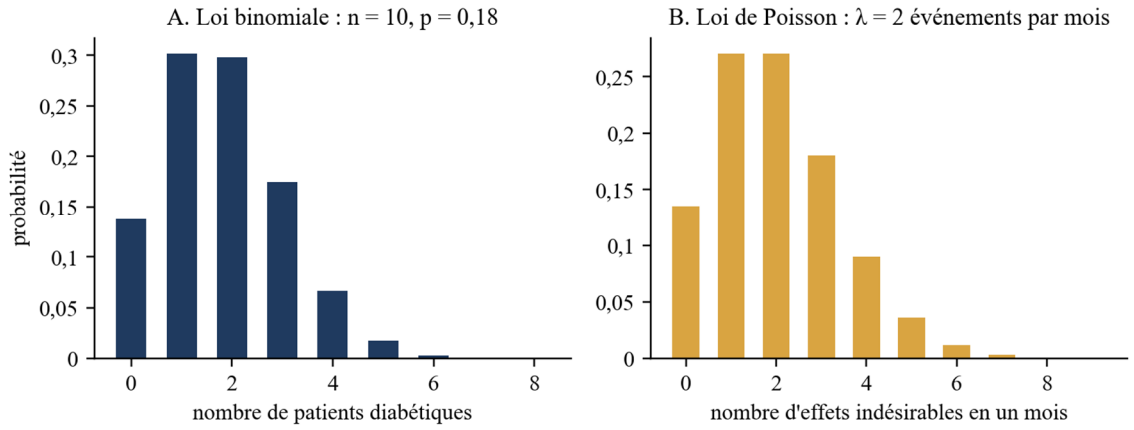


Figure 3.1. Lois discrètes. A : loi binomiale du nombre de patients diabétiques dans un groupe de 10, pour une prévalence de 18% ; la moyenne est de 1,8. B : loi de Poisson du nombre mensuel d'effets indésirables dans un cabinet, avec λ = 2. Les deux sont asymétriques à droite. Valeurs théoriques.

3.3 La loi normale

3.3.1 Pourquoi elle apparaît si souvent

La loi normale, appelée aussi loi de Gauss ou de Laplace-Gauss, décrit des variables continues qui résultent de la somme de nombreuses influences petites et indépendantes. La taille d'un adulte

dépend de dizaines de gènes, de l'alimentation pendant l'enfance, des maladies, du milieu social ; la pression artérielle dépend de facteurs génétiques, du poids, du sel, du stress, du moment de la mesure, de la technique de l'infirmière. Lorsque de nombreuses petites causes s'additionnent, le résultat tend à se distribuer de façon symétrique, en forme de cloche. L'explication mathématique est le théorème central limite, que nous rencontrerons à la section 3.7.

Il faut dire d'emblée que de nombreuses variables biologiques ne sont PAS normales. Les concentrations sériques (triglycérides, enzymes hépatiques, CRP, insuline), les durées (hospitalisation, durée des interventions), les titres d'anticorps, le nombre de jours d'arrêt maladie présentent en règle générale une asymétrie positive. Aucune loi de la nature n'oblige les données médicales à être normales. La loi normale est un modèle, souvent bon, parfois mauvais, et c'est précisément pour cette raison qu'il faut le vérifier à chaque fois [42,53].

3.3.2 Les propriétés de la loi normale

La densité de la loi normale a pour formule :

$$f(x) = \left(\frac{1}{\sigma \times \sqrt{2\pi}}\right) \times e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Vous n'avez pas besoin de la mémoriser. Vous devez cependant retenir ses conséquences :

- La courbe est symétrique par rapport à la moyenne μ. La moitié de l'aire se trouve à gauche de la moyenne, l'autre moitié à droite.
- La moyenne, la médiane et le mode coïncident. Dans un échantillon réel, ils seront seulement proches : pour la PAS de mg2026, la moyenne est de 131,13 mmHg et la médiane de 131 mmHg.
- La loi est entièrement déterminée par deux paramètres : la moyenne μ, qui donne la position, et l'écart-type σ, qui donne la largeur. Deux lois normales de même moyenne et de σ différents ont le même centre, mais l'une est haute et étroite, l'autre basse et large.
- L'aire totale sous la courbe vaut 1.
- Les "queues" de la courbe se rapprochent de l'axe horizontal à mesure que l'on s'éloigne de la moyenne, mais ne le touchent jamais. En théorie, toute valeur est possible, aussi grande ou aussi petite soit-elle, mais les valeurs très éloignées de la moyenne ont des probabilités extrêmement faibles.
- Les points d'inflexion de la courbe se situent en μ - σ et μ + σ.

Cette dernière propriété des queues a une conséquence pratique que les étudiants découvrent d'eux-mêmes lorsqu'ils appliquent le modèle normal à des variables qui ne peuvent pas être négatives. Une loi normale de moyenne 11,6 jours et d'écart-type 7,0 jours prédirait qu'environ 5% des patients restent hospitalisés un nombre négatif de jours, et 2,3% même moins de -2,4 jours, ce qui est absurde. Nous reviendrons à la section 3.8 sur cet exemple, tiré d'un article publié.

3.3.3 La règle 68,3%-95,4%-99,7%

Pour toute loi normale, quels que soient μ et σ, la proportion des valeurs situées à une certaine distance de la moyenne, mesurée en écarts-types, est la même (figure 3.2) :

- entre μ - σ et μ + σ se trouvent 68,3% des valeurs ;
- entre μ - 2σ et μ + 2σ se trouvent 95,4% des valeurs ;
- entre μ - 3σ et μ + 3σ se trouvent 99,7% des valeurs.

Deux valeurs dérivées sont utilisées tout aussi souvent : entre μ - 1,96σ et μ + 1,96σ se trouvent exactement 95% des valeurs, et entre μ - 2,576σ et μ + 2,576σ, 99%. La différence entre 2 et 1,96 semble une subtilité, mais vous la rencontrerez dans chaque intervalle de confiance à 95%.

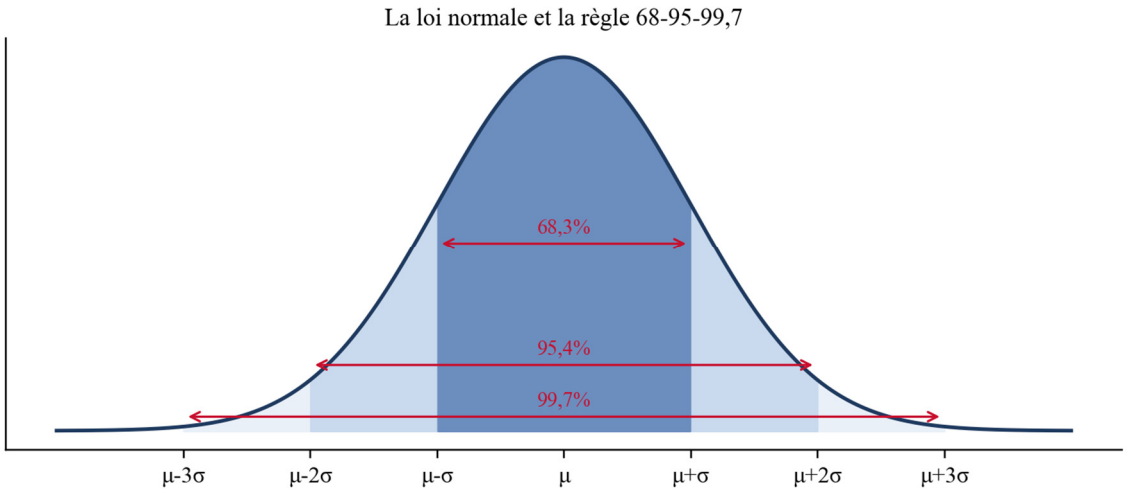


Figure 3.2. La loi normale et la règle 68-95-99,7 : l'aire comprise entre μ - σ et μ + σ représente 68,3% de l'aire totale, celle comprise entre μ - 2σ et μ + 2σ 95,4%, et celle comprise entre μ - 3σ et μ + 3σ 99,7%. Courbe théorique.

Vérifions la règle sur le jeu de données mg2026. La pression systolique initiale a une moyenne de 131,13 mmHg et un écart-type de 13,88 mmHg. L'intervalle $m \pm s$ est de 117,3-145,0 mmHg et contient 70,0% des 160 patients (en théorie, 68,3%). L'intervalle $m \pm 2s$, soit 103,4-158,9 mmHg, en contient 93,8% (en théorie, 95,4%). L'intervalle $m \pm 3s$, 89,5-172,8 mmHg, les contient tous (en théorie, 99,7%). L'ajustement n'est pas parfait et ne peut pas l'être, car 160 patients constituent un échantillon, et non la population. Il est toutefois assez bon pour que nous puissions utiliser le modèle normal pour la PAS. Attention à la notation : μ et σ sont les paramètres de la population, que nous ne connaissons pas ; m et s sont leurs estimations tirées de l'échantillon. En pratique, nous remplaçons les paramètres par les estimations.

3.3.4 Proportion observée et proportion prédite par le modèle normal

Un test simple et très instructif du modèle normal consiste à comparer la proportion observée de patients au-delà d'un seuil clinique à la proportion prédite par une loi normale ayant la moyenne et l'écart-type de l'échantillon. Pour une PAS ≥ 140 mmHg, le seuil classique de l'hypertension, le score z est (140 - 131,13) / 13,88 = 0,64. D'après la table E.1, Φ(0,64) = 0,7389, donc la proportion théorique au-delà du seuil est de 1 - 0,7389 = 0,261, soit 26,1%. Dans l'échantillon, 41 des 160 patients ont une PAS ≥ 140, soit 25,6%. La différence d'un demi-point de pourcentage est négligeable.

Le même calcul, répété pour d'autres variables, nous en dit long sur la forme de la distribution (tableau 3.3).

Tableau 3.3. Proportions observées et proportions prédites par le modèle normal (jeu de données mg2026, n = 160).

Variable et seuil	Score z du seuil	Proportion théorique	Proportion observée	Commentaire
PAS ≥ 140 mmHg	0,64	26,1%	25,6%	très bon ajustement
PAS < 120 mmHg	-0,80	21,1%	20,0%	très bon ajustement
Âge ≥ 65 ans	0,99	16,0%	16,3%	très bon ajustement
Cholestérol ≥ 240 mg/dL	0,83	20,3%	23,8%	acceptable
Glycémie ≥ 126 mg/dL	1,30	9,6%	11,3%	queue droite un peu plus épaisse

Variable et seuil	Score z du seuil	Proportion théorique	Proportion observée	Commentaire
$\text{IMC} \geq \frac{30\ \text{kg}}{m^2}$	0,25	40,3%	33,8%	modèle inadapté
$\text{IMC} \geq \frac{40\ \text{kg}}{m^2}$	2,80	0,3%	1,9%	modèle inadapté

Pour l'IMC, le modèle normal se trompe dans les deux sens : il surestime la proportion d'obèses (40,3% contre 33,8%), car la moyenne est tirée vers le haut par quelques valeurs très élevées, et il sous-estime fortement la proportion de personnes atteintes d'obésité de grade III (0,3% contre 1,9%, soit 3 patients observés contre moins de 1 prédit). C'est la signature typique d'une distribution asymétrique à droite, à longue queue. Pour la pression artérielle et l'âge, le modèle fonctionne très bien. Nous verrons à la section 3.8 que les tests formels de normalité aboutissent exactement à la même conclusion.

3.4 La loi normale centrée réduite et le score z

3.4.1 La standardisation

Il existe une infinité de lois normales, une pour chaque couple (μ, σ). Il serait impossible de disposer d'une table pour chacune. La solution consiste à les ramener toutes à une seule, la loi normale centrée réduite, de moyenne 0 et d'écart-type 1, notée N(0, 1). Cette transformation s'appelle la standardisation (ou centrage-réduction) et produit le score z :

$$z = \frac{x - \mu}{\sigma}$$

ou, avec les estimations tirées de l'échantillon, $z = \frac{x-m}{s}$. Le score z indique combien d'écarts-types séparent une valeur de la moyenne, et dans quel sens. Un z positif correspond à une valeur supérieure à la moyenne, un z négatif à une valeur inférieure. Le score z n'a pas d'unité de mesure : les mmHg s'annulent. C'est pourquoi nous pouvons comparer les scores z de variables complètement différentes. Un patient avec z = 2,1 pour la glycémie et z = 0,3 pour le cholestérol est "plus inhabituel" pour la glycémie que pour le cholestérol, même si les deux se mesurent sur des échelles différentes.

Si une variable est normale, ses scores z suivent la loi normale centrée réduite. La moyenne des scores z d'un échantillon est toujours 0, et leur écart-type 1, même si la variable n'est pas normale. La standardisation ne modifie cependant pas la forme de la distribution : une variable asymétrique reste tout aussi asymétrique après transformation en scores z. C'est une erreur fréquente de croire que "la standardisation normalise les données".

3.4.2 Un exemple complet : une PAS de 160 mmHg

Un patient de mg2026 a une PAS de 160 mmHg. À quel point cette valeur est-elle inhabituelle dans l'échantillon ?

$$z = \frac{160 - 131{,}13}{13{,}88} = \frac{28{,}87}{13{,}88} = 2{,}08$$

La valeur se situe à 2,08 écarts-types au-dessus de la moyenne (figure 3.3). Dans la table E.1 de l'annexe E, nous cherchons la ligne 2,0 et la colonne 0,08 et lisons Φ(2,08) = 0,9812. Cela signifie que, selon le modèle normal, 98,1% des patients ont une PAS inférieure ou égale à 160 mmHg : le patient se situe au 98e percentile. La proportion de ceux qui ont des valeurs plus élevées est de 1 - 0,9812 = 0,0188, soit 1,9%, et l'aire comprise entre la moyenne et 160 mmHg est de 0,9812 - 0,5 = 0,4812.

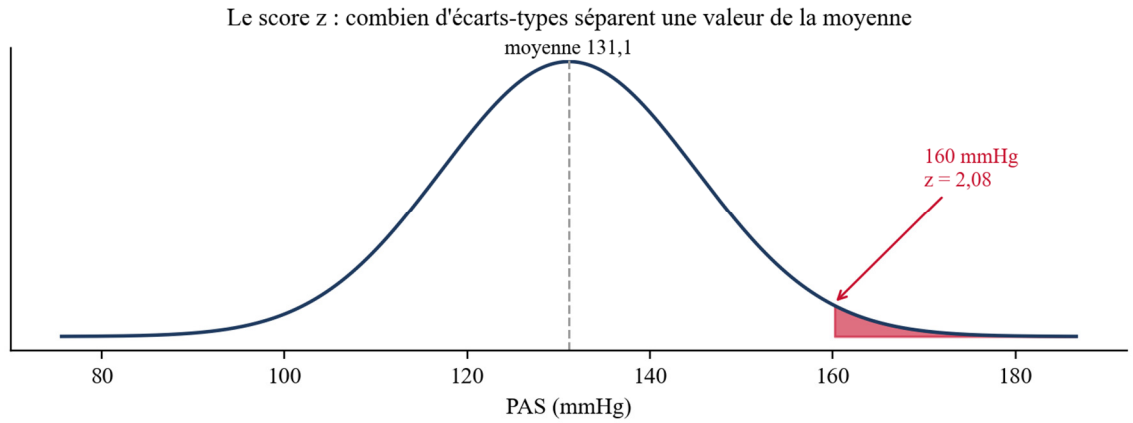


Figure 3.3. Le score z d'une pression systolique de 160 mmHg dans le jeu de données mg2026 (moyenne 131,1 mmHg, écart-type 13,9 mmHg) est de 2,08. L'aire colorée à droite de la valeur représente la proportion théorique de patients ayant une PAS plus élevée, environ 1,9%. Données synthétiques (jeu de données mg2026).

Pour les valeurs inférieures à la moyenne, nous utilisons la symétrie de la courbe. Un patient dont la PAS est de 108 mmHg a $z = \frac{108-131,13}{13,88} = -1,67$. La table donne Φ(1,67) = 0,9525, et par symétrie Φ(-1,67) = 1 - 0,9525 = 0,0475. Le patient se situe approximativement au 5e percentile. Dans l'échantillon, 6,9% des patients ont une PAS ≤ 108 mmHg, là encore près de la valeur théorique.

3.4.3 Le problème inverse : du pourcentage à la valeur

La question est parfois inverse : quelle est la valeur de PAS au-dessous de laquelle se trouvent 90% des patients ? Nous cherchons dans la table le z pour lequel Φ(z) = 0,90 et trouvons z = 1,28 (plus précisément 1,2816, valeur donnée dans la table E.2). Nous refaisons ensuite le chemin inverse :

$$x = m + z \times s = 131,13 + 1,28 \times 13,88 = 148,9 \text{ mmHg}$$

Le 90e percentile observé dans l'échantillon est de 148,2 mmHg, et la valeur théorique de 148,9 mmHg. Les valeurs de z qu'il vaut la peine de connaître par cœur sont peu nombreuses : 1,28 (90e percentile), 1,645 (95e percentile), 1,96 (97,5e percentile), 2,33 (99e percentile) et 2,576 (99,5e percentile). Vous les utiliserez pour les intervalles de confiance, pour les tests unilatéraux et bilatéraux et pour le calcul de la taille de l'échantillon.

3.4.4 La loi normale dans Google Sheets

Dans les feuilles de calcul, la table n'est plus nécessaire. Google Sheets propose quatre fonctions pour la loi normale, plus la fonction de standardisation [14,43]. Avec les paramètres régionaux France (ou Roumanie), le séparateur d'arguments est ";" et la virgule est le séparateur décimal.

Tableau 3.4. Les fonctions de Google Sheets pour la loi normale, avec des exemples sur la PAS de mg2026.

Fonction	Ce qu'elle calcule	Exemple	Résultat
=STANDARDIZE(x; moyenne; ÉT)	le score z	=STANDARDIZE(160; 131,13; 13,88)	2,08
=NORM.DIST(x; moyenne; ÉT; TRUE)	l'aire à gauche de x (fonction de répartition)	=NORM.DIST(160; 131,13; 13,88; TRUE)	0,981
=NORM.DIST(x; moyenne; ÉT; FALSE)	la hauteur de la courbe (densité), rarement utilisée	=NORM.DIST(131; 131,13; 13,88; FALSE)	0,029
=NORM.S.DIST(z; TRUE)	l'aire à gauche de z sous la loi normale centrée réduite	=NORM.S.DIST(2,08; TRUE)	0,9812
=NORM.INV(p; moyenne;	la valeur x au-dessous de	=NORM.INV(0,9; 131,13;	148,9

Fonction	Ce qu'elle calcule	Exemple	Résultat
ÉT)	laquelle se trouve la proportion p	13,88)	
=NORM.S.INV(p)	la valeur z au-dessous de laquelle se trouve la proportion p	=NORM.S.INV(0,975)	1,96

En pratique, on remplace les valeurs numériques par des références : =STANDARDIZE(M2; AVERAGE($M$2:$M$161); STDEV($M$2:$M$161)) calcule le score z du premier patient et peut être recopiée sur toute la colonne. Le dernier argument de la fonction NORM.DIST est un piège classique en travaux pratiques : avec TRUE, nous obtenons la probabilité cumulée (qui nous intéresse presque toujours) ; avec FALSE, nous obtenons la densité, un nombre qui n'est pas une probabilité et qui n'a pas d'interprétation directe. Les fonctions NORMDIST et NORMSDIST des anciennes versions d'Excel existent également, avec la même logique.

3.4.5 Le score z en clinique

Les scores z ne sont pas un simple exercice de travaux dirigés. Les courbes de croissance pédiatriques expriment le poids et la taille de l'enfant sous forme de score z (ou de percentile) par rapport à une population de référence de même âge et de même sexe ; un enfant dont la taille est à z = -2 se situe au 2,3e percentile. L'ostéodensitométrie rapporte le T-score, qui est un score z calculé par rapport à la moyenne des jeunes adultes, et le Z-score, calculé par rapport à des personnes du même âge. L'échocardiographie pédiatrique exprime les dimensions des cavités et des vaisseaux sous forme de scores z rapportés à la surface corporelle. Dans toutes ces situations, le médecin compare le patient à une distribution de référence, et la question est toujours la même : à quelle distance du centre se trouve sa valeur, mesurée en écarts-types ?

3.5 Les intervalles de référence en biologie médicale

3.5.1 D'où viennent les "valeurs normales"

Sur le compte rendu d'analyses, à côté de chaque résultat, figure un intervalle de référence, couramment appelé "valeurs normales". Il est construit, en règle générale, sur un échantillon de personnes en bonne santé, comme l'intervalle qui contient les 95% centraux de leurs valeurs [64]. Si la variable est approximativement normale, l'intervalle est :

$$m - 1{,}96 \times s \; ; \; m + 1{,}96 \times s$$

Si la variable n'est pas normale, on utilise directement les 2,5e et 97,5e percentiles de l'échantillon de référence (méthode non paramétrique), ou bien on transforme les données, on calcule l'intervalle sur l'échelle transformée et on rétrotransforme les bornes [42,53].

La conséquence logique de cette construction en surprend plus d'un : par définition, 5% des personnes en parfaite santé ont une valeur "en dehors de la normale", 2,5% au-dessous de la borne inférieure et 2,5% au-dessus de la borne supérieure. Un résultat légèrement en dehors de l'intervalle ne signifie pas une maladie. Il signifie seulement que la valeur est inhabituelle par rapport à la population de référence.

Le problème s'aggrave lorsqu'on prescrit de nombreuses analyses. Si les k analyses étaient indépendantes, la probabilité qu'une personne en bonne santé ait tous ses résultats dans l'intervalle ar fi $0{,}95^k$. Pour 5 analyses, 0,95^5 = 0,774, donc 22,6% des personnes en bonne santé auraient au moins un résultat "anormal". Pour 12 analyses, la proportion monte à 46%, et pour 20 analyses à 64%. Les analyses réelles sont corrélées entre elles, de sorte que les chiffres exacts diffèrent, mais l'idée demeure : un "dépistage" portant sur des dizaines de paramètres trouvera presque chez chaque personne en bonne santé quelque chose à explorer. C'est l'une des raisons pour lesquelles on prescrit des analyses en ayant une question clinique en tête, et non "en vrac".

3.5.2 Un intervalle à 95% sur le jeu de données mg2026

Construisons l'intervalle $m \pm 1{,}96s$ pour la PAS de mg2026 :

borne inférieure = 131,13 - 1,96 x 13,88 = 131,13 - 27,20 = 103,9 mmHg borne supérieure = 131,13 + 1,96 x 13,88 = 131,13 + 27,20 = 158,3 mmHg

L'intervalle 103,9-158,3 mmHg contient 150 des 160 patients, soit 93,8%, près de 95%. Les 2,5e et 97,5e percentiles calculés directement à partir des données sont 103,0 et 156,1 mmHg, très proches des bornes obtenues avec le modèle normal. Lorsque la variable est normale, les deux méthodes donnent des résultats semblables, et la méthode paramétrique a l'avantage d'utiliser toutes les données, et pas seulement les quelques valeurs extrêmes.

Une précision de fond s'impose : les patients de mg2026 sont à risque cardiométabolique ; ils ne forment donc pas une population de référence en bonne santé. L'intervalle calculé décrit cet échantillon et nous sert d'exercice, non d'intervalle de référence pour la pression artérielle. Les valeurs cibles de PA ne se définissent pas statistiquement, mais en fonction du risque cardiovasculaire associé, dans les recommandations de pratique clinique. La distinction entre "normal statistique" (fréquent) et "normal clinique" (sans risque accru) mérite d'être retenue : une cholestérolémie de 220 mg/dL peut être très fréquente dans une population et néanmoins associée à un risque accru.

3.5.3 Ce qui se passe avec une variable asymétrique

Pour les triglycérides (moyenne 161,18 mg/dL, écart-type 76,30 mg/dL), la formule $m \pm 1{,}96s$ donne 11,6-310,7 mg/dL. La borne inférieure n'a aucun sens clinique : personne dans l'échantillon n'a des triglycérides inférieurs à 44 mg/dL, et 11,6 mg/dL est une valeur pratiquement inexistante chez l'adulte. À droite, 6 patients dépassent la borne supérieure, tous dans la longue queue. L'intervalle contient 96,3% des données, mais il est décalé : trop large vers le bas, trop étroit vers le haut.

Si nous passons aux logarithmes (section 3.9), la moyenne des logarithmes est de 4,970 et leur écart-type de 0,488. L'intervalle sur l'échelle logarithmique est $4{,}970 \pm 1{,}96 \times 0{,}488$ et, après retransformation par la fonction exponentielle, nous obtenons 55,3-374,9 mg/dL. Les percentiles 2,5 et 97,5 observés sont de 53,8 et 364,0 mg/dL, presque identiques aux limites obtenues sur l'échelle logarithmique. L'intervalle retransformé n'est plus symétrique par rapport à la moyenne, et c'est correct : il reflète la forme réelle de la distribution.

> **Étude de cas : seuils sur des distributions continues en élastographie hépatique**
>
> Dans le diagnostic non invasif des maladies hépatiques, les paramètres mesurés sont continus, alors que la décision clinique exige des seuils. Moga et ses collaborateurs ont évalué l'élastographie time-harmonic par rapport à l'élastographie impulsionnelle chez 165 patients et ont proposé deux seuils pour la cirrhose : un seuil d'exclusion (en anglais rule-out) inférieur à 1,61 m/s (7,77 kPa), avec une sensibilité de 90,7% et une spécificité de 66,6%, et un seuil de confirmation (en anglais rule-in) supérieur à 1,83 m/s (10 kPa), avec une sensibilité de 65,1% et une spécificité de 96,7% ; l'aire sous la courbe ROC était de 0,90 (IC à 95% 0,82-0,93) [3]. Deux seuils au lieu d'un ne sont pas une complication inutile : ils reconnaissent que les distributions des valeurs chez les patients avec et sans cirrhose se chevauchent. Sous le premier seuil se trouvent surtout des patients sans cirrhose, au-dessus du second surtout des patients cirrhotiques, et entre les deux subsiste une "zone grise" où la décision requiert d'autres investigations. Nica et ses collaborateurs ont procédé de façon semblable pour la fraction de graisse mesurée par échographie (USFF), chez 169 patients, avec des seuils supérieurs à 5% pour une stéatose au moins légère (AUROC 0,90), à 7,5% pour une stéatose modérée (AUROC 0,92) et à 10% pour une stéatose sévère (AUROC 0,91), en prenant comme référence le CAP,

> avec des seuils de 248, 268 et 280 dB/m [65]. Ce que les résumés ne nous disent pas : la forme des distributions dans chaque groupe et la proportion de patients situés dans la zone grise, deux informations utiles pour juger de l'applicabilité des seuils dans un autre centre. Nous les reprendrons au chapitre 12, avec les courbes ROC.

3.6 Les distributions d'échantillonnage : t, khi-deux et F

Jusqu'ici, nous avons parlé de la distribution des valeurs individuelles : la façon dont se répartissent les pressions artérielles des patients. La statistique inférentielle utilise pourtant un autre type de distribution, tout aussi important : la distribution d'une statistique calculée sur l'échantillon, par exemple la moyenne, si nous répétions l'étude un très grand nombre de fois. On l'appelle distribution d'échantillonnage (en anglais sampling distribution). Trois distributions d'échantillonnage reviennent sans cesse en biostatistique ; nous les présentons ici uniquement par leur forme, pour que vous puissiez les reconnaître dans les chapitres suivants (figure 3.4).

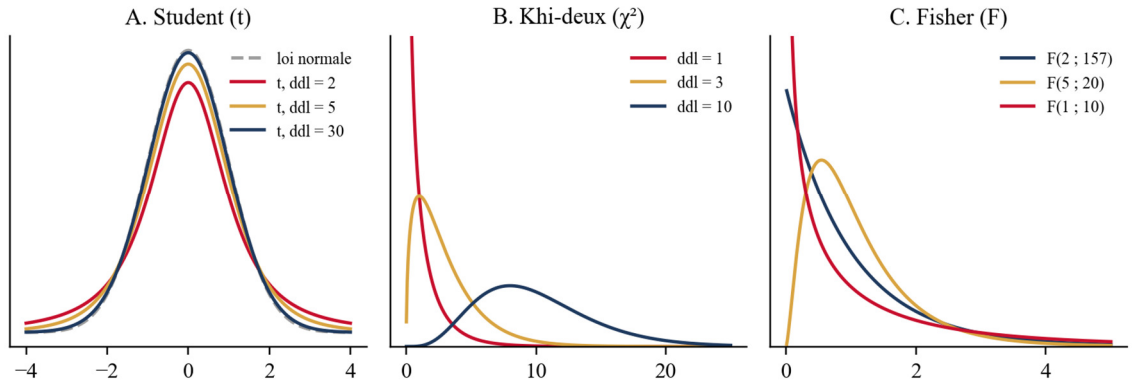


Figure 3.4. Trois distributions d'échantillonnage. A : la loi de Student (t) à 2, 5 et 30 degrés de liberté, comparée à la loi normale centrée réduite (ligne discontinue) ; moins il y a de ddl, plus les queues sont épaisses. B : la loi du khi-deux à 1, 3 et 10 degrés de liberté, asymétrique vers la droite. C : la loi de Fisher (F) pour quelques couples de degrés de liberté. Courbes théoriques.

La loi de Student (t) a été publiée par W. S. Gosset sous le pseudonyme "Student". Elle apparaît lorsque nous standardisons une moyenne en utilisant l'écart-type estimé sur l'échantillon, s, au lieu de l'écart-type réel σ, que nous ne connaissons pas. Elle a une forme en cloche, elle est symétrique par rapport à 0, mais ses queues sont plus épaisses que celles de la loi normale : estimer σ à partir de peu de données ajoute de l'incertitude. Sa forme dépend d'un seul paramètre, le nombre de degrés de liberté (ddl), qui vaut n - 1 pour une moyenne. Quand les ddl augmentent, la loi t se rapproche de la loi normale centrée réduite. La valeur critique bilatérale pour 95% est de 2,776 à 4 ddl, 2,262 à 9 ddl, 2,045 à 29 ddl, 1,975 à 159 ddl et 1,96 à l'infini (annexe E, tableau E.3). Ces valeurs remplaceront 1,96 dans les intervalles de confiance pour les petits échantillons (chapitre 4) et dans les tests t (chapitre 6).

Les degrés de liberté méritent une explication intuitive. Si nous connaissons la moyenne de 10 valeurs et 9 d'entre elles, la dixième est déterminée : elle n'est plus "libre". Lorsque nous calculons l'écart-type autour de la moyenne de l'échantillon, nous utilisons déjà une information extraite des données (la moyenne), de sorte qu'il reste n - 1 valeurs libres. C'est de là que vient le dénominateur n - 1 de la formule de l'écart-type, dont nous parlerons au chapitre 4.

La loi du khi-deux (χ^2) apparaît lorsque nous additionnons les carrés de variables normales centrées réduites. Elle ne prend que des valeurs positives, elle est asymétrique vers la droite, et son asymétrie diminue quand les ddl augmentent. Nous l'utiliserons pour le test du khi-deux sur les tableaux de contingence (chapitre 7), où ddl = (r - 1) x (c - 1), et pour l'intervalle de confiance d'une variance. La valeur critique pour $\alpha = 0{,}05$ est de 3,84 à 1 ddl, 5,99 à 2 ddl et 7,81 à 3 ddl.

La loi de Fisher (F) est le rapport de deux variables du khi-deux, chacune divisée par ses degrés de liberté. Elle a deux nombres de degrés de liberté, l'un pour le numérateur et l'autre pour le dénominateur, ne prend que des valeurs positives et est asymétrique vers la droite. Nous l'utiliserons pour l'analyse de la variance (ANOVA, chapitre 6), pour le test de Levene d'égalité des variances et en régression. Par exemple, pour une ANOVA à trois bras et 160 patients, les ddl sont 2 et 157, et la valeur critique pour $\alpha = 0{,}05$ est de 3,05.

Vous n'avez pas à calculer ces distributions à la main. Vous devez savoir ce qu'est chacune d'elles, dans quel test elle intervient et que les valeurs critiques diminuent à mesure que les degrés de liberté augmentent.

3.7 Le théorème central limite

3.7.1 L'énoncé, sans formules

Le théorème central limite est probablement le résultat le plus important de la statistique appliquée. En termes simples : si nous tirons des échantillons de taille n d'une population de moyenne μ et d'écart-type σ, quelle que soit la forme de la distribution de la population, les moyennes de ces échantillons ont une distribution qui :

- est centrée sur μ (les moyennes des échantillons ne surestiment ni ne sous-estiment μ de façon systématique) ;
- a pour écart-type $\frac{\sigma}{\sqrt{n}}$, appelé erreur standard de la moyenne ;
- se rapproche d'une distribution normale à mesure que n augmente.

Le dernier point est la partie surprenante. La population peut être asymétrique, bimodale, de n'importe quelle forme : les moyennes des échantillons tendront malgré tout vers une courbe en cloche. Plus la population est asymétrique, plus il faut un n élevé. Pour des distributions modérément asymétriques, 30 à 50 observations suffisent généralement ; pour des distributions très asymétriques (par exemple les coûts d'hospitalisation, où quelques patients coûtent des dizaines de fois plus que les autres), il peut en falloir des centaines.

3.7.2 Une simulation sur les triglycérides

Pour voir le théorème à l'œuvre, nous avons traité les 160 valeurs de triglycérides de mg2026 comme une "population" et nous y avons tiré, par ordinateur, 10 000 échantillons aléatoires (avec remise) de taille 2, 5, 10, 30 et 50. Pour chaque échantillon, nous avons calculé la moyenne, puis examiné la distribution des 10 000 moyennes (figure 3.5 et tableau 3.5).

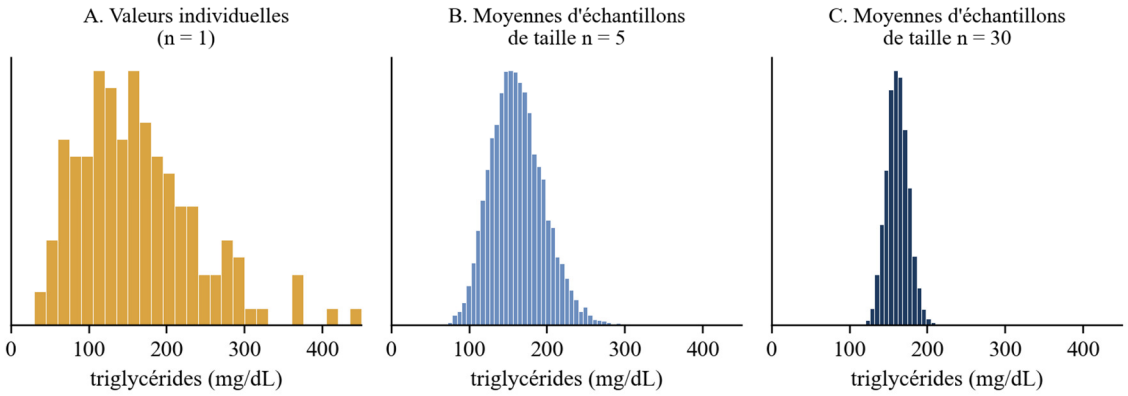


Figure 3.5. Le théorème central limite appliqué aux triglycérides. A : la distribution des valeurs individuelles, asymétrique vers la droite (asymétrie 0,95). B : la distribution des moyennes de 10 000 échantillons de 5 patients, plus symétrique. C : la distribution des moyennes des échantillons de 30 patients, presque normale et beaucoup plus étroite. Simulation sur le jeu de données mg2026.

Tableau 3.5. Simulation du théorème central limite : 10 000 échantillons tirés des valeurs de triglycérides (mg2026).

Taille de l'échantillon n	**Moyenne des 10 000 moyennes (mg/dL)**	**Écart-type des moyennes (mg/dL)**	$\frac{\sigma}{\sqrt{n}}$ théorique (mg/dL)	**Asymétrie des moyennes**
1 (valeurs individuelles)	161,2	76,3	-	0,95
2	162,2	54,3	53,8	0,65
5	160,6	33,7	34,0	0,39
10	161,1	24,0	24,1	0,25
30	161,3	14,1	13,9	0,21
50	161,2	10,8	10,8	0,19

Trois choses sautent aux yeux. La moyenne des moyennes reste autour de 161 mg/dL, c'est-à-dire la moyenne de la "population", quel que soit n. L'écart-type des moyennes diminue exactement comme le prédit la formule $\frac{\sigma}{\sqrt{n}}$: pour n = 30, il est de 14,1 mg/dL, contre 76,3 mg/dL pour les valeurs individuelles. L'asymétrie passe de 0,95 à environ 0,2. Les moyennes des échantillons de 30 patients sont déjà presque normales, alors que les triglycérides individuels ne le sont pas.

3.7.3 Pourquoi c'est important pour le médecin

Le théorème central limite explique pourquoi le test t et l'intervalle de confiance de la moyenne fonctionnent bien même lorsque les données ne sont pas parfaitement normales, à condition que l'échantillon soit suffisamment grand. Il explique aussi pourquoi une moyenne calculée sur 160 patients est bien plus stable qu'une valeur individuelle. Au chapitre 4, nous transformerons ces observations en intervalles de confiance. Retenez cependant aussi sa limite : le théorème parle de moyennes, pas de valeurs individuelles. Un intervalle de référence pour les patients (section 3.5) exige que les valeurs individuelles soient normales ; le théorème central limite ne nous aide pas dans ce cas.

Vous pouvez refaire vous-mêmes la simulation dans Google Sheets, sans aucun logiciel particulier : lors des travaux pratiques 3, on tire des échantillons avec la fonction RAND et on calcule leurs moyennes, et lors des travaux pratiques 4, on suit l'évolution de la largeur des intervalles de confiance.

3.8 La vérification de la normalité

3.8.1 Pourquoi vérifier et que vérifier exactement

De nombreuses méthodes des chapitres suivants (l'intervalle de confiance fondé sur la loi t, le test t, l'ANOVA, la corrélation de Pearson, la régression linéaire) supposent que les données ou les résidus proviennent d'une distribution normale. Si cette hypothèse est gravement violée, surtout dans les petits échantillons, les valeurs p et les intervalles de confiance peuvent être trompeurs. C'est pourquoi, avant de choisir un test, nous examinons la distribution. La vérification comporte trois volets, qui s'utilisent ensemble : les méthodes graphiques, les indicateurs numériques et les tests formels [49,53].

Une précision qui évite bien des confusions : ce qui doit être normal dépend du test. Pour le test t pour échantillons indépendants, c'est la distribution de la variable dans chaque groupe qui nous intéresse, pas dans l'échantillon total. Pour le test t apparié, c'est la distribution des différences, pas celle des valeurs initiales et finales. En régression, c'est la distribution des résidus. Tester la normalité sur "toute la colonne" est parfois sans rapport avec le test que nous allons appliquer.

3.8.2 Les méthodes graphiques

L'histogramme est la première chose que nous regardons. Pour une variable normale, il présente un seul pic, à peu près au milieu, et deux "épaules" qui descendent symétriquement. L'histogramme de la PAS de la figure 3.6A correspond à cette description ; celui des triglycérides de la figure 3.6B a son pic décalé vers la gauche et une longue queue vers la droite. L'histogramme a toutefois un défaut : son aspect dépend du nombre et de la largeur des classes, et dans les petits échantillons (moins de 30 à 40 valeurs) il peut paraître irrégulier même si les données proviennent d'une population parfaitement normale.

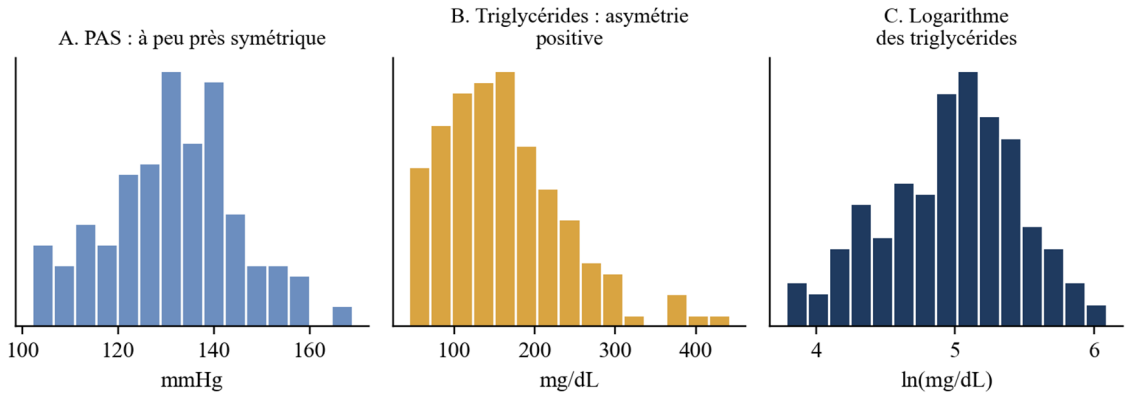


Figure 3.6. La forme de la distribution dans le jeu de données mg2026. A : la pression systolique initiale, approximativement symétrique (asymétrie 0,01). B : les triglycérides, avec une asymétrie positive (0,95) et une longue queue vers la droite. C : le logarithme népérien des triglycérides, approximativement symétrique (asymétrie -0,28). Données synthétiques (jeu de données mg2026).

Le diagramme Q-Q (en anglais quantile-quantile plot) compare les quantiles observés aux quantiles que nous attendrions si les données étaient normales. Chaque point est une observation : sur l'axe horizontal, la valeur qu'elle devrait avoir si la distribution était normale ; sur l'axe vertical, la valeur réelle. Si les données sont normales, les points s'alignent sur une droite. Les écarts suivent des schémas faciles à reconnaître :

- les points s'incurvent vers le haut à l'extrémité droite et restent au-dessus de la droite à l'extrémité gauche (forme de "banane" ou de "U") : asymétrie positive, comme pour les triglycérides ;
- la courbure inverse (forme de "n") : asymétrie négative ;
- les points descendent sous la droite à gauche et montent au-dessus de la droite à droite, comme un "S" couché : queues plus épaisses que celles de la loi normale (aplatissement positif) ;
- un seul point loin de la droite, à une extrémité : une valeur extrême.

Dans la figure 3.7, les points de la PAS suivent la droite presque parfaitement, avec de petits écarts aux extrémités, normaux pour 160 observations. Les points des triglycérides forment clairement une courbe : les grandes valeurs sont plus grandes que ne le prédirait le modèle normal, et les petites valeurs sont moins petites qu'elles ne devraient l'être.

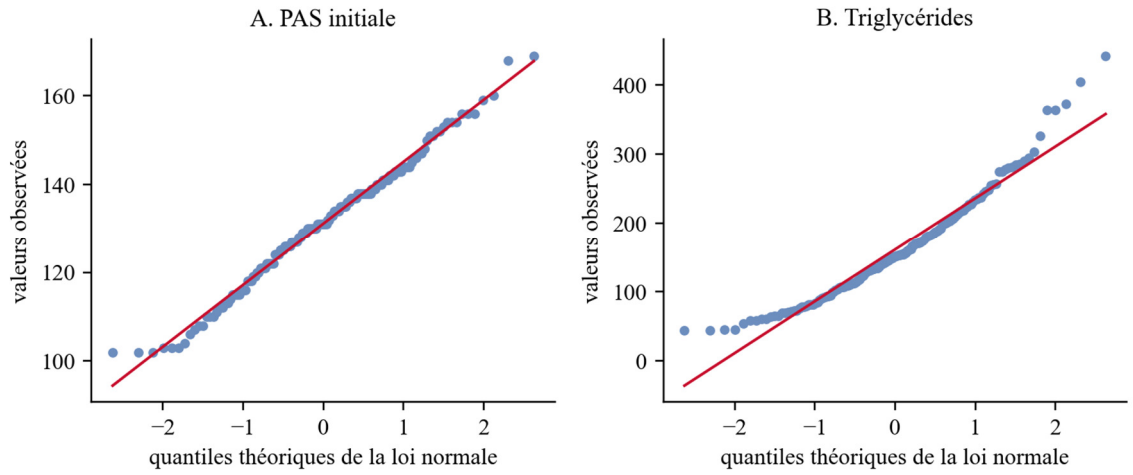


Figure 3.7. Les diagrammes Q-Q de la pression systolique initiale (A), où les points suivent la droite, et des triglycérides (B), où les points forment une courbe caractéristique de l'asymétrie positive. Données synthétiques (jeu de données mg2026).

La boîte à moustaches (box-plot), décrite au chapitre 2, montre l'asymétrie par la position de la médiane dans la boîte (plus proche de Q1 quand l'asymétrie est positive), par la longueur inégale des "moustaches" et par les points individuels situés au-delà de celles-ci. Elle est excellente pour comparer plusieurs groupes et pour repérer les valeurs extrêmes, mais elle dit peu de chose sur la forme exacte de la distribution.

3.8.3 Les indicateurs numériques : asymétrie et aplatissement

L'asymétrie (en anglais skewness) mesure à quel point la distribution est déséquilibrée par rapport à la moyenne. Pour une distribution symétrique, l'asymétrie vaut 0. Les valeurs positives indiquent une queue plus longue vers la droite (des valeurs élevées rares, mais très éloignées), les valeurs négatives, une queue vers la gauche. L'aplatissement (en anglais excess kurtosis), souvent appelé simplement "kurtosis" dans les logiciels, mesure l'épaisseur des queues par rapport à une loi normale : la loi normale a un aplatissement de 0, une distribution à queues épaisses et à pic pointu a un aplatissement positif, une distribution "écrasée", à queues minces, a un aplatissement négatif. Google Sheets (fonctions SKEW et KURT), JASP et MedCalc fournissent les versions corrigées pour l'échantillon de ces indicateurs ; ils donnent donc les mêmes valeurs sur les mêmes données.

À partir de quelle valeur l'asymétrie doit-elle nous inquiéter ? La réponse dépend de n, car dans les petits échantillons l'asymétrie varie beaucoup d'un échantillon à l'autre même si la population est normale. C'est pourquoi l'asymétrie et l'aplatissement sont rapportés à leurs erreurs standard. Pour n grand, approximativement :

$$\mathrm{ES(\ asymétrie\ )} \approx \sqrt{\frac{6}{n}} \qquad \mathrm{ES(\ kurtosis\ )} \approx \sqrt{\frac{24}{n}}$$

Les formules exactes, utilisées par les logiciels, donnent pour n = 160 une erreur standard de 0,192 pour l'asymétrie et de 0,381 pour l'aplatissement (les approximations donnent 0,194 et 0,387). La règle pratique veut qu'un rapport |asymétrie / ES| ou |aplatissement / ES| supérieur à 2 (certains auteurs utilisent 1,96 ou, pour les grands échantillons, 3) signale un écart à la normalité qui mérite d'être pris en compte.

Deux autres indicateurs simples complètent le tableau. Le premier est la comparaison de la moyenne et de la médiane : dans une distribution symétrique, elles sont proches ; dans une distribution à asymétrie positive, la moyenne est supérieure à la médiane (tirée par la queue

droite). Le second est le test de la "moyenne moins deux écarts-types" : pour une variable qui ne peut pas être négative, si m - 2s est inférieur à zéro, la distribution ne peut pas être normale, car le modèle normal prédirait au moins 2,3% de valeurs négatives, impossibles [53].

3.8.4 Les tests formels de normalité

Les tests de normalité formulent les hypothèses ainsi :

H0 : les données proviennent d'une population à distribution normale ; H1 : les données proviennent d'une population dont la distribution n'est pas normale.

Une petite valeur p (inférieure à 0,05) nous conduit à rejeter H0 et à dire que la distribution s'écarte significativement de la normalité. Une grande valeur p signifie que nous n'avons pas suffisamment de preuves contre la normalité. Remarquez la formulation : non pas "les données sont normales", mais "nous n'avons pas trouvé de preuves qu'elles ne le soient pas". Cette différence deviendra centrale au chapitre 5.

Le test de Shapiro-Wilk [66] calcule une statistique W, comprise entre 0 et 1, qui mesure dans quelle mesure les valeurs ordonnées s'alignent sur les quantiles normaux attendus ; c'est essentiellement une corrélation sur le diagramme Q-Q. Des valeurs de W proches de 1 indiquent la normalité. C'est le test le plus puissant pour la plupart des écarts à la normalité, et c'est le test par défaut dans JASP. Il est recommandé surtout pour les petits et moyens échantillons.

Le test de Kolmogorov-Smirnov compare la fonction de répartition observée (les fréquences cumulées des données) à la fonction de répartition de la loi normale et utilise comme statistique D la plus grande distance verticale entre les deux courbes. La version classique suppose que la moyenne et l'écart-type de la population sont connus à l'avance. Lorsque nous les estimons à partir des mêmes données, comme c'est presque toujours le cas, le test classique devient trop "indulgent" : il donne des valeurs p trop grandes. La correction de Lilliefors résout ce problème. Dans mg2026, le test de Kolmogorov-Smirnov classique appliqué à l'IMC donne p = 0,082, ce qui suggérerait, à tort, la normalité, tandis que la version avec correction de Lilliefors donne p = 0,002, et Shapiro-Wilk p < 0,001. Certains logiciels affichent le test classique sans le dire explicitement ; il vaut la peine de vérifier dans la documentation.

Le test d'Anderson-Darling est une variante de l'idée de Kolmogorov-Smirnov qui accorde davantage de poids aux queues de la distribution. La statistique A^2 est comparée à une valeur critique (pour n = 160 et α = 0,05, environ 0,75). Vous le rencontrerez dans MedCalc et dans les logiciels de contrôle de qualité des laboratoires.

Le tableau 3.6 rassemble toutes ces informations pour les principales variables de mg2026.

Tableau 3.6. Vérification de la normalité pour quelques variables du jeu de données mg2026 (n = 160 ; ES de l'asymétrie 0,192 ; ES de l'aplatissement 0,381).

Variable	Moyenne	Médiane	Asymétrie (rapport à l'ES)	Aplatissement (rapport à l'ES)	Shapiro-Wilk W ; p	Lilliefors p	Anderson-Darling A^2
Âge (ans)	56,29	56	0,01 (0,06)	-0,09 (-0,24)	0,993; 0,687	0,835	0,26
PAS0 (mmHg)	131,13	131	0,01 (0,03)	-0,10 (-0,26)	0,990; 0,303	0,534	0,38
Glicemie_0(n	105,44	104,5	0,37 (1,92)	-0,23 (-0,60)	0,982; 0,037	0,127	0,63
$\text{IMC}_0\left(\frac{\text{kg}}{m^2}\right)$	29,04	28,25	1,42 (7,42)	3,68 (9,64)	0,904; < 0,001	0,002	2,80
Triglycérides (mg/dL)	161,18	151,5	0,95 (4,97)	1,12 (2,94)	0,945; < 0,001	0,016	1,78

Variable	Moyenne	Médiane	Asymétrie (rapport à l'ES)	Aplatissement (rapport à l'ES)	Shapiro-Wilk W ; p	Lilliefors p	Anderson-Darling A^2
ln(triglycérides)	4,97	5,02	-0,28 (-1,45)	-0,28 (-0,73)	0,988; 0,200	0,364	0,50

Si l'on lit le tableau ligne par ligne, les conclusions sont claires pour quatre des variables. L'âge et la PAS sont compatibles avec la normalité selon tous les critères : asymétrie et aplatissement pratiquement nuls, moyenne égale à la médiane, tous les tests non significatifs (Shapiro-Wilk p = 0,687 pour l'âge et p = 0,303 pour la PAS ; Kolmogorov-Smirnov classique pour la PAS p = 0,823). L'IMC est nettement asymétrique : l'asymétrie de 1,42 est plus de 7 fois supérieure à son erreur standard, l'aplatissement de 3,68 indique des queues épaisses (quelques patients atteints d'obésité sévère), la moyenne dépasse la médiane de 0,8 kg/m², et tous les tests rejettent la normalité. Les triglycérides sont eux aussi asymétriques, mais leur logarithme se comporte comme une variable normale : nous disons que les triglycérides suivent une distribution log-normale.

3.8.5 Le cas limite : la glycémie à jeun

La glycémie initiale mérite une discussion à part, car ici les critères ne concordent plus. Shapiro-Wilk donne p = 0,037 ; strictement parlant, nous rejetons donc la normalité au seuil de 0,05. En revanche, le test de Lilliefors (p = 0,127) et le test d'Anderson-Darling ($A^2 = 0{,}63,$, sous la valeur critique de 0,75) ne la rejettent pas. L'asymétrie est de 0,37, avec un rapport à l'erreur standard de 1,92, juste sous le seuil de 2. L'aplatissement est négligeable. La moyenne (105,44 mg/dL) et la médiane (104,5 mg/dL) diffèrent de moins de 1 mg/dL. L'histogramme de la glycémie de la figure 2.3C (chapitre 2) montre un léger allongement vers la droite.

Que faire ? D'abord, comprendre d'où vient l'écart. L'échantillon comprend 29 patients atteints d'un diabète non diagnostiqué, chez qui la glycémie est plus élevée ; ils forment une "bosse" dans la queue droite, sur une distribution par ailleurs symétrique. C'est une explication biologique plausible, pas un artefact. Ensuite, se demander quelle analyse va suivre. Pour un intervalle de confiance de la moyenne ou un test t sur 160 patients, une asymétrie de 0,37 est pratiquement sans importance : le théorème central limite garantit que la moyenne est distribuée de façon presque normale (section 3.7). Pour un intervalle de référence des valeurs individuelles, l'écart compterait davantage, car il touche précisément les queues. Enfin, rapporter les choses de façon transparente : "la glycémie présentait une légère asymétrie positive (0,37 ; Shapiro-Wilk p = 0,037) ; nous avons utilisé des tests paramétriques, et les résultats ont été confirmés par des tests non paramétriques". Au chapitre 6, vous verrez que, pour la comparaison de la glycémie entre les sexes, le test t et le test de Mann-Whitney conduisent à la même conclusion.

La leçon générale est que la décision "normale ou non" ne se prend pas mécaniquement, d'après une seule valeur p, mais en pesant tous les indices, compte tenu de la taille de l'échantillon et de l'analyse qui suit [18,42].

3.8.6 Les pièges de l'interprétation des tests de normalité

Les tests de normalité ont un problème de fond : leur puissance dépend fortement de la taille de l'échantillon, exactement à l'inverse de nos besoins.

Dans les grands échantillons (des centaines ou des milliers d'observations), les tests deviennent extrêmement sensibles et rejettent la normalité pour des écarts minuscules, sans importance pratique. Dans une simulation portant sur des séries de 2000 valeurs tirées d'une distribution dont l'asymétrie n'était que de 0,2, le test de Shapiro-Wilk a rejeté la normalité dans plus de 90% des répétitions, alors qu'un test t aurait parfaitement fonctionné sur de telles données.

C'est précisément dans les grands échantillons, là où le théorème central limite nous protège, que le test nous alarme le plus souvent.

Dans les petits échantillons (moins de 20 à 30 observations), les tests ont une faible puissance et ne détectent même pas des écarts importants. Nous avons tiré de mg2026 trois échantillons aléatoires de 12 valeurs de triglycérides, une variable que nous savons asymétrique. Shapiro-Wilk a donné p = 0,084, p = 0,313 et p = 0,259 : aucun échantillon n'a "échoué" au test, alors que l'un d'eux avait une asymétrie de 1,40. C'est précisément dans les petits échantillons, là où la normalité compte le plus, que le test nous rassure à tort.

Pire encore, une valeur p > 0,05 ne prouve pas la normalité. Elle dit seulement qu'avec les données disponibles, nous ne pouvons pas rejeter l'hypothèse. Des formulations comme "le test de Shapiro-Wilk a confirmé la normalité" sont fausses et apparaissent fréquemment dans les mémoires de licence, et même dans des articles. La formulation correcte est "le test de Shapiro-Wilk n'a pas indiqué d'écart significatif à la normalité (p = 0,303)".

La conclusion pratique : utilisez les tests comme un indice, non comme un verdict. Regardez toujours les graphiques. Pour n grand, jugez surtout d'après le graphique, l'asymétrie et la différence moyenne-médiane ; pour n petit, soyez prudents même lorsque le test n'est pas significatif et tenez compte de ce que vous savez de la variable grâce à d'autres études (par exemple, que les durées d'hospitalisation sont presque toujours asymétriques).

> **Étude de cas : l'écart-type qui "trahit" l'asymétrie**
>
> Cioca et ses collaborateurs ont comparé l'évolution de la COVID-19 chez 486 patients : 200 sans diabète, 62 atteints de diabète de type 1 et 224 de diabète de type 2 [8]. La durée d'hospitalisation a été rapportée sous forme de moyenne ± ÉT : 11,6 ± 7,0 jours pour le diabète de type 2 et 9,1 ± 5,8 jours pour le type 1. Appliquez le test "m - 2s" : 11,6 - 2 x 7,0 = -2,4 jours et 9,1 - 2 x 5,8 = -2,5 jours. Si les durées étaient normales, environ 5% des patients auraient séjourné à l'hôpital un nombre négatif de jours, ce qui est impossible. La distribution est donc asymétrique vers la droite, comme c'est presque toujours le cas pour les durées d'hospitalisation : la plupart des patients restent une durée typique, et quelques-uns, avec des complications, très longtemps. La médiane et l'écart interquartile auraient décrit plus fidèlement les données, et la comparaison aurait dû être faite avec un test non paramétrique ou sur une échelle logarithmique. Le même schéma apparaît dans l'étude de Popa et de ses collaborateurs sur les biomarqueurs de l'accident vasculaire cérébral, chez 85 patients [61] : la GFAP était de $142{,}91 \pm 102{,}19$ pg/mL chez les patients ayant eu un AVC, et UCH $-L_1$ $1307{,}68 \pm 967{,}54$ pg/mL, soit des écarts-types de 70 à 75% de la moyenne, avec un m - 2s bien inférieur à zéro. Pour les biomarqueurs sériques, un écart-type proche de la moyenne est presque toujours le signe d'une distribution log-normale. Il ne s'agit pas d'une erreur de calcul, mais d'un choix de présentation qui mérite d'être discuté : la moyenne ± ÉT ne décrit bien que les données symétriques.

3.9 La transformation des données

3.9.1 Le principe de la transformation

Lorsqu'une variable est asymétrique, nous avons trois options : utiliser des méthodes qui ne supposent pas la normalité (les méthodes non paramétriques, chapitre 7), transformer la variable de façon à la rendre approximativement normale puis appliquer les méthodes paramétriques, ou utiliser des modèles plus avancés qui acceptent directement d'autres distributions. Transformer signifie appliquer la même fonction mathématique à chaque valeur : logarithme, racine carrée, inverse. L'ordre des valeurs est conservé (celui qui avait des triglycérides plus élevés les a

toujours plus élevés), mais les distances entre elles changent : les grandes valeurs sont davantage "comprimées" que les petites, et la longue queue se raccourcit [42,53].

3.9.2 La transformation logarithmique

La transformation logarithmique est la plus utilisée en médecine, car de nombreux processus biologiques sont multiplicatifs : une concentration est "multipliée par deux", et non "augmentée de 50 unités". On peut utiliser le logarithme népérien (ln, base e) ou le logarithme décimal (log10) ; la forme de la distribution obtenue est la même, seule l'échelle diffère. Dans Google Sheets : =LN(R2) ou =LOG10(R2). Le logarithme exige des valeurs strictement positives ; pour les variables qui peuvent valoir zéro (par exemple le nombre de jours d'arrêt maladie), on utilise parfois ln(x + 1), avec prudence dans l'interprétation.

Pour les triglycérides de mg2026, le passage aux logarithmes a réduit l'asymétrie de 0,95 à -0,28 et l'aplatissement de 1,12 à -0,28 ; Shapiro-Wilk passe de p < 0,001 à p = 0,200 (figure 3.6C). Sur l'échelle logarithmique, il n'y a plus non plus de valeurs extrêmes : aucun patient n'a un score z supérieur à 2,3 en valeur absolue.

3.9.3 Autres transformations et "échelle des puissances"

Les transformations usuelles peuvent être vues comme les barreaux d'une échelle, de la plus douce à la plus forte : racine carrée $\left(\sqrt{x}\right)$, logarithme (ln x), inverse de la racine carrée $\left(\frac{1}{\sqrt{x}}\right)$, inverse (1/x). Plus l'asymétrie positive est forte, plus nous montons haut sur l'échelle. Pour une asymétrie négative (rare en médecine, par exemple les scores d'un test facile, où la plupart obtiennent presque la note maximale), on transforme d'abord la variable "réfléchie" (valeur maximale + 1 - x). La transformation racine carrée est traditionnellement utilisée pour les comptages (par exemple le nombre de colonies, le nombre de lésions), et la transformation arc sinus pour les proportions, bien qu'aujourd'hui celles-ci soient généralement analysées avec des modèles dédiés.

Tableau 3.7. Effet des transformations sur les triglycérides (mg2026, n = 160).

Transformation	Asymétrie	Aplatissement	Shapiro-Wilk p
sans transformation	0,95	1,12	< 0,001
racine carrée $\sqrt{x}$	0,32	-0,10	0,181
logarithme ln x	-0,28	-0,28	0,200
Box-Cox, λ = 0,22	-0,01	-	0,710
inverse 1/x	1,48	2,30	< 0,001

La racine carrée et le logarithme sont tous deux acceptables, avec des asymétries de signes opposés et d'ampleur similaire. L'inverse est trop fort : il transforme une asymétrie positive modérée en une asymétrie encore plus grande (l'inverse renverse l'ordre, et les petites valeurs, de 44 à 60 mg/dL, deviennent maintenant la longue queue). La transformation de Box-Cox généralise l'échelle : elle cherche automatiquement l'exposant λ qui rapproche le plus possible les données de la normalité (λ = 1 signifie aucune transformation, λ = 0,5 la racine carrée, λ = 0 le logarithme, λ = -1 l'inverse). Pour les triglycérides, le λ optimal est de 0,22, entre la racine carrée et le logarithme. En pratique, nous choisissons une transformation facile à interpréter (généralement le logarithme) plutôt qu'un exposant "exotique" comme 0,22, même si celui-ci donne un p un peu plus grand ; un résultat exprimé en "triglycérides à la puissance 0,22" ne peut être communiqué à personne.

La transformation ne résout pas toujours le problème. L'IMC de mg2026 reste asymétrique après passage aux logarithmes (asymétrie 0,89, Shapiro-Wilk p < 0,001), et Box-Cox exige un exposant de -2,5, c'est-à-dire une transformation très forte et sans interprétation clinique, après laquelle Shapiro-Wilk donne p = 0,076. L'asymétrie de l'IMC provient en bonne partie de

quelques patients atteints d'obésité sévère, et non d'un processus multiplicatif. Ici, le meilleur choix est une méthode non paramétrique ou la présentation de la médiane et de l'EIQ.

3.9.4 La retransformation : la moyenne géométrique

Après avoir analysé les données sur l'échelle logarithmique, il faut ramener les résultats sur l'échelle d'origine pour qu'un clinicien puisse les comprendre. La moyenne des logarithmes des triglycérides est de 4,970. La fonction inverse du logarithme népérien est l'exponentielle :

$$e^{4,970} = 144,0 \text{ mg/dL}$$

Cette valeur n'est pas la moyenne arithmétique des triglycérides (161,18 mg/dL). C'est la moyenne géométrique, c'est-à-dire la racine n-ième du produit des n valeurs. Pour une distribution log-normale, la moyenne géométrique est proche de la médiane (151,5 mg/dL) et décrit mieux la valeur "typique" que la moyenne arithmétique, tirée vers le haut par la queue droite.

L'écart-type sur l'échelle logarithmique (0,488) ne se retransforme pas en écart-type. Son exponentielle, $e^{0,488} = 1,63$, est un facteur multiplicatif, appelé écart-type géométrique : environ deux tiers des patients ont des triglycérides compris entre 144,0 / 1,63 et 144,0 x 1,63, c'est-à-dire entre 88 et 235 mg/dL. Remarquez l'asymétrie de l'intervalle par rapport à 144 : 56 mg/dL en dessous, 91 mg/dL au-dessus.

L'intervalle de confiance de la moyenne géométrique se calcule sur l'échelle logarithmique, puis se retransforme. L'erreur standard de la moyenne des logarithmes est de 0,488 / √160 = 0,0386, et la valeur critique t pour 159 ddl est de 1,975 ; donc :

$4,9699 \pm 1,975 \times 0,03859 = 4,9699 \pm 0,0762$, c'est-à-dire de 4,8937 à 5,0461

La retransformation des limites donne $e^{4,8937} = 133,44$ et $e^{5,0461} = 155,42$ mg/dL ; l'IC à 95% de la moyenne géométrique est donc 133,44-155,42 mg/dL. Nous avons volontairement conservé quatre décimales sur l'échelle logarithmique : l'exponentielle amplifie les petites erreurs, et un arrondi à deux décimales avant la retransformation déplacerait chaque limite de près de 0,5 mg/dL. La règle pratique est que, sur l'échelle logarithmique, on n'arrondit qu'à la fin, après la retransformation. L'intervalle retransformé n'est pas symétrique autour de 144,0, et c'est normal. Nous reprendrons ce calcul au chapitre 4.

Une dernière précision : une différence sur l'échelle logarithmique devient, après retransformation, un rapport. Si la moyenne des logarithmes des triglycérides était supérieure de 0,10 dans un groupe par rapport à un autre, la retransformation donne $e^{0,10} = 1,105$: la moyenne géométrique est supérieure d'environ 10,5%, et non "de 0,10 mg/dL". C'est la raison pour laquelle les études qui analysent des biomarqueurs sur une échelle logarithmique rapportent des rapports de moyennes géométriques.

3.9.5 Quand choisir des méthodes non paramétriques plutôt qu'une transformation

La transformation est utile lorsqu'il existe une transformation naturelle, facile à interpréter (le logarithme pour les concentrations), lorsque nous voulons estimer des paramètres avec des intervalles de confiance et lorsque des analyses multivariées (régression) vont suivre. Les méthodes non paramétriques, fondées sur les rangs, sont préférables lorsqu'aucune transformation simple ne normalise les données (l'IMC de mg2026), lorsque la variable est ordinale (scores de Likert, stades), lorsque l'échantillon est petit et que nous ne pouvons pas juger de la forme de la distribution, lorsqu'il existe des valeurs extrêmes que nous ne pouvons pas exclure ou lorsque les données présentent un "plafond" ou un "plancher" (beaucoup de valeurs égales au minimum ou au maximum de l'échelle). Le prix des méthodes non paramétriques est une puissance un peu plus faible lorsque les données sont malgré tout normales, et le fait qu'elles comparent des distributions

ou des médianes, et non des moyennes. L'algorithme complet de choix du test est présenté au chapitre 8.

> **Étude de cas : données asymétriques présentées sous forme de médiane (EIQ)**
>
> Gherbon et ses collaborateurs ont suivi pendant un an 200 patients atteints de diabète de type 2 traités par inhibiteurs du SGLT-2 [59]. La glycémie est passée de 180,00 mg% (EIQ 154,50-207,00) à 130,00 mg% (EIQ 117,50-150,00), l'HbA1c de 8,40% (EIQ 7,98-9,15%) à 7,30% (EIQ 6,90-7,95%), et la PAS de 140,00 mmHg (130,00-160,00) à 120,00 mmHg (115,50-130,00), toutes avec p < 0,001. Remarquez que les auteurs ont choisi la médiane et l'EIQ pour les variables asymétriques, mais la moyenne ± ÉT pour le tour de taille ($107{,}05 \pm 16{,}39$ cm, puis $102{,}50 \pm 15{,}11$ cm; $p = 0{,}042$), probablement distribué de façon presque symétrique. C'est le bon choix : l'indicateur est adapté à la distribution de chaque variable. Regardez les quartiles de la glycémie initiale : 180 - 154,5 = 25,5 mg% sous la médiane, 207 - 180 = 27 mg% au-dessus, une légère asymétrie. Pour la PAS, Q1 = 130 se situe 10 mmHg sous la médiane de 140, et Q3 = 160, 20 mmHg au-dessus : la distribution est étirée vers la droite. Un seul détail de lecture critique : le résumé appelle ces valeurs des "moyennes", alors que ce sont des médianes ; ce genre d'inattention terminologique est fréquent, et il est bon de savoir la reconnaître. Frandeş et ses collaborateurs ont procédé de la même manière dans une étude transversale portant sur 304 adultes diabétiques, en présentant les scores d'implication numérique sous forme de médiane [Q1-Q3] : 1,71 [1,71-2,39] contre 3,86 [3,86-4,43] [7]. Le fait que la médiane coïncide avec Q1 dans les deux groupes montre qu'au moins un quart des participants ont exactement la valeur minimale de leur groupe, une distribution avec "plancher" pour laquelle la moyenne ±ÉT ar fi aurait été trompeuse ; les auteurs ont utilisé, de façon cohérente, la régression bêta, adaptée aux scores bornés. C'est aussi avec des médianes qu'ont été comparées les valeurs de l'indice TyG dans trois phénotypes endocrino-métaboliques, chez 387 patients évalués pour une dysfonction thyroïdienne : 8,24, 8,74 et 8,87, p < 0,001 [67].

3.10 Les valeurs extrêmes

3.10.1 Comment les reconnaître

Une valeur extrême (en anglais outlier) est une observation qui se situe anormalement loin du reste des données. Elle peut être une erreur (de saisie, d'unité de mesure, d'appareil), une valeur réelle mais rare (un patient atteint d'obésité sévère dans une étude de prévention) ou le signe que la distribution a, en réalité, une longue queue. La première étape consiste toujours à l'identifier ; la seconde, à comprendre d'où elle vient. Dans mg2026_brut.csv, la valeur de PAS 1540 et la taille 1,72 sont des erreurs évidentes (un zéro de trop, des mètres au lieu de centimètres), corrigées au chapitre 2. Mais que faire des valeurs plausibles, quoique éloignées ?

Trois critères simples sont couramment utilisés :

- Le score z : les valeurs avec $|z| > 3$ sont suspectes. Ce critère suppose une distribution approximativement normale ; dans une loi normale, seulement 0,3% des valeurs dépassent 3 écarts-types. Il a un défaut : la valeur extrême elle-même gonfle la moyenne et l'écart-type, et "masque" donc en partie son propre caractère extrême.
- La règle 1,5 x EIQ (Tukey) : on calcule les "barrières" Q1 - 1,5 x EIQ et Q3 + 1,5 x EIQ ; les valeurs situées au-delà sont signalées comme extrêmes (les points individuels de la boîte à moustaches). Pour les valeurs très extrêmes, on utilise 3 x EIQ. L'avantage est que la règle repose sur les quartiles, qui ne sont pas influencés par les valeurs extrêmes.
- Le test de Grubbs : un test formel pour une seule valeur extrême dans un échantillon supposé normal, avec la statistique $G = \frac{|x-m|}{s}$ comparée à une valeur critique qui dépend de n (pour n = 160 et α = 0,05, environ 3,54). Il existe aussi des variantes pour plusieurs valeurs extrêmes. Le test

s'applique à des données presque normales ; sur une variable asymétrique, il trouvera des "valeurs extrêmes" qui sont en fait la queue naturelle de la distribution.

Tableau 3.8. Valeurs extrêmes dans le jeu de données mg2026 selon les critères du score z et de l'EIQ.

Variable	Q1-Q3	Barrières 1,5 x EIQ	Valeurs au-delà des barrières	Valeurs avec z en valeur absolue > 3	Valeur maximale et son score z
PAS0 (mmHg)	122-140	95-167	2 (168, 169)	0	169; 2,73
Âge (ans)	50,75-62	33,9-78,9	1 (79)	0	79; 2,59
Glicemie_0(mg/dl	93-116	58,5-150,5	1 (154)	1	154; 3,08
$\text{IMC}_0\left(\frac{\text{kg}}{m^2}\right)$	26,2-31,4	18,4-39,2	3 (40,0; 43,4; 48,3)	2	48,3; 4,92
Triglycérides (mg/dL)	107-203,25	-37,4-347,6	5 (364; 364; 373; 405; 442)	2	442; 3,68
ln(triglycérides)	4,67-5,31	3,71-6,28	0	0	6,09; 2,30

Remarquez que la règle de l'EIQ signale une ou deux valeurs même pour des variables normales, comme la PAS et l'âge. C'est attendu : même dans une distribution parfaitement normale, environ 0,7% des valeurs sortent des barrières de Tukey, soit un ou deux patients sur 160. Pour l'IMC, la statistique de Grubbs de la valeur 48,3 est de 4,92, bien au-dessus de la valeur critique, et pour les triglycérides, la valeur 442 a un G = 3,68, juste au-dessus du seuil. Après passage aux logarithmes, en revanche, les triglycérides n'ont plus aucune valeur extrême : les "valeurs extrêmes" étaient la queue normale d'une distribution log-normale.

3.10.2 Que faire de ces valeurs

La règle d'or : les valeurs extrêmes ne sont jamais supprimées en silence. L'ordre des étapes est le suivant. Nous vérifions d'abord si la valeur est une erreur, en la comparant au document source (dossier du patient, compte rendu de laboratoire). S'il s'agit d'une erreur et que la valeur correcte peut être retrouvée, nous la corrigeons et consignons la correction. S'il s'agit d'une erreur et qu'elle ne peut pas être retrouvée, nous la traitons comme une valeur manquante et le signalons. Si la valeur est biologiquement plausible, nous la conservons. Un IMC de 48,3 kg/m² est rare, mais il existe ; un patient atteint d'obésité de grade III est exactement le type de patient qui fait partie de la population étudiée.

Pour les valeurs plausibles, nous choisissons des méthodes qui les tolèrent : médiane et EIQ au lieu de moyenne et ÉT, tests non paramétriques, transformation logarithmique, méthodes robustes. Nous faisons ensuite une analyse de sensibilité : nous répétons l'analyse avec et sans les valeurs extrêmes et indiquons si la conclusion change. Si nous excluons le patient dont l'IMC est de 48,3, la moyenne de l'IMC passe de 29,04 à 28,92 kg/m², l'écart-type de 3,91 à 3,61 et l'asymétrie de 1,42 à 0,95 ; Shapiro-Wilk reste néanmoins significatif ($p < 0,001$). Exclure une seule valeur ne "répare" pas la distribution, et continuer à exclure des valeurs jusqu'à ce que le test devienne non significatif est une forme de manipulation des données. Au chapitre 9, vous verrez à quel point une seule valeur extrême peut modifier un coefficient de corrélation.

Toute exclusion doit être fixée à l'avance dans le protocole (par exemple "les valeurs de glycémie supérieures à 600 mg/dL seront vérifiées dans le compte rendu de laboratoire") et rapportée dans l'article, avec le nombre d'observations exclues et le motif.

> **Étude de cas : délais, scores et tests fondés sur les rangs**
>
> Les variables de temps en médecine d'urgence sont presque toujours asymétriques : elles ne peuvent pas être négatives, la plupart des patients se situent dans un intervalle typique, et quelques-uns arrivent avec beaucoup de retard. Popa et ses collaborateurs ont analysé 258

patients éligibles atteints d'AVC ischémique aigu (sur 355 évalués), dont seulement 46 ont reçu une thrombolyse intraveineuse, soit un taux de 17,83% [62]. Les délais ont été décrits par des médianes : une médiane de 100 minutes entre le début des symptômes et l'arrivée aux urgences ($p < 0,001$) et un délai porte-scanner de 15,5 minutes ($p = 0,009$). Le choix de la médiane est correct ici : une moyenne aurait été tirée vers le haut par les quelques patients arrivés après de nombreuses heures. En maladies infectieuses, Marin et ses collaborateurs ont comparé les marqueurs immunologiques entre trois types de réponse au traitement antirétroviral, chez 462 adultes vivant avec le VIH, à l'aide du test de Kruskal-Wallis, avec $H = 153,62$ ($p < 0,001$, $\varepsilon^2 = 0,33$) et $H = 115,10$ ($p < 0,001$, $\varepsilon^2 = 0,25$) [68]. Le rapport CD4/CD8, rapport de deux numérations cellulaires, a par nature une distribution asymétrique, et le test fondé sur les rangs, accompagné d'une taille d'effet (epsilon carré), est le choix naturel. Dans le même esprit, Gherbon et ses collaborateurs ont utilisé le test de Kruskal-Wallis ($p = 0,003$) pour les scores de détresse liée au diabète chez 453 adultes atteints de diabète de type 2, dont 19 (4,2%) présentaient une détresse sévère [69] : les scores d'un questionnaire sont ordinaux et souvent asymétriques, et un test comparant des moyennes aurait été difficile à justifier. Ce qu'il faut remarquer dans ces trois études : le choix de la méthode suit la forme de la distribution et le type de variable, non l'habitude.

3.11 Distributions qui se chevauchent : du score z aux décisions cliniques

Jusqu'ici, nous avons examiné une seule distribution. En diagnostic, nous en avons généralement deux : la distribution d'un marqueur chez les patients malades et sa distribution chez les sujets indemnes. La capacité du marqueur à séparer les deux groupes dépend de la distance entre les moyennes, mesurée en écarts-types, c'est-à-dire exactement de l'idée du score z.

Lupuşoru et ses collaborateurs ont évalué l'échographie de contraste avec imagerie paramétrique (CEUS-PAT) pour le diagnostic de la cirrhose, chez 64 personnes : 37 patients cirrhotiques et 27 volontaires sains [2]. L'indice mesuré était de 0,83 ± 0,09 chez les patients cirrhotiques et de 0,49 ± 0,11 chez les volontaires ($p < 0,0001$). La présentation sous forme de moyenne ± ÉT est appropriée ici, car les intervalles $m \pm 2s$ restent dans le domaine plausible de l'indice. La différence entre les moyennes, 0,34, représente plus de trois écarts-types : les deux distributions ne se chevauchent presque pas. Il n'est pas étonnant que le seuil > 0,7 ait donné une aire sous la courbe ROC de 0,98, une sensibilité de 89,19% et une spécificité de 100%. Nous vérifierons dans le problème résolu 3.6 que le modèle normal, appliqué aux deux moyennes et aux deux écarts-types, prédit des chiffres très proches. Ce que l'étude ne dit pas, c'est dans quelle mesure l'indice fonctionnerait chez des patients atteints de fibrose intermédiaire : la comparaison entre des malades évidents et des volontaires sains maximise la distance entre les distributions, un phénomène appelé biais de spectre, dont nous parlerons au chapitre 11.

Bende et ses collaborateurs ont mesuré le paramètre d'atténuation UGAP chez 179 patients, en le comparant au CAP [58]. Les moyennes ont augmenté d'un grade de stéatose à l'autre : S_0 198,3 $\pm \frac{25,7 d_B}{m}$, S1 216,86 ± 26,3, S2 237,79 ± 26,3, S_3 270,8 $\pm$ 31,62($p < 0,001$). La différence entre S0 et S1 n'est que de 18,6 dB/m, moins d'un écart-type : les distributions des deux grades se chevauchent largement. La différence entre S0 et S3 dépasse 70 dB/m, soit environ 2,5 écarts-types. Les aires ROC rapportées suivent exactement cette logique : 0,83 pour le seuil S1 (192,5 dB/m), 0,90 pour S2 (231 dB/m) et 0,91 pour S3 (248 dB/m). Plus les distributions sont éloignées, mieux un seuil les sépare.

Étude de cas : la distribution d'une différence

Muntean et ses collaborateurs ont suivi 921 étudiants de première année et ont mesuré la variation de l'IMC entre le début et la fin de l'année [5]. La différence moyenne était de -0,45 ±

1,07 kg/m² (IC à 95% -0,52 à -0,38 ; p < 0,001). Regardez ces deux chiffres : l'écart-type (1,07) est plus de deux fois supérieur à la moyenne en valeur absolue (0,45). Cela signifie que les variations individuelles ont été très diverses et se sont réparties de part et d'autre de zéro, même si la moyenne baisse nettement. Le résumé le confirme : 48,2% des étudiants ont vu leur IMC diminuer, 38,7% sont restés stables et 13,1% ont pris du poids, avec une prise de poids cliniquement significative chez 7,2%. Une moyenne des différences décrit le groupe, pas l'individu ; une étude qui rapporterait seulement que "l'IMC a diminué en moyenne de 0,45 kg/m²" masquerait le fait qu'un étudiant sur huit a pris du poids. C'est la distribution des différences, et non celle des valeurs initiales ou finales, qui compte pour le test t apparié (chapitre 6). Et le fait que les auteurs aient analysé la même variable à la fois comme une différence continue (régression linéaire) et sous forme dichotomisée comme "prise de poids cliniquement significative" (régression logistique) montre qu'une distribution continue peut être envisagée sous deux angles, chacun avec sa propre question.

3.12 Problèmes résolus

Problème résolu 3.1. Pourcentages sous la courbe normale pour la PAS

Énoncé. Dans le jeu de données mg2026, la PAS initiale a une moyenne de 131,13 mmHg et un écart-type de 13,88 mmHg, et la distribution est compatible avec la normalité (Shapiro-Wilk p = 0,303). (a) Quelle proportion de patients devrait, en théorie, avoir une PAS comprise entre 110 et 150 mmHg ? (b) Comparez-la à la proportion observée, 83,8%. (c) Quelle valeur de PAS délimite les 10% de patients ayant les valeurs les plus basses ?

Solution. Nous vérifions d'abord la condition : le modèle normal peut être utilisé, car l'asymétrie (0,01) et l'aplatissement (-0,10) sont pratiquement nuls, et les tests de normalité ne sont pas significatifs.

(a) Nous calculons les scores z des deux limites :

$$z_1 = \frac{110 - 131{,}13}{13{,}88} = -\frac{21{,}13}{13{,}88} = -1{,}52 z_2 = \frac{150 - 131{,}13}{13{,}88} = \frac{18{,}87}{13{,}88} = 1{,}36$$

D'après le tableau E.1 : Φ(1,36) = 0,9131 et Φ(1,52) = 0,9357 ; donc Φ(-1,52) = 1 - 0,9357 = 0,0643. La proportion comprise entre les deux limites est :

$$P(110 \leq \text{PAS} \leq 150) = \Phi(1{,}36) - \Phi(-1{,}52) = 0{,}9131 - 0{,}0643 = 0{,}8488$$

En théorie, environ 84,9% des patients ont une PAS comprise entre 110 et 150 mmHg. Dans Google Sheets : =NORM.DIST(150; 131,13; 13,88; TRUE) - NORM.DIST(110; 131,13; 13,88; TRUE).

(b) La proportion observée est de 83,8% (134 patients sur 160). La différence de 1,1 point de pourcentage est de l'ordre des fluctuations d'échantillonnage attendues pour n = 160. Le modèle normal décrit bien les données.

(c) Nous cherchons z tel que Φ(z) = 0,10. Par symétrie, c'est l'opposé de la valeur pour laquelle Φ(z) = 0,90, c'est-à-dire z = -1,28. Alors x = 131,13 - 1,28 x 13,88 = 131,13 - 17,77 = 113,4 mmHg. Dans Sheets : =NORM.INV(0,1; 131,13; 13,88) donne 113,3 mmHg (la différence provient de l'arrondi de z). Le 10e percentile observé dans l'échantillon est de 112,0 mmHg.

Interprétation clinique. Près de cinq patients à risque cardiométabolique sur six dans cet échantillon ont une PAS comprise entre 110 et 150 mmHg, et un sur dix a une PAS inférieure à 113 mmHg. Ces chiffres décrivent l'échantillon et ne remplacent pas les seuils cliniques des recommandations.

Problème résolu 3.2. La loi binomiale dans le dépistage du diabète

Énoncé. La prévalence du diabète non diagnostiqué dans une population d'adultes à risque cardiométabolique est de 18%. Un médecin généraliste adresse 20 patients choisis au hasard pour une HGPO. (a) Quels sont le nombre attendu de cas et l'écart-type ? (b) Quelle est la probabilité de ne trouver aucun cas ? (c) Et la probabilité d'en trouver 5 ou plus ? (d) Le médecin trouve 8 cas. Ce résultat est-il surprenant ?

Solution. Les conditions de la loi binomiale sont remplies : n = 20 épreuves indépendantes, deux résultats possibles, p = 0,18 constant.

(a) Moyenne = n x p = 20 x 0,18 = 3,6 cas. Variance = 20 x 0,18 x 0,82 = 2,952, et écart-type = √2,952 = 1,72 cas.

(b) $P(X = 0) = 0{,}82^{20}$. Calculons pas à pas : 0,82² = 0,6724 ; 0,82^4 = 0,4521 ; 0,82^8 = 0,2044 ; 0,82^16 = 0,0418 ; $0{,}82^{20} = 0{,}0418 \times 0{,}4521 = 0{,}0189$. La probabilité de ne trouver aucun cas est de 1,9%.

(c) P(X ≥ 5) = 1 - P(X ≤ 4). Le calcul à la main des cinq termes est laborieux ; dans Google Sheets, nous utilisons =1 - BINOM.DIST(4; 20; 0,18; TRUE), qui donne 0,2849. Dans environ 28% des groupes de 20 patients, nous trouverons au moins 5 cas.

(d) P(X ≥ 8) = 1 - BINOM.DIST(7; 20; 0,18; TRUE) = 0,0177. Un tel résultat se produit dans moins de 2% des groupes si la prévalence réelle est de 18%. Il est donc inhabituel et suggère soit que les patients de ce médecin ont un risque plus élevé que la population de référence (par exemple, il a sélectionné des patients plus âgés ou plus obèses), soit un hasard rare. Remarquez que nous avons fait, sans le nommer, un test statistique : nous avons calculé la probabilité d'un résultat au moins aussi extrême sous une hypothèse donnée. C'est exactement la logique de la valeur p du chapitre 5.

Problème résolu 3.3. La loi de Poisson et le seuil d'alerte

Énoncé. Dans un centre de vaccination, on signale en moyenne 1,5 réaction indésirable mineure par mois. (a) Quelle est la probabilité d'un mois sans aucune réaction ? (b) L'équipe veut un seuil d'alerte tel que les fausses alertes surviennent au plus dans 5% des mois. Quel est le plus petit nombre de réactions qui déclenche l'alerte ?

Solution. Des événements rares, dénombrés sur un intervalle de temps, avec une moyenne connue : loi de Poisson avec λ = 1,5.

(a) $P(X = 0) = e^{-1{,}5} = 0{,}2231$. Environ un mois sur quatre ou cinq ne comporte aucune réaction signalée.

(b) Nous calculons les probabilités successives avec la formule $P(X = k) = \frac{e^{-\lambda} \times \lambda^k}{k!}$:

$$P(1) = 0{,}2231 \times 1{,}5 = 0{,}3347 P(2) = \frac{0{,}2231 \times 1{,}5^2}{2} = 0{,}2510 P(3) = \frac{0{,}2231 \times 1{,}5^3}{6}$$
$$= 0{,}1255 P(4) = \frac{0{,}2231 \times 1{,}5^4}{24} = 0{,}0471$$

P(X ≤ 3) = 0,2231 + 0,3347 + 0,2510 + 0,1255 = 0,9343, donc P(X ≥ 4) = 0,0657 (6,6%), au-dessus de la limite de 5%. P(X ≤ 4) = 0,9814, donc P(X ≥ 5) = 0,0186 (1,9%). Le plus petit seuil donnant moins de 5% de fausses alertes est de 5 réactions par mois. Dans Sheets : =1 - POISSON.DIST(4; 1,5; TRUE) = 0,0186.

Interprétation. Un seuil de 4 réactions déclencherait une fausse alerte environ un mois sur quinze ; le seuil de 5, environ un mois sur cinquante. Le choix du seuil est un compromis : un

seuil plus bas détecte plus vite un problème réel, mais produit davantage de fausses alertes. Nous retrouverons le même dilemme avec les tests diagnostiques (sensibilité contre spécificité).

Problème résolu 3.4. Un intervalle de référence et les analyses multiples

Énoncé. Dans un échantillon de 400 femmes en bonne santé, l'hémoglobine a une moyenne de 13,8 g/dL et un écart-type de 1,0 g/dL, avec une distribution approximativement normale (données fictives). (a) Construisez l'intervalle de référence à 95%. (b) Une patiente a 11,6 g/dL. Calculez le score z et le percentile. (c) Une femme en bonne santé fait 12 analyses indépendantes, chacune avec un intervalle de référence à 95%. Quelle est la probabilité que toutes soient "dans les limites de la normale" ?

Solution. $(a)\, m \pm 1{,}96s = 13{,}8 \pm 1{,}96 \times 1{,}0 = 13{,}8 \pm 1{,}96$, c'est-à-dire 11,8-15,8 g/dL (arrondi à une décimale).

(b) $z = \frac{11{,}6 - 13{,}8}{1{,}0} = -2{,}20$. D'après la table, Φ(2,20) = 0,9861, donc Φ(-2,20) = 1 - 0,9861 = 0,0139. La patiente se situe au percentile 1,4 : seulement 1,4% des femmes en bonne santé de la population de référence ont une valeur aussi basse ou plus basse. La valeur est inhabituelle et justifie la recherche d'une anémie, mais la statistique à elle seule ne pose pas le diagnostic.

(c) $P(\text{ toutes les } 12 \text{ normales }) = 0{,}95^{12} = 0{,}540$. Près de la moitié (46%) des femmes en bonne santé qui font 12 analyses indépendantes auront au moins un résultat en dehors de l'intervalle de référence. D'où le conseil pratique : un résultat isolé, légèrement en dehors de l'intervalle, chez une personne asymptomatique, se contrôle et s'interprète dans son contexte avant toute conclusion.

Problème résolu 3.5. Le choix d'une transformation pour les triglycérides

Énoncé. Les triglycérides de mg2026 ont une moyenne de 161,18 mg/dL, une médiane de 151,5 mg/dL, un écart-type de 76,30 mg/dL, une asymétrie de 0,95, un aplatissement de 1,12 et un Shapiro-Wilk p < 0,001. (a) Expliquez pourquoi la distribution n'est pas normale. (b) Choisissez une transformation à partir du tableau 3.7. (c) Calculez la moyenne géométrique et un intervalle dans lequel se situent environ 95% des patients.

Solution. (a) Quatre indices convergent. L'asymétrie divisée par son erreur standard vaut 0,95 / 0,192 = 4,95, bien au-dessus de 2. L'aplatissement rapporté à son erreur standard vaut 1,12 / 0,381 = 2,94, également au-dessus de 2. La moyenne dépasse la médiane de près de 10 mg/dL. Le test de Shapiro-Wilk rejette la normalité. Le diagramme Q-Q (figure 3.7B) présente la forme incurvée typique de l'asymétrie positive.

(b) D'après le tableau 3.7, la racine carrée (asymétrie 0,32 ; p = 0,181) et le logarithme (asymétrie -0,28 ; p = 0,200) normalisent tous deux suffisamment. L'inverse est trop fort. Box-Cox donne le meilleur résultat numérique (λ = 0,22), mais sans interprétation clinique. Nous choisissons le logarithme : il est aussi efficace que la racine carrée, il a un sens biologique (les concentrations varient de façon multiplicative) et il permet la retransformation en moyenne géométrique et en rapports.

(c) La moyenne des logarithmes est de 4,970, leur écart-type de 0,488. La moyenne géométrique est $e^{4{,}970} = 144{,}0$ mg/dL. L'intervalle à 95% sur l'échelle logarithmique est $4{,}970 \pm 1{,}96 \times 0{,}488 = 4{,}970 \pm 0{,}957$, c'est-à-dire 4,013-5,927. Après retransformation : $e^{4{,}013} = 55{,}3$ et $e^{5{,}927} = 374{,}9$ mg/dL. Environ 95% des patients ont des triglycérides compris entre 55 et 375 mg/dL ; dans l'échantillon, 95,6% des valeurs se situent dans cet intervalle, et les percentiles 2,5 et 97,5 observés sont de 53,8 et 364,0 mg/dL.

Interprétation. Sur l'échelle d'origine, l'intervalle n'est pas symétrique autour de la moyenne géométrique (89 mg/dL en dessous, 231 mg/dL au-dessus), exactement comme le montre l'histogramme. L'intervalle calculé naïvement, 11,6-310,7 mg/dL, aurait été faux à ses deux extrémités.

Problème résolu 3.6. Le modèle normal appliqué à des données publiées

Énoncé. Dans une étude diagnostique de la cirrhose [2], l'indice CEUS-PAT était de 0,83 ± 0,09 chez 37 patients cirrhotiques et de 0,49 ± 0,11 chez 27 volontaires sains. Les auteurs ont rapporté, au seuil > 0,7, une sensibilité de 89,19% et une spécificité de 100%. En supposant que l'indice soit normalement distribué dans chaque groupe, quelles sensibilité et spécificité le modèle prédirait-il ?

Solution. La sensibilité est la proportion de malades dont la valeur est supérieure au seuil. Chez les patients cirrhotiques :

$z = \frac{0,7-0,83}{0,09} = -1,44$; Φ(-1,44) = 1 - 0,9251 = 0,0749

La proportion au-dessus du seuil est de 1 - 0,0749 = 0,925 ; la sensibilité prédite est donc de 92,5%.

La spécificité est la proportion de sujets indemnes dont la valeur est inférieure au seuil. Chez les volontaires :

$z = \frac{0,7-0,49}{0,11} = 1,91$; Φ(1,91) = 0,9719

La spécificité prédite est de 97,2%.

Interprétation. Le modèle normal prédit 92,5% et 97,2% ; l'étude a observé 89,19% et 100%. La concordance est bonne, compte tenu des petits effectifs : 100% sur 27 volontaires signifie qu'aucun n'a dépassé le seuil, et le modèle indique que nous nous attendrions à 0,8 volontaire, c'est-à-dire à zéro ou un. L'exercice montre que, si nous connaissons la moyenne et l'écart-type dans deux groupes et que les distributions sont normales, nous pouvons anticiper la performance d'un seuil. Il en montre aussi la limite : le résultat dépend de l'hypothèse de normalité et de la composition des groupes (ici, des malades avérés face à des volontaires sains, ce qui exagère la séparation).

Problème résolu 3.7. La décision de normalité pour l'HbA1c

Énoncé. Pour l'HbA1c initiale de mg2026 : n = 160, moyenne 5,58%, écart-type 0,45%, médiane 5,6%, asymétrie 0,14, aplatissement 0,16, Shapiro-Wilk W = 0,990, p = 0,304. (a) Le modèle normal est-il raisonnable ? (b) Combien de patients devraient, en théorie, avoir une HbA1c ≥ 6,5% ? Il y en a 4 dans l'échantillon.

Solution. (a) Le rapport asymétrie / ES $= \frac{0,14}{0,192} = 0,73$; le rapport aplatissement / ES $= \frac{0,16}{0,381} = 0,42$; tous deux bien inférieurs à 2. La moyenne et la médiane sont pratiquement égales. Le test de Shapiro-Wilk n'indique pas d'écart significatif à la normalité. Avec 160 observations, le test a une bonne puissance ; le résultat non significatif est donc informatif. Le modèle normal est raisonnable.

(b) $z = \frac{6,5-5,58}{0,45} = 2,04$; Φ(2,04) = 0,9793 ; la proportion théorique au-dessus du seuil est de 2,1% (avec le z exact, 1,9%), soit environ 3 patients sur 160. Nous en observons 4 (2,5%). Très bonne concordance.

Interprétation. Le fait que seulement 4 patients aient une HbA1c ≥ 6,5%, alors que 29 ont un diabète confirmé par HGPO, n'est pas une contradiction statistique, mais clinique : l'HbA1c et la

glycémie à jeun identifient des personnes en partie différentes, et le diabète non diagnostiqué est souvent à un stade précoce. La corrélation entre la glycémie et l'HbA1c sera étudiée au chapitre 9.

3.13 Exercices proposés

1. La PAS de mg2026 a une moyenne de 131,13 mmHg et un ÉT de 13,88 mmHg. Calculez les scores z des valeurs 125 et 150 mmHg et les percentiles correspondants, à l'aide du tableau E.1, puis de la fonction NORM.S.DIST.
2. Le cholestérol total a une moyenne de 209,39 mg/dL et un ÉT de 36,91 mg/dL (Shapiro-Wilk p = 0,927). Calculez la proportion théorique de patients ayant un cholestérol ≥ 240 mg/dL et un cholestérol < 190 mg/dL. Dans l'échantillon, les proportions observées sont de 23,8% et 30,0%. Commentez.
3. Le HDL a une moyenne de 49,52 mg/dL et un ÉT de 11,78 mg/dL. Construisez l'intervalle $m \pm 1{,}96s$. Quelle proportion de patients devrait-il contenir ? Calculez la proportion théorique de ceux dont le HDL est < 40 mg/dL (observé : 19,4%).
4. Dans la population dont provient mg2026, la proportion de fumeurs actifs est d'environ 0,33. Pour un groupe de 6 patients choisis au hasard, calculez : la moyenne et la variance du nombre de fumeurs, la probabilité qu'il n'y ait aucun fumeur et la probabilité qu'il y en ait au moins 2.
5. Dans un service, on enregistre en moyenne 0,5 événement indésirable grave par mois. Calculez la probabilité qu'au moins un événement survienne au cours d'un mois et la probabilité qu'il en survienne 3 ou plus.
6. Dans une population de référence en bonne santé, la kaliémie a une moyenne de 4,3 mmol/L et un ÉT de 0,35 mmol/L, avec une distribution normale (données fictives). Construisez l'intervalle de référence à 95%. Un patient a 5,1 mmol/L : calculez le score z et le percentile.
7. Dans un article, la durée d'hospitalisation des patients atteints de diabète de type 1 est rapportée sous la forme 9,1 ± 5,8 jours. Sans autre information, que pouvez-vous dire de la forme de la distribution ? Quels indicateurs auraient été plus appropriés ?
8. Pour le score FINDRISC de mg2026 (n = 160) : moyenne 10,37, médiane 10, asymétrie 0,39, aplatissement 0,74, Shapiro-Wilk p = 0,034. Calculez les rapports asymétrie / ES et aplatissement / ES (ES 0,192 et 0,381) et indiquez, en le justifiant, quels indicateurs descriptifs et quel type de test vous choisiriez, en tenant compte du fait que le score ne prend que des valeurs entières entre 0 et 26.
9. La moyenne des logarithmes népériens d'une variable est de 4,97, et leur écart-type de 0,49. Calculez la moyenne géométrique, l'écart-type géométrique (en tant que facteur) et l'intervalle dans lequel se situent environ 95% des valeurs sur l'échelle d'origine.
10. La glycémie initiale de mg2026 a Q1 = 93 et $Q_3 = 116$ mg/dL, une moyenne de 105,44 et un ÉT de 15,76 mg/dL ; la valeur maximale est de 154 mg/dL. La valeur 154 est-elle extrême selon la règle 1,5 x EIQ ? Et selon le critère |z| > 3 ? Que feriez-vous de cette valeur ?
11. Décrivez à quoi ressemblerait le diagramme Q-Q pour : (a) une variable normale ; (b) la durée d'hospitalisation ; (c) une variable à queues plus épaisses que celles de la loi normale ; (d) une variable normale comportant une seule erreur de saisie (un zéro de trop).

Pièges fréquents

- Croire qu'un test de normalité non significatif "prouve" la normalité. Cela signifie seulement que nous n'avons pas trouvé de preuves contre elle, et dans les petits échantillons le test a une puissance réduite.

- Rejeter automatiquement les méthodes paramétriques dans un grand échantillon uniquement parce que Shapiro-Wilk donne p < 0,05. Regardez le graphique, l'ampleur de l'asymétrie et l'analyse qui suit.
- Tester la normalité sur l'ensemble de l'échantillon alors que le test exige la normalité dans chaque groupe ou la normalité des différences.
- Confondre la fonction de répartition et la densité : dans Google Sheets, NORM.DIST avec l'argument FALSE ne donne pas une probabilité.
- Croire que la standardisation (les scores z) transforme une distribution asymétrique en une distribution normale.
- Rapporter la moyenne ± ÉT pour des variables dont le m - 2s est négatif, alors que la variable ne peut pas être négative.
- Retransformer l'écart-type calculé sur l'échelle logarithmique comme s'il s'agissait d'un écart-type sur l'échelle d'origine, ou arrondir les limites de l'intervalle sur l'échelle logarithmique avant la retransformation.
- Supprimer les valeurs extrêmes sans vérification et sans le signaler, ou exclure des valeurs jusqu'à ce que le test de normalité devienne non significatif.
- Utiliser le test de Kolmogorov-Smirnov classique, sans la correction de Lilliefors, lorsque la moyenne et l'écart-type sont estimés à partir des mêmes données.
- Interpréter un résultat de laboratoire légèrement en dehors de l'intervalle de référence comme une preuve de maladie.

À retenir

- La distribution normale est entièrement déterminée par μ et σ ; elle est symétrique, sa moyenne, sa médiane et son mode coïncident, et 68,3%, 95,4% et 99,7% des valeurs se situent à 1, 2 et 3 écarts-types de la moyenne (95% à 1,96 écart-type).
- Le score z = (x - m) / s exprime la distance à la moyenne en écarts-types ; le tableau E.1 ou la fonction NORM.S.DIST le convertit en percentile. Dans mg2026, une PAS de 160 mmHg a un z = 2,08 (98e percentile).
- La loi binomiale (n épreuves, probabilité p, moyenne n x p) et la loi de Poisson (événements rares, moyenne = variance = λ) décrivent des comptages.
- L'intervalle de référence $m \pm 1{,}96s$ laisse, par construction, 5% des personnes en bonne santé en dehors de ses limites ; pour les variables asymétriques, on utilise des percentiles ou l'échelle logarithmique.
- Les distributions t, χ^2 et F sont des distributions d'échantillonnage, dont la forme dépend des degrés de liberté ; le théorème central limite fait que les moyennes sont presque normales dans des échantillons suffisamment grands.
- La normalité se juge à l'aide des graphiques, des indicateurs (l'asymétrie et l'aplatissement rapportés à leurs erreurs standard, la moyenne comparée à la médiane) et des tests (Shapiro-Wilk, Lilliefors, Anderson-Darling), pris ensemble.
- La transformation logarithmique normalise souvent les concentrations sériques ; la moyenne retransformée est la moyenne géométrique (triglycérides : 144,0 mg/dL), et les différences deviennent des rapports.
- Les valeurs extrêmes s'identifient (|z| > 3, 1,5 x EIQ, Grubbs), se vérifient à la source et se signalent ; on ne les supprime pas en silence.

Questions d'autoévaluation

1. (QRU) Quelle proportion des valeurs d'une distribution normale se situe entre μ - 2σ et μ + 2σ ?
    a) 68,3%
    b) 90,0%
    c) 95,4%
    d) 99,7%

e) 50,0%

2. (QRU) Dans le jeu de données mg2026, la PAS a une moyenne de 131,13 mmHg et un écart-type de 13,88 mmHg. Le score z d'une valeur de 160 mmHg est :
   a) 1,64
   b) 2,08
   c) 2,58
   d) 28,87
   e) -2,08
3. (QRU) Lequel des tests suivants est recommandé en premier lieu pour vérifier la normalité dans les petits et moyens échantillons ?
   a) le test du khi-deux d'association
   b) le test de Shapiro-Wilk
   c) le test t pour échantillons indépendants
   d) le test de Levene
   e) le test de Mann-Whitney
4. (QRU) Un article rapporte la durée d'hospitalisation sous la forme 11,6 ± 7,0 jours. La forme la plus probable de la distribution est :
   a) normale, puisque c'est la moyenne qui est rapportée
   b) asymétrique vers la droite (asymétrie positive)
   c) asymétrique vers la gauche (asymétrie négative)
   d) uniforme
   e) avec un aplatissement négatif et sans asymétrie
5. (QRU) Si l'on calcule la moyenne des logarithmes népériens d'une variable puis qu'on applique la fonction exponentielle, on obtient :
   a) la moyenne arithmétique
   b) la médiane exacte
   c) la moyenne géométrique
   d) la moyenne harmonique
   e) le mode
6. (QRM) À propos de la distribution normale, les affirmations suivantes sont vraies :
   a) elle est symétrique par rapport à la moyenne
   b) la moyenne, la médiane et le mode coïncident
   c) elle est entièrement déterminée par la moyenne et l'écart-type
   d) les queues de la courbe touchent l'axe horizontal à $\mu \pm 3\sigma$
   e) l'aire totale sous la courbe est égale à 1
7. (QRM) Sur un échantillon de 12 patients, le test de Shapiro-Wilk donne p = 0,30. Les affirmations correctes sont :
   a) le test prouve que les données sont normales
   b) on ne rejette pas l'hypothèse de normalité
   c) le test peut avoir une faible puissance sur un échantillon aussi petit
   d) il est obligatoire d'utiliser un test paramétrique
   e) il faut aussi examiner les graphiques (histogramme, Q-Q)
8. (QRM) Lesquels des éléments suivants indiquent une asymétrie positive ?

a) la moyenne est supérieure à la médiane
b) le rapport asymétrie / erreur standard est supérieur à 2
c) sur le diagramme Q-Q, les points s'incurvent vers le haut à l'extrémité droite
d) pour une variable qui ne peut pas être négative, m - 2s est négatif
e) l'aplatissement est égal à zéro

9. (QRM) À propos des valeurs extrêmes, les affirmations correctes sont :
a) elles sont supprimées automatiquement si leur score z dépasse 3 en valeur absolue
b) elles sont vérifiées dans le document source avant toute décision
c) une analyse de sensibilité, avec et sans elles, est recommandée
d) la règle 1,5 x EIQ repose sur les quartiles
e) le test de Grubbs suppose une distribution approximativement normale

10. (QRM) À propos d'un intervalle de référence construit comme $m \pm 1{,}96s$ sur une population en bonne santé, les affirmations correctes sont :

a) il contient environ 95% des personnes en bonne santé
b) environ 5% des personnes en bonne santé ont des valeurs en dehors de cet intervalle
c) une valeur en dehors de l'intervalle prouve la maladie
d) pour les variables asymétriques, on peut utiliser les percentiles 2,5 et 97,5 ou l'échelle logarithmique
e) pour 20 analyses indépendantes, environ 64% des personnes en bonne santé auront au moins un résultat en dehors de l'intervalle

Chapitre 4. Estimation des paramètres et intervalles de confiance ; principes du calcul de la taille de l'échantillon

Objectifs

À la fin de ce chapitre, l'étudiant sera capable de :

- distinguer l'estimation ponctuelle de l'estimation par intervalle et expliquer les propriétés d'un bon estimateur (absence de biais, convergence, efficacité) ;
- calculer l'erreur standard de la moyenne et ne pas la confondre avec l'écart-type ;
- construire, sur les mêmes données, les intervalles à 68,3%, 95,4% et 99,7% pour les valeurs individuelles et pour la moyenne, et expliquer pourquoi ils diffèrent autant ;
- calculer à la main un intervalle de confiance de la moyenne à l'aide de la loi t et le vérifier dans JASP, MedCalc et Google Sheets ;
- interpréter correctement un intervalle de confiance et reconnaître les interprétations erronées ;
- calculer l'intervalle de confiance d'une proportion par les méthodes de Wald et de Wilson et appliquer la règle de trois en l'absence d'événements ;
- construire des intervalles de confiance pour la différence de deux moyennes et de deux proportions, et lire les intervalles des OR, RR, AUC et ceux de données asymétriques ;
- calculer la taille de l'échantillon pour l'estimation d'une moyenne, d'une proportion et pour la comparaison de deux moyennes, en tenant compte de la puissance, de la taille de l'effet et des perdus de vue ;
- lire de façon critique les intervalles de confiance rapportés dans les articles publiés.

4.1 Estimation ponctuelle et estimation par intervalle

4.1.1 Paramètres, estimateurs, estimations

Au chapitre 1, nous avons distingué la population de l'échantillon. La population possède des paramètres, des valeurs fixes mais inconnues : la moyenne de la pression systolique μ de tous les adultes à risque cardiométabolique de l'ouest du pays, la proportion π de ceux qui ont un diabète non diagnostiqué. L'échantillon nous fournit des statistiques, calculées à partir des données : la moyenne m, la proportion p. Lorsque nous utilisons une statistique pour approcher un paramètre, nous l'appelons estimateur, et la valeur concrète obtenue sur notre échantillon s'appelle une estimation.

L'estimation ponctuelle fournit un seul nombre : "la moyenne de la PAS dans la population est estimée à 131,13 mmHg". C'est simple, mais incomplet. Si nous répétions l'étude sur 160 autres patients, nous obtiendrions presque certainement une autre moyenne, peut-être 129,8 ou 132,5 mmHg. Un nombre isolé ne dit rien de la confiance que nous pouvons lui accorder. L'estimation par intervalle complète l'information : "la moyenne de la PAS dans la population se situe, avec une confiance de 95%, entre 128,96 et 133,30 mmHg". L'intervalle exprime la précision de l'estimation, et sa largeur nous dit ce que nous savons réellement [42,53].

Les recommandations actuelles pour la présentation des études exigent que les résultats principaux soient présentés sous forme d'estimations assorties d'intervalles de confiance, et pas seulement de valeurs p. La raison deviendra claire d'ici la fin du chapitre : l'intervalle de confiance contient tout ce que contient la valeur p et, en plus, la taille de l'effet et sa précision [70].

4.1.2 Les propriétés d'un bon estimateur

Un estimateur est sans biais (en anglais unbiased) si, en moyenne, sur tous les échantillons possibles, il atteint le paramètre. La moyenne de l'échantillon est un estimateur sans biais de la moyenne de la population : certains échantillons la surestiment, d'autres la sous-estiment, mais sans tendance systématique. La simulation du chapitre 3 (tableau 3.5) a montré exactement cela : la moyenne des 10 000 moyennes est restée autour de 161 mg/dL pour tout n.

C'est ici qu'apparaît l'explication d'un détail qui intrigue les étudiants dès la première séance de travaux dirigés : pourquoi divisons-nous par n - 1, et non par n, dans la formule de l'écart-type ?

$$s^2 = \frac{\sum(x_i - m)^2}{n - 1}$$

La réponse tient à l'absence de biais. Les écarts sont calculés par rapport à la moyenne de l'échantillon m, et non par rapport à la vraie moyenne μ, que nous ne connaissons pas. La moyenne de l'échantillon est, par construction, le point par rapport auquel la somme des carrés des écarts est minimale ; les écarts par rapport à m sont donc systématiquement un peu plus petits que ceux par rapport à μ. Diviser par n sous-estimerait la variance de la population. Nous l'avons vérifié par simulation : sur 20 000 échantillons de 5 valeurs tirés d'une population normale avec σ = 14 mmHg (donc σ² = 196), la variance calculée avec n - 1 valait en moyenne 196,2, et celle calculée avec n seulement 157,0, soit exactement 196 x 4/5. La correction n - 1 (correction de Bessel) élimine ce biais. Pour n grand, la différence devient négligeable : pour la PAS de mg2026, l'écart-type avec n - 1 est de 13,88 mmHg, et avec n il serait de 13,84 mmHg. Dans Google Sheets, STDEV (ou STDEV.S) divise par n - 1, et STDEVP (STDEV.P) par n ; pour des échantillons, on utilise toujours la première.

Un estimateur est convergent s'il se rapproche de plus en plus du paramètre à mesure que n augmente. La moyenne de l'échantillon est convergente : son erreur standard $\frac{\sigma}{\sqrt{n}}$ tend vers zéro quand n augmente. Un échantillon plus grand ne garantit pas une estimation correcte si l'échantillon est biaisé (par exemple, uniquement des patients d'un hôpital universitaire), mais il garantit une estimation de plus en plus précise de ce qu'il mesure.

Un estimateur est efficace si, parmi les estimateurs sans biais, il a la plus faible variabilité. Pour des données normales, la moyenne comme la médiane estiment le centre de la distribution, mais la moyenne est plus efficace : l'erreur standard de la médiane vaut environ $\frac{1{,}25 \times \sigma}{\sqrt{n}}$. Pour la PAS de mg2026, elle serait de 1,25 x 1,097 = 1,37 mmHg, contre 1,10 mmHg pour la moyenne. Pour obtenir la même précision avec la médiane, il nous faudrait environ 57% de patients en plus. Pour des données asymétriques ou comportant des valeurs extrêmes, la situation peut s'inverser, et la médiane devient l'estimateur le plus stable. Le choix de l'estimateur dépend donc de la forme de la distribution, c'est-à-dire de tout ce dont nous avons parlé au chapitre 3.

4.2 La distribution d'échantillonnage de la moyenne et l'erreur standard

4.2.1 L'erreur standard de la moyenne

Le théorème central limite (section 3.7) nous a appris que les moyennes des échantillons de taille n ont une distribution approximativement normale, centrée sur μ, d'écart-type $\frac{\sigma}{\sqrt{n}}$. Cet écart-type des moyennes s'appelle l'erreur standard de la moyenne (en anglais standard error of the mean), notée ES ou SEM. Comme σ est inconnu, nous l'estimons par s :

$$\text{ES} = \frac{s}{\sqrt{n}}$$

Pour la PAS de mg2026 :

$$\text{ES} = \frac{13{,}88}{\sqrt{160}} = \frac{13{,}88}{12{,}649} = 1{,}097 \text{ mmHg}$$

L'erreur standard est près de 13 fois plus petite que l'écart-type, car √160 ≈ 12,65. Elle ne décrit pas la dispersion des patients, mais celle des moyennes que nous obtiendrions si nous répétions l'étude de nombreuses fois avec 160 patients chaque fois. Dans Google Sheets, elle se calcule avec =STDEV(M2:M161)/SQRT(COUNT(M2:M161)) ; JASP l'affiche dans Descriptives, option "Std. error of mean", et MedCalc la fournit automatiquement dans Summary statistics.

4.2.2 Écart-type ou erreur standard ?

C'est probablement la confusion la plus fréquente de la littérature médicale, et elle revient à l'examen année après année. Les deux grandeurs répondent à des questions différentes.

L'écart-type (ÉT) décrit la variabilité des données : à quel point les patients diffèrent les uns des autres. Il ne diminue pas quand l'échantillon grandit ; avec davantage de patients, nous l'estimons seulement de façon plus précise. Il est rapporté dans la partie descriptive (le tableau 1 d'un article), sous la forme "PAS $131{,}13 \pm 13{,}88$ mmHg".

L'erreur standard (ES) décrit la précision de la moyenne : à quel point la moyenne varierait d'un échantillon à l'autre. Elle diminue quand n augmente. Elle n'est généralement pas rapportée seule, mais sert à construire l'intervalle de confiance, beaucoup plus facile à interpréter.

Rapporter "moyenne ± ES" au lieu de "moyenne ± ÉT" fait paraître les données moins variables qu'elles ne le sont, et les barres d'erreur d'un graphique semblent plus courtes. Lorsque vous lisez un article, vérifiez toujours ce que représente le signe ±. Si ce n'est pas précisé et que n est connu, vous pouvez le déduire : un "ÉT" de 1,1 mmHg pour la pression systolique dans un échantillon de 160 patients est presque certainement une erreur standard. La règle de présentation recommandée est simple : moyenne ± ÉT pour décrire les données, moyenne avec IC à 95% pour exprimer la précision de l'estimation.

4.3 L'intervalle pour les valeurs individuelles et l'intervalle pour la moyenne

Dans la tradition de notre discipline, les premiers exercices d'estimation sont formulés ainsi : "estimez, avec une confiance de 95,4%, les valeurs de la taille des étudiants" et "estimez, avec une confiance de 95,4%, la moyenne de la taille dans la population des étudiants". Les formulations se ressemblent, mais elles demandent des choses différentes, et la différence est exactement celle qui sépare l'ÉT de l'ES.

L'intervalle pour les valeurs individuelles (appelé aussi intervalle de prédiction ou, au laboratoire, intervalle de référence) utilise l'écart-type :

$m \pm k \times s$, avec k = 1 (68,3%), 2 (95,4%) ou 3 (99,7%)

Il répond à la question : où se situent la plupart des patients ? Il suppose des valeurs individuelles approximativement normales (chapitre 3).

L'intervalle pour la moyenne utilise l'erreur standard :

$m \pm k \times$ ES, avec k = 1 (68,3%), 2 (95,4%) ou 3 (99,7%)

Il répond à la question : où se situe probablement la moyenne de la population ? Grâce au théorème central limite, il fonctionne aussi pour des données modérément asymétriques, si n est suffisamment grand.

Tableau 4.1. Les intervalles pour les valeurs individuelles et pour la moyenne de la PAS (mg2026 : m = 131,13 mmHg ; s = 13,88 mmHg ; ES = 1,097 mmHg ; n = 160).

Niveau de confiance	Pour les valeurs individuelles : $m \pm k \times s$	Largeur	Pour la moyenne : $m \pm k \times \text{ES}$	Largeur
68,3% ($k = 1$)	117,3-145,0 mmHg	27,8 mmHg	130,03-132,23 mmHg	2,2 mmHg
95,4% ($k = 2$)	103,4-158,9 mmHg	55,5 mmHg	128,94-133,33 mmHg	4,4 mmHg
99,7% ($k = 3$)	89,5-172,8 mmHg	83,3 mmHg	127,84-134,42 mmHg	6,6 mmHg

La différence entre les deux colonnes est spectaculaire (figure 4.1). L'intervalle à 95,4% pour les valeurs a une largeur de plus de 55 mmHg et englobe des patients allant d'une pression normale-basse jusqu'à une hypertension de grade II. L'intervalle à 95,4% pour la moyenne fait moins de 4,5 mmHg : nous savons assez précisément où se trouve la moyenne de la population, même si les patients diffèrent beaucoup les uns des autres. Le rapport entre les largeurs est $\sqrt{160} \approx 12{,}6$.

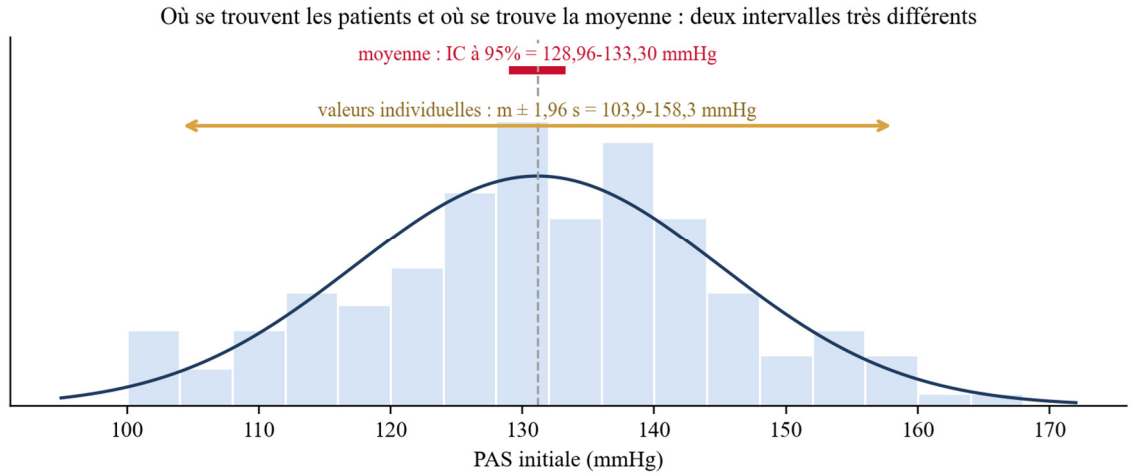


Figure 4.1. Histogramme de la PAS initiale du jeu de données mg2026, avec la courbe normale ajustée. La flèche jaune indique l'intervalle m ± 1,96 s, dans lequel se situent environ 95% des patients (103,9-158,3 mmHg). La courte barre rouge indique l'intervalle de confiance à 95% de la moyenne de la population (128,96-133,30 mmHg). Données synthétiques (jeu de données mg2026).

Pour l'intervalle exact à 95%, on utilise 1,96 au lieu de 2. Pour les valeurs individuelles : 103,9-158,3 mmHg (chap. 3) ; pour la moyenne, avec z : $131{,}13 \pm 1{,}96 \times 1{,}097 = 128{,}98 - 133{,}28$ mmHg. Les logiciels utilisent cependant pour la moyenne la loi t, et non la loi normale, et la différence apparaît à la deuxième décimale, comme nous allons le voir tout de suite.

4.4 L'intervalle de confiance de la moyenne avec la loi t

4.4.1 Pourquoi t et non z

La formule $m \pm 1{,}96 \times \text{ES}$ suppose que nous connaissons σ. En réalité, nous l'estimons par s, et cette estimation comporte elle aussi une incertitude, plus grande dans les petits échantillons. La loi t (section 3.6), avec ses queues plus épaisses, tient compte de cette incertitude supplémentaire. L'intervalle de confiance de la moyenne devient :

$$\text{IC} = m \pm \frac{t\left(\text{ddl};\, 1-\frac{\alpha}{2}\right) \times s}{\sqrt{n}}, \text{ avec ddl = n - 1}$$

où t(ddl ; 1 - α/2) est la valeur critique lue dans le tableau E.3 de l'annexe E. Pour un IC à 95%, α = 0,05, et nous cherchons la colonne 0,05 bilatérale (ou 0,975 cumulée).

Tableau 4.2. Valeurs critiques de t pour des intervalles de confiance à 95% et 99%.

n	ddl $= n - 1$	t pour un IC à 95%	t pour un IC à 99%
5	4	2,776	4,604
10	9	2,262	3,250
20	19	2,093	2,861
30	29	2,045	2,756
40	39	2,023	2,708
60	59	2,001	2,662
100	99	1,984	2,626
160	159	1,975	2,607
très grand	infini	1,960	2,576

Pour n = 160, la différence entre t (1,975) et z (1,96) est faible. Pour n = 10, la valeur t de 2,262 élargit l'intervalle de 15% par rapport à celui calculé avec z. Pour n = 5, de 42%. L'ancienne règle "pour n ≥ 30 on utilise z, pour n < 30 on utilise t" date de l'époque des tables imprimées ; aujourd'hui, les logiciels utilisent toujours t, parce que c'est correct pour tout n.

4.4.2 Le calcul complet pour la PAS de mg2026

Étape 1. Les statistiques descriptives : n = 160, m = 131,13 mmHg, s = 13,88 mmHg.

Étape 2. L'erreur standard : ES $= \frac{13{,}88}{\sqrt{160}} = 1{,}097$ mmHg.

Étape 3. La valeur critique : ddl = 159 ; t(159 ; 0,975) = 1,975.

Étape 4. La marge d'erreur : E = 1,975 x 1,097 = 2,167 mmHg.

Étape 5. L'intervalle : 131,13 - 2,167 = 128,96 mmHg et 131,13 + 2,167 = 133,30 mmHg.

IC à 95% de la moyenne de la PAS : 128,96-133,30 mmHg.

Pour un IC à 99%, t(159 ; 0,995) = 2,607, la marge est de 2,607 x 1,097 = 2,861 mmHg, et l'intervalle est 128,27-133,99 mmHg. Pour un IC à 90%, t(159 ; 0,95) = 1,654, et l'intervalle est 129,32-132,95 mmHg. Remarquez la règle : plus nous exigeons une confiance élevée, plus l'intervalle est large. Il n'existe pas d'intervalle à la fois très sûr et très étroit, sauf si nous augmentons la taille de l'échantillon.

Dans Google Sheets, la marge d'erreur s'obtient directement avec =CONFIDENCE.T(0,05; STDEV(M2:M161); COUNT(M2:M161)), qui donne 2,167. La fonction CONFIDENCE.NORM (ou l'ancienne CONFIDENCE) utilise z au lieu de t et donne 2,151. Les limites se calculent ensuite comme =AVERAGE(M2:M161) - CONFIDENCE.T(...) et =AVERAGE(M2:M161) + CONFIDENCE.T(...). Dans JASP, l'intervalle apparaît dans T-Tests > One Sample T-Test, avec les options "Location parameter" et "Confidence interval" cochées [48,49] ; dans MedCalc, dans Statistics > Summary statistics, à la ligne "95% CI for the mean" [50].

Pour la glycémie initiale, avec $m = 105{,}44$ mg/dL, $s = 15{,}76$ mg/dL et n = 160, ES = 1,246 mg/dL, et l'IC à 95% est $105{,}44 \pm 1{,}975 \times 1{,}246$, c'est-à-dire 102,98-107,90 mg/dL. La légère asymétrie de la glycémie (chapitre 3) n'affecte pratiquement pas cet intervalle, car, pour n = 160, la distribution de la moyenne est presque normale.

4.4.3 Un petit échantillon : 10 patients

Supposons que nous n'ayons eu que 10 patients, tirés au hasard de mg2026. Leurs valeurs de PAS, ordonnées, sont : 102, 103, 110, 130, 131, 136, 137, 147, 150, 154 mmHg.

La moyenne : la somme est de 1300, donc m = 130,0 mmHg.

L'écart-type : les écarts à la moyenne sont -28, -27, -20, 0, 1, 6, 7, 17, 20, 24 ; leurs carrés sont 784, 729, 400, 0, 1, 36, 49, 289, 400, 576, de somme 3264. La variance $s^2 = \frac{3264}{9} = 362{,}67$, et s = 19,04 mmHg.

L'erreur standard : ES $= \frac{19{,}04}{\sqrt{10}} = \frac{19{,}04}{3{,}162} = 6{,}022$ mmHg.

La valeur critique : ddl = 9, t = 2,262.

La marge : 2,262 x 6,022 = 13,62 mmHg.

IC à 95% : 130,0 ± 13,62, c'est-à-dire 116,38-143,62 mmHg, avec une largeur de 27,2 mmHg.

Avec z au lieu de t, nous aurions obtenu 118,20-141,80 mmHg, un intervalle trop étroit, qui aurait exagéré la précision. Comparez avec l'intervalle calculé sur 160 patients, d'une largeur de 4,3 mmHg. Avec 10 patients, la moyenne de la population pourrait se trouver n'importe où entre "normale" et "hypertension de grade I" ; avec 160, nous la connaissons avec une précision de ±2 mmHg.

4.5 L'interprétation correcte d'un intervalle de confiance

4.5.1 Ce que signifie "95%"

L'interprétation correcte d'un IC à 95% est de type fréquentiste et porte sur la procédure, non sur un intervalle particulier : si nous répétions l'étude un très grand nombre de fois, chaque fois avec un nouvel échantillon et en calculant l'intervalle de la même façon, environ 95% des intervalles obtenus contiendraient la vraie moyenne de la population. Les 5% restants la manqueraient, en raison des fluctuations d'échantillonnage.

La figure 4.2 rend cette idée visible. Nous avons simulé 50 études, chacune portant sur 25 patients tirés d'une population dont la moyenne de PAS est connue (131 mmHg) et l'écart-type de 14 mmHg, et nous avons calculé un IC à 95% pour chacune. Les intervalles ont des positions et des largeurs différentes, car chaque échantillon a une moyenne et un écart-type différents. 48 d'entre eux contiennent la vraie moyenne ; 2 (signalés en rouge) ne la contiennent pas. La proportion de 96% est très proche de 95%, comme prévu. Dans la vie réelle, nous ne faisons qu'une seule étude et nous ne savons pas si notre intervalle fait partie des 95% "bons" ou des 5% "ratés".

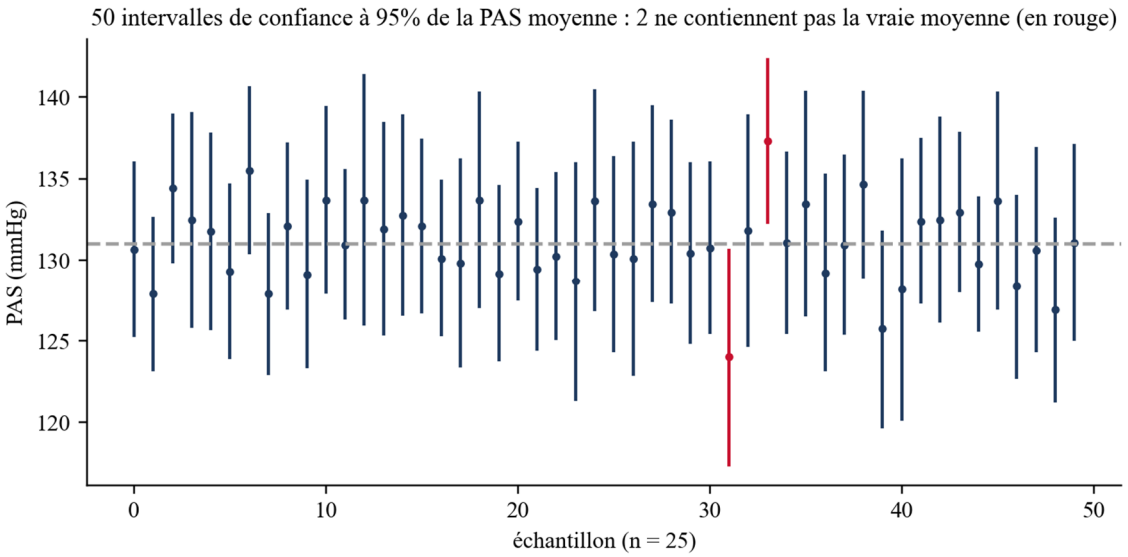


Figure 4.2. Cinquante intervalles de confiance à 95% de la moyenne de la PAS, calculés sur 50 échantillons simulés de 25 patients tirés d'une population de moyenne 131 mmHg (ligne discontinue). Deux intervalles, signalés en rouge, ne contiennent pas la vraie moyenne. Simulation.

Dans le langage courant, nous pouvons dire : "nous pouvons affirmer, avec une confiance de 95%, que la moyenne de la population se situe entre 128,96 et 133,30 mmHg", étant entendu que la "confiance" porte sur la méthode de construction, qui réussit dans 95% des cas. L'approche bayésienne, qui ne fait pas l'objet de ce cours, permet des affirmations directes de probabilité sur le paramètre, mais elle exige des informations supplémentaires (une distribution a priori).

4.5.2 Les interprétations erronées

Les formulations suivantes apparaissent fréquemment dans les travaux et à l'examen, et elles sont fausses :

- "Il y a une probabilité de 95% que la moyenne de la population soit comprise entre 128,96 et 133,30 mmHg." Dans l'approche fréquentiste, la moyenne de la population est une constante : soit elle est dans l'intervalle, soit elle n'y est pas. La probabilité de 95% porte sur la procédure.
- "95% des patients ont une PAS comprise entre 128,96 et 133,30 mmHg." C'est la confusion entre l'intervalle pour la moyenne et l'intervalle pour les valeurs. En réalité, seulement 25 des 160 patients (15,6%) ont une PAS dans cet intervalle étroit.
- "Si nous répétons l'étude, la moyenne du nouvel échantillon se trouvera dans cet intervalle avec une probabilité de 95%." Non : le nouvel échantillon aura sa propre variabilité, et la probabilité que sa moyenne tombe dans notre intervalle est, en moyenne, d'environ 83%, et non de 95%.
- "Les valeurs du centre de l'intervalle sont sûres, celles des extrémités impossibles." Les valeurs proches de l'estimation ponctuelle sont plus compatibles avec les données que celles des extrémités, mais il n'existe pas de frontière brutale aux limites de l'intervalle [70].
- "Un intervalle plus large signifie un effet plus important." La largeur exprime la précision (qui dépend de n et de la variabilité), non la taille de l'effet.

4.6 La largeur de l'intervalle de confiance et la taille de l'échantillon

La largeur d'un IC de la moyenne est $\frac{2 \times t \times s}{\sqrt{n}}$. Elle dépend de trois éléments : le niveau de confiance (par t), la variabilité des données (par s) et la taille de l'échantillon (par $\sqrt{n}$). Le premier, c'est nous qui le choisissons ; le deuxième relève de la biologie et de la qualité de la mesure ; le troisième relève de la conception de l'étude. Le message le plus important est que la largeur

diminue avec la racine carrée de n, et non avec n : pour diviser par deux la largeur de l'intervalle, il faut multiplier la taille de l'échantillon par quatre (figure 4.3).

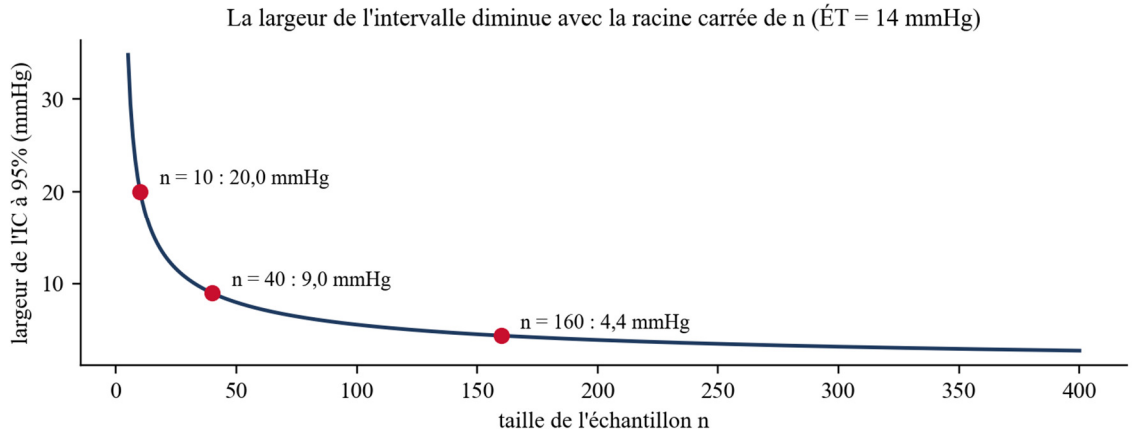


Figure 4.3. Largeur de l'intervalle de confiance à 95% de la moyenne de la PAS en fonction de la taille de l'échantillon, pour un écart-type de 14 mmHg. La courbe descend rapidement au début, puis de plus en plus lentement : doubler la précision exige quatre fois plus de patients. Courbe théorique.

Sur les données réelles de mg2026, nous avons tiré au hasard un sous-échantillon de 10 patients et un autre de 40 patients, et nous avons calculé l'IC à 95% de la moyenne de la PAS (tableau 4.3).

Tableau 4.3. L'intervalle de confiance à 95% de la moyenne de la PAS en fonction de la taille de l'échantillon (jeu de données mg2026).

n	Moyenne (mmHg)	IC à 95% (mmHg)	Largeur (mmHg)	Largeur théorique pour s = 14 mmHg
10	130,0	116,38-143,62	27,2	20,0
40	128,4	123,08-133,67	10,6	9,0
160	131,1	128,96-133,30	4,3	4,4

Les largeurs observées diffèrent des largeurs théoriques, car chaque sous-échantillon a son propre écart-type : le sous-échantillon de 10 patients a eu, par hasard, s = 19,04 mmHg, plus que celui de l'échantillon complet, et celui de 40, s = 16,54 mmHg. La tendance est pourtant claire : de 10 à 40 patients (quatre fois plus), la largeur passe de 27,2 à 10,6 mmHg ; de 40 à 160 (de nouveau quatre fois plus), de 10,6 à 4,3 mmHg. Remarquez aussi que les trois intervalles contiennent tous la moyenne de l'échantillon complet, et que les intervalles des plus petits échantillons l'"attrapent" avec une marge plus large. Lors des travaux pratiques 4, vous répéterez cette simulation des dizaines de fois dans Google Sheets et vous verrez comment les intervalles varient d'un échantillon à l'autre.

La conséquence pratique pour la recherche est importante : le gain de précision a un rendement décroissant. Passer de 40 à 160 patients fait gagner 6,3 mmHg de largeur, passer de 160 à 640 ne ferait gagner que 2,2 mmHg. C'est pourquoi le calcul de la taille de l'échantillon (section 4.13) part de la question "de quelle précision avons-nous besoin ?", et non de "combien de patients pouvons-nous réunir ?".

Étude de cas : l'intervalle de confiance d'une variation moyenne

Muntean et ses collaborateurs ont suivi 921 étudiants de première année et ont rapporté une variation moyenne de l'IMC de -0,45 ± 1,07 kg/m², avec un IC à 95% allant de -0,52 à -0,38 kg/m² ($p < 0,001$) [5]. Nous pouvons vérifier l'intervalle à la main. L'erreur standard de la différence moyenne est de 1,07 / √921 = 1,07 / 30,35 = 0,0353 kg/m². La marge d'erreur est $1,96 \times 0,0353 = \frac{0,069\ \text{kg}}{m^2}$ (pour 920 degrés de liberté, t = 1,963, pratiquement 1,96). L'intervalle

> est -0,45 ± 0,069, c'est-à-dire de -0,519 à -0,381, ce qui, une fois arrondi, coïncide avec l'intervalle publié. La leçon des chiffres : l'écart-type des variations (1,07) est grand par rapport à la moyenne, mais avec 921 participants l'erreur standard est 30 fois plus petite, et l'intervalle est très étroit. Nous savons très précisément que l'IMC moyen a diminué, d'environ une demi-unité ; dans le même temps, 13,1% des étudiants ont pris du poids. La précision d'une moyenne ne dit rien de l'hétérogénéité des individus. Un second détail mérite l'attention : l'intervalle exclut zéro, et p < 0,001 ; nous verrons à la section 4.12 que ces deux informations sont liées.

4.7 L'intervalle de confiance d'une proportion

4.7.1 La méthode de Wald

Une proportion calculée sur un échantillon est elle aussi une estimation, avec sa propre erreur standard. Si, dans l'échantillon, k sujets sur n présentent la caractéristique étudiée, la proportion observée est $p_i = \frac{k}{n}$, et son erreur standard est :

$$\mathrm{ES}(p_i) = \sqrt{\frac{p_i \times (1 - p_i)}{n}}$$

La méthode la plus simple de calcul de l'intervalle, appelée méthode de Wald, repose sur l'approximation normale de la loi binomiale (section 3.2.1) :

$$\mathrm{IC}\,95\% = p_i \pm 1{,}96 \times \sqrt{\frac{p_i \times (1 - p_i)}{n}}$$

Dans mg2026, 29 des 160 patients ont un diabète confirmé par HGPO : $p_i = \frac{29}{160} = 0{,}181 (18{,}1\%)$. L'erreur standard est $\sqrt{\frac{0{,}181 \times 0{,}819}{160}} = \sqrt{0{,}000927} = 0{,}0305$. Marge d'erreur : 1,96 x 0,0305 = 0,0597. Intervalle de Wald : 0,181 ± 0,060, c'est-à-dire 12,2%-24,1%.

4.7.2 La méthode de Wilson et pourquoi elle est préférée

La méthode de Wald est facile à calculer à la main, mais elle a des défauts connus. L'intervalle est toujours symétrique, alors que la distribution d'une petite proportion ne l'est pas (elle ne peut pas descendre sous 0, mais peut monter beaucoup). Pour des proportions proches de 0 ou de 1 et pour de petits échantillons, l'intervalle de Wald peut descendre sous 0% ou dépasser 100% et, plus grave, il couvre le vrai paramètre moins souvent qu'il ne le promet : un "IC à 95%" de Wald peut avoir, en réalité, une couverture de 85 à 90%. C'est pourquoi les recommandations et les logiciels modernes préconisent la méthode de Wilson (score de Wilson) ou la méthode exacte de Clopper-Pearson [42,53].

La formule de Wilson paraît intimidante, mais elle peut se calculer pas à pas :

centre = (p_i + z^2 / (2n)) / (1 + z^2 / n) marge = $\left(\frac{z}{1+\frac{z^2}{n}}\right) \times \sqrt{\frac{p_i \times (1-p_i)}{n} + \frac{z^2}{4\,n^2}}$

Pour 29/160, avec $z = 1{,}96 (z^2 = 3{,}8416)$:

$\frac{z^2}{n} = \frac{3{,}8416}{160} = 0{,}0240$; $\frac{z^2}{2n} = 0{,}0120$ centre = (0,1813 + 0,0120) / 1,0240 = 0,1887 sous la racine : $\frac{0{,}1813 \times 0{,}8187}{160} + \frac{3{,}8416}{102400} = 0{,}000927 + 0{,}0000375 = 0{,}000965$; racine = 0,0311 marge = $\left(\frac{1{,}96}{1{,}0240}\right) \times 0{,}0311 = 1{,}914 \times 0{,}0311 = 0{,}0595$; IC 95% Wilson : 0,1887 ± 0,0595, c'est-à-dire 12,9%-24,8%

Remarquez deux choses. Le centre de l'intervalle de Wilson (18,9%) est légèrement déplacé vers 50% par rapport à la proportion observée (18,1%), et l'intervalle n'est plus symétrique autour de 18,1% : 5,2 points de pourcentage en dessous et 6,7 au-dessus. Pour n = 160 et p_i = 18%, la différence avec Wald est faible. La méthode exacte de Clopper-Pearson donne 12,5%-25,0%, un peu plus large, parce qu'elle est conservatrice.

Les différences deviennent importantes dans les petits échantillons. Si, sur 20 patients, 2 présentent une complication (10%), l'intervalle de Wald est de 10% ± 13,1%, c'est-à-dire de -3,1% à 23,1%, une limite inférieure impossible, que les logiciels "coupent" à 0. L'intervalle de Wilson est de 2,8%-30,1%, et l'intervalle exact de 1,2%-31,7%. Tous deux sont asymétriques et réalistes. Conclusion pratique : pour les proportions, rapportez l'intervalle de Wilson (ou exact) ; la méthode de Wald n'est utile que pour comprendre l'idée et pour des calculs rapides à la main, avec un n grand et un p_i éloigné de 0 et de 1. Dans MedCalc, l'intervalle exact apparaît dans Statistics > Summary statistics pour les variables dichotomiques ou dans le calculateur pour une proportion ; dans JASP, dans Frequencies > Binomial Test, qui affiche l'intervalle exact de Clopper-Pearson. Dans Google Sheets, il n'existe pas de fonction dédiée, mais la formule de Wilson peut s'écrire dans une cellule, comme nous le ferons lors des travaux pratiques 4.

4.7.3 Zéro événement : la règle de trois

Une situation particulière, très fréquente en médecine : une étude portant sur 50 patients traités par un nouveau médicament n'observe aucune réaction indésirable grave. La proportion observée est de 0/50 = 0%, et l'intervalle de Wald est de 0% ± 0, c'est-à-dire "exactement zéro", ce qui est évidemment absurde. Personne ne soutiendrait que le risque réel est nul simplement parce que nous ne l'avons pas observé chez 50 patients.

La règle de trois offre une solution simple : lorsqu'aucun événement n'est observé dans un échantillon de n sujets, la limite supérieure de l'IC à 95% de la proportion réelle est d'environ 3 / n. Pour 50 patients : 3/50 = 6%. Le risque réel pourrait donc atteindre environ 6%, soit jusqu'à 1 patient sur 17. La méthode exacte donne 7,1% ; la règle fonctionne mieux pour des n plus grands (pour 100 patients, 3% contre 3,6% avec la méthode exacte). Pour 30 patients sans événement, la limite est de 3/30 = 10%. La règle explique pourquoi les réactions indésirables rares (1 sur 10 000) ne peuvent pas être détectées dans les essais cliniques de phase III portant sur quelques milliers de patients, et pourquoi la pharmacovigilance après autorisation est indispensable.

Étude de cas : intervalles de Wilson pour des prévalences

Moleriu et ses collaborateurs ont étudié 136 patients d'une clinique de gastro-entérologie, dont 52 présentant un phénotype de trouble de l'interaction intestin-cerveau et 84 témoins asymptomatiques [71]. Seulement 4,4% d'entre eux (IC à 95% 2,0-9,3%) avaient un diagnostic formel de syndrome de l'intestin irritable, alors que 47,8% (IC à 95% 39,6-56,1%) remplissaient les critères d'un ensemble de symptômes compatible, d'où un rapport de sous-diagnostic de 10,8. Regardez le premier intervalle : 4,4% se situe à 2,4 points de la limite inférieure et à 4,9 points de la limite supérieure. C'est l'asymétrie typique d'un intervalle de Wilson ou exact pour une petite proportion. Nous pouvons vérifier : 4,4% de 136 correspond à 6 patients, et l'intervalle de Wilson pour 6/136 est exactement 2,0-9,3% ; pour 65/136 (47,8%), l'intervalle de Wilson est 39,6-56,1%, là encore identique à celui qui a été publié. Un intervalle de Wald pour 6/136 aurait donné 1,0-7,9%, symétrique et trop étroit dans sa partie haute. Le second intervalle, pour une proportion proche de 50%, est presque symétrique, car à ce niveau les méthodes donnent des résultats similaires. Mischie et ses collaborateurs ont explicitement rapporté des prévalences avec des intervalles de Wilson dans une enquête en ligne auprès de 753 jeunes de 18 à 30 ans, avec 11,6% de personnes en insuffisance pondérale et 29,3% en surpoids ou obèses [20]. À titre d'exercice, l'intervalle de Wilson que nous avons calculé pour

> 29,3% de 753 est d'environ 26,2-32,7%. La précision est bonne grâce au grand effectif, mais l'intervalle ne corrige pas un autre défaut, que les auteurs reconnaissent dans le titre même : l'échantillon est un échantillon de commodité, avec 93,5% de femmes. L'intervalle de confiance ne mesure que l'erreur aléatoire d'échantillonnage, pas l'erreur systématique de sélection.

4.8 L'intervalle de confiance de la différence de deux moyennes et de deux proportions

4.8.1 La différence de deux moyennes

Lorsque nous comparons deux groupes indépendants, le paramètre qui nous intéresse est la différence entre les moyennes des populations, µ1 - µ2. L'estimation ponctuelle est m1 - m2, et son erreur standard combine les erreurs standard des deux moyennes :

$$\mathrm{ES}(m_1 - m_2) = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

L'intervalle de confiance est $(m_1 - m_2) \pm t \times \mathrm{ES}(m_1 - m_2)$, avec un nombre de degrés de liberté calculé par la formule de Welch (les logiciels le calculent automatiquement) ou, si les variances sont égales, n1 + n2 - 2.

Dans mg2026, la PAS est de 130,85 ± 13,25 mmHg chez les 74 hommes et de 131,37 ± 14,47 mmHg chez les 86 femmes. La différence hommes moins femmes est de -0,52 mmHg. L'erreur standard :

$$\mathrm{ES} = \sqrt{\frac{13{,}25^2}{74} + \frac{14{,}47^2}{86}} = \sqrt{2{,}372 + 2{,}435} = \sqrt{4{,}807} = 2{,}193 \text{ mmHg}$$

Avec 157,4 degrés de liberté (Welch), t = 1,975, et la marge est de 1,975 x 2,193 = 4,33 mmHg. IC à 95% de la différence : -0,52 ± 4,33, c'est-à-dire de -4,9 à 3,8 mmHg.

L'intervalle contient zéro : les données sont compatibles aussi bien avec une PAS légèrement plus élevée chez les hommes (jusqu'à 3,8 mmHg) qu'avec une PAS plus élevée chez les femmes (jusqu'à 4,9 mmHg), et bien sûr avec une absence de différence. L'intervalle dit toutefois quelque chose de plus que "la différence n'est pas significative" (p = 0,81, chapitre 6) : une différence réelle de 10 mmHg entre les sexes, dans cette population, est très peu probable, car elle se situe bien en dehors de l'intervalle. Un résultat non significatif avec un intervalle étroit est informatif ; un résultat avec un intervalle très large ne nous dit presque rien.

4.8.2 La différence de deux proportions

Pour deux proportions indépendantes $p_1 = \frac{k_1}{n_1}$ et $p_2 = \frac{k_2}{n_2}$:

$$\mathrm{ES}(p_1 - p_2) = \sqrt{\frac{p_1 \times (1 - p_1)}{n_1} + \frac{p_2 \times (1 - p_2)}{n_2}}$$

Dans l'étude simulée mg2026, une perte de poids d'au moins 5% à 6 mois a été obtenue chez 16 des 54 patients du bras C (29,6%) et chez 6 des 53 patients du bras A (11,3%). La différence absolue est de 18,3 points de pourcentage. L'erreur standard :

$$\mathrm{ES} = \sqrt{\frac{0{,}296 \times 0{,}704}{54} + \frac{0{,}113 \times 0{,}887}{53}} = \sqrt{0{,}00386 + 0{,}00189} = \sqrt{0{,}00576} = 0{,}0759$$

IC à 95% : $0{,}183 \pm 1{,}96 \times 0{,}0759 = 0{,}183 \pm 0{,}149$, c'est-à-dire de 3,4 à 33,2 points de pourcentage. L'intervalle exclut zéro ; la différence est donc statistiquement significative. Il est cependant très large : le bénéfice réel de la téléconsultation pourrait être modeste (3 patients de plus sur 100) ou très important (33 sur 100). C'est à partir de cette différence que l'on calcule le nombre de sujets à traiter, $\text{NNT} = \frac{1}{0{,}183} = 5{,}5$, avec un intervalle de 3,0 à 29,1 patients, dont nous parlerons au chapitre 10.

4.9 Intervalles de confiance pour des données asymétriques

4.9.1 L'échelle logarithmique et la moyenne géométrique

Pour une variable log-normale, comme les triglycérides de mg2026, l'intervalle de confiance se calcule sur l'échelle logarithmique, puis se retransforme (section 3.9.4). La moyenne des logarithmes est de 4,970, leur écart-type de 0,488, l'erreur standard de 0,488 / √160 = 0,0386, et l'IC à 95% sur l'échelle logarithmique est $4{,}9699 \pm 1{,}975 \times 0{,}0386$, c'est-à-dire de 4,8937 à 5,0461. Après retransformation, nous obtenons $e^{4{,}8937} = 133{,}44$ et $e^{5{,}0461} = 155{,}42$ mg/dL, autour de la moyenne géométrique de 144,0 mg/dL (les limites sur l'échelle logarithmique sont conservées avec toutes leurs décimales jusqu'à la retransformation). L'intervalle de la moyenne arithmétique, calculé directement avec la formule t, serait de 149,26-173,09 mg/dL. Les deux intervalles ne se contredisent pas : ils estiment des paramètres différents. Le premier estime le centre "typique" d'une distribution asymétrique, le second la moyenne arithmétique, tirée vers le haut par la queue droite. Lorsque vous rapportez vos résultats, dites explicitement ce que vous avez estimé.

4.9.2 L'intervalle de confiance de la médiane

Pour les variables qui n'admettent pas de transformation adaptée, on peut estimer directement l'intervalle de confiance de la médiane. La méthode classique, fondée sur la loi binomiale, ne suppose aucune forme particulière de distribution. On range les valeurs par ordre croissant et on calcule les rangs :

rang inférieur $\approx \frac{n}{2} - \frac{1{,}96 \times \sqrt{n}}{2}$; rang supérieur $\approx \frac{n}{2} + \frac{1{,}96 \times \sqrt{n}}{2}$ (arrondis vers l'extérieur)

Pour n = 160 : $80 \pm \frac{1{,}96 \times 12{,}65}{2} = 80 \pm 12{,}4$, donc les rangs 67 et 93. Les valeurs situées à ces positions donnent l'intervalle. Pour les triglycérides (médiane 151,5 mg/dL), la 67e et la 93e valeur ordonnée sont 134 et 162 mg/dL : IC à 95% de la médiane 134-162 mg/dL. Pour la PAS (médiane 131 mmHg), l'intervalle est 129-134 mmHg, un peu plus large que celui de la moyenne (128,96-133,30 mmHg, largeur de 4,3 contre 5), ce qui illustre l'efficacité moindre de la médiane pour des données normales. MedCalc calcule automatiquement l'intervalle de confiance de la médiane dans Summary statistics.

4.10 Intervalles de confiance des rapports (OR, RR) et de l'aire ROC

4.10.1 Pourquoi les intervalles des OR et des RR sont asymétriques

Le rapport de cotes (OR) et le risque relatif (RR), que nous étudierons au chapitre 10, sont des rapports : ils prennent des valeurs entre 0 et l'infini, la valeur 1 correspondant à "aucune association". Leur distribution d'échantillonnage est asymétrique, mais leur logarithme a une distribution approximativement normale. L'intervalle se calcule donc sur l'échelle logarithmique, puis se retransforme, comme pour la moyenne géométrique.

Pour l'association obésité-diabète dans mg2026 (tableau 2 x 2 avec a = 18, b = 36, c = 11, d = 95), $\mathrm{OR} = \frac{18\times95}{36\times11} = 4{,}32$. Sur l'échelle logarithmique :

$$\ln(\mathrm{OR}) = \ln(4{,}32) = 1{,}463;\ \mathrm{ES}(\ln \mathrm{OR}) = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{18}+\frac{1}{36}+\frac{1}{11}+\frac{1}{95}} =$$

$\sqrt{0{,}1848} = 0{,}430 \ln(\mathrm{OR}) \pm 1{,}96 \times 0{,}430 = 1{,}463 \pm 0{,}843$, c'est-à-dire de 0,620 à 2,306 IC à 95% de l'OR : de $e^{0{,}620} = 1{,}86$ à $e^{2{,}306} = 10{,}03$

L'intervalle est fortement asymétrique autour de 4,32 : 2,46 unités en dessous, 5,71 au-dessus. Cette asymétrie n'est pas un défaut, mais la conséquence de l'échelle multiplicative. Une vérification utile : l'estimation ponctuelle est la moyenne géométrique des limites, $\sqrt{1{,}86 \times 10{,}03} = 4{,}32$. Pour le RR de la perte de poids dans le bras C par rapport au bras A (16/54 contre 6/53), RR = 2,62, et l'IC à 95%, calculé de la même façon, est 1,11-6,17.

4.10.2 L'intervalle de l'aire sous la courbe ROC

L'aire sous la courbe ROC (AUC), que nous étudierons au chapitre 12, prend des valeurs entre 0,5 (test inutile) et 1 (test parfait). Son intervalle de confiance est lui aussi souvent asymétrique lorsque l'AUC est proche de 1, car la limite supérieure ne peut pas dépasser 1. Dans mg2026, le score FINDRISC a une AUC de 0,852 (IC à 95% 0,762-0,943) pour le diabète confirmé par HGPO. Gherbon et ses collaborateurs rapportent pour l'âge, comme prédicteur de la détresse sévère liée au diabète, une AUC de 0,726 (IC à 95% 0,602-0,827), avec seulement 19 cas, ce qui explique la grande largeur de l'intervalle [69]. La règle de lecture est la même que pour toute estimation : nous regardons d'abord la limite inférieure. Une AUC de 0,726 avec une limite inférieure de 0,602 décrit un prédicteur allant de "presque inutile" à "acceptable".

Moga et ses collaborateurs rapportent pour l'élastographie time-harmonic, dans le diagnostic de la cirrhose, une AUROC de 0,90 (IC à 95% 0,82-0,93) [3] : l'estimation est beaucoup plus proche de la limite supérieure que de la limite inférieure (0,08 en dessous, 0,03 au-dessus), une asymétrie naturelle pour une grandeur bornée à 1. La même étude rapporte aussi, avec des intervalles de confiance, la reproductibilité des mesures, sur un sous-groupe de 27 patients : un accord global entre examinateurs de 0,94 (IC à 95% 0,89-0,97) et des coefficients de corrélation intraclasse de 0,92 (0,82-0,96), 0,94 (0,87-0,97) et 0,97 (0,95-0,99). L'intervalle de 0,92 s'étend sur 14 centièmes, ce qui reflète la petite taille du sous-groupe ; une reproductibilité "excellente" estimée sur 27 patients reste compatible avec une reproductibilité seulement "bonne".

Rațiu et ses collaborateurs ont comparé trois méthodes de diagnostic de l'obstruction biliaire maligne chez 160 patients, avec une AUROC de 0,88 (IC à 95% 0,75-0,90) pour la cholangio-IRM, de 0,94 (0,85-0,99) pour la CPRE et de 0,80 (0,70-0,82) pour l'histologie [57]. Deux observations de lecture critique. Les intervalles se chevauchent en partie, mais le chevauchement de deux intervalles ne signifie pas automatiquement que la différence n'est pas significative ; la comparaison de deux AUC mesurées sur les mêmes patients exige un test dédié, qui tient compte de la corrélation entre elles (chapitre 12). Ensuite, pour la cholangio-IRM et l'histologie, l'estimation ponctuelle se situe très près de la limite supérieure (0,88 dans un intervalle qui ne monte que jusqu'à 0,90 ; 0,80 dans un intervalle qui monte jusqu'à 0,82). Ce schéma est inhabituel et peut provenir de la méthode de calcul ou d'une erreur de rédaction ; il vaut la peine de le vérifier dans le texte intégral avant de réutiliser ces chiffres.

> **Étude de cas : des rapports de cotes avec intervalles de confiance, lus correctement**
>
> Trois études de collègues de la discipline, menées sur des étudiants et des patients de Timișoara, illustrent la lecture des intervalles des rapports. Moleriu et ses collaborateurs ont analysé 921 étudiants, dont 24,1% en surpoids ou obèses, et ont rapporté des rapports de cotes ajustés de 7,20 (IC à 95% 4,26-12,17) et de 1,98 (IC à 95% 1,26-3,10), ainsi qu'un OR de 1,43

pour chaque point d'un score de risque alimentaire (IC à 95% 1,23-1,66) [4]. Vérifiez la propriété de la moyenne géométrique : $\sqrt{4{,}26 \times 12{,}17} = 7{,}20$, $\sqrt{1{,}26 \times 3{,}10} = 1{,}98$, $\sqrt{1{,}23 \times 1{,}66} = 1{,}43$. Les trois intervalles excluent tous 1 ; les associations sont donc significatives, mais le premier est large : les données sont compatibles avec des cotes multipliées par 4 comme par 12. Piroș et ses collaborateurs ont trouvé, sur 693 étudiants, une prévalence du prédiabète de 21,1% et un OR de 1,78 par catégorie de profil alimentaire (IC à 95% 1,38-2,30 ; p < 0,001) [6], un intervalle plus étroit, qui exclut confortablement la valeur 1. À l'opposé, Popa et ses collaborateurs, en comparant 100 femmes et 113 hommes traités aux urgences pour des infections bactériennes suspectées, ont rapporté un OR de 0,71 (IC à 95% 0,34-1,49) pour un critère clinique observé chez 14 femmes (14,0%) et 21 hommes (18,6%), et un OR de 0,87 (IC à 95% 0,31-2,43) pour la mortalité hospitalière, avec 7 et 9 décès [9]. Ces intervalles incluent 1 et sont très larges : pour la mortalité, les données sont compatibles aussi bien avec des cotes trois fois plus faibles qu'avec des cotes plus de deux fois plus élevées chez les femmes. La conclusion correcte n'est pas "il n'y a pas de différence entre les sexes", mais "l'étude ne comporte pas assez d'événements pour répondre à la question". La même étude a trouvé, pour l'origine urinaire de l'infection, un OR de 2,14 (IC à 95% 1,20-3,81). Par contraste, les coefficients de régression linéaire ont des intervalles symétriques, car ils se situent sur une échelle additive : Cîrstea et ses collaborateurs rapportent B = 0,076 (IC à 95% 0,017-0,135 ; p = 0,011) pour l'association entre la satisfaction vis-à-vis des services pharmaceutiques et l'observance thérapeutique, chez 220 patients de dermatologie [72] ; l'estimation est exactement au milieu, 0,076 ± 0,059.

4.11 Le bootstrap expliqué simplement

Pour de nombreuses statistiques (la médiane, le rapport de deux médianes, le coefficient de corrélation de Spearman, l'AUC), il n'existe pas de formule simple pour l'erreur standard, ou bien les formules exigent des hypothèses que les données ne remplissent pas. La méthode du bootstrap, proposée dans les années 70, contourne le problème grâce à une idée presque naïve : si notre échantillon est représentatif de la population, nous pouvons simuler la "répétition de l'étude" en tirant de nouveaux échantillons à partir de l'échantillon lui-même.

La procédure est la suivante. Parmi les 160 valeurs de triglycérides, nous tirons au hasard 160 valeurs avec remise : certaines valeurs apparaîtront deux ou trois fois, d'autres pas du tout. Nous calculons la médiane de cet "échantillon bootstrap". Nous répétons l'opération 10 000 fois. Nous obtenons 10 000 médianes, dont la distribution approche la distribution d'échantillonnage de la médiane. Les percentiles 2,5 et 97,5 de cette distribution forment l'intervalle de confiance bootstrap à 95% (méthode des percentiles).

Pour la médiane des triglycérides de mg2026, le résultat est de 137-162 mg/dL (figure 4.4), très proche de l'intervalle obtenu par la méthode des rangs (134-162 mg/dL). La distribution des médianes bootstrap est irrégulière, avec des "sauts", car la médiane ne peut prendre que des valeurs du jeu de données ou des moyennes de deux valeurs voisines. Pour la moyenne de la PAS, le bootstrap donne 128,99-133,26 mmHg, presque identique à l'intervalle t classique (128,96-133,30 mmHg), et l'écart-type des 10 000 moyennes bootstrap est de 1,09 mmHg, pratiquement égal à l'erreur standard calculée par la formule (1,097 mmHg). Lorsque les hypothèses de la méthode classique sont remplies, les deux méthodes concordent ; lorsqu'elles ne le sont pas, le bootstrap reste valide.

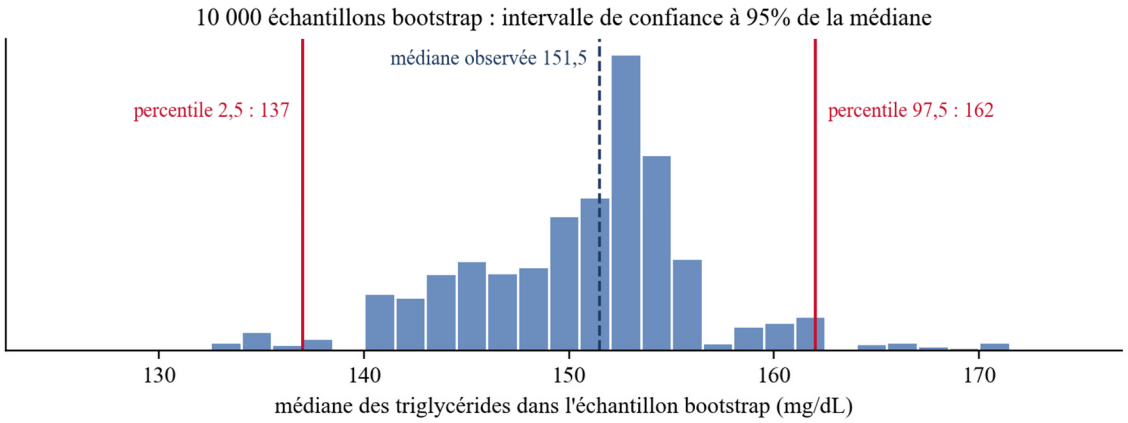


Figure 4.4. Distribution des médianes des triglycérides dans 10 000 échantillons bootstrap tirés du jeu de données mg2026. Les lignes rouges marquent les percentiles 2,5 (137 mg/dL) et 97,5 (162 mg/dL), qui forment l'intervalle de confiance bootstrap à 95% de la médiane ; la ligne discontinue est la médiane observée (151,5 mg/dL). Simulation sur le jeu de données mg2026.

Le bootstrap n'est pas une solution miracle : il ne peut pas corriger un échantillon non représentatif ou trop petit (avec 10 valeurs, les échantillons bootstrap répètent à l'infini les mêmes 10 nombres). Il est cependant de plus en plus utilisé dans les articles médicaux, par exemple pour les intervalles des AUC ou pour les modèles de médiation. Dicu et ses collaborateurs ont utilisé 5000 rééchantillonnages bootstrap pour valider un instrument d'évaluation des consommateurs, sur 504 participants [73]. JASP propose l'option "Bootstrap" dans de nombreuses analyses, et MedCalc l'utilise pour certains intervalles.

4.12 L'intervalle de confiance et la valeur p

Le chapitre 5 est consacré aux tests d'hypothèses, mais le lien avec les intervalles de confiance mérite d'être annoncé dès maintenant, car il simplifie beaucoup la lecture des articles. Pour une comparaison (différence de moyennes, différence de proportions, OR, RR), il existe une "valeur nulle" qui signifie "aucun effet" : 0 pour les différences, 1 pour les rapports. La règle est la suivante :

- si l'IC à 95% ne contient pas la valeur nulle, le test bilatéral correspondant donne $p < 0,05$;
- si l'IC à 95% contient la valeur nulle, $p \geq 0,05$;
- si la valeur nulle se trouve exactement à une limite de l'intervalle, $p \approx 0,05$.

Dans mg2026 : la différence de PAS hommes-femmes, IC à 95% de -4,9 à 3,8 mmHg, contient 0, et $p = 0,81$. Le RR de la perte de poids C par rapport à A, IC à 95% 1,11-6,17, ne contient pas 1, et $p < 0,05$. L'OR obésité-diabète, IC à 95% 1,86-10,03, ne contient pas 1, et $p < 0,001$; le fait que la limite inférieure soit loin de 1 correspond à une valeur p très petite. La règle fonctionne exactement pour les méthodes qui utilisent la même erreur standard pour l'intervalle et pour le test ; pour d'autres combinaisons (par exemple un intervalle de Wilson et un test exact), de petites discordances peuvent apparaître à la limite.

L'intervalle contient cependant plus d'information que la valeur p. La valeur p nous dit seulement si les données sont compatibles avec "aucun effet". L'intervalle nous dit quelles tailles d'effet sont compatibles avec les données : si l'intervalle est étroit et proche de zéro, l'effet est petit, même s'il est significatif ; s'il est large, le résultat est imprécis, quelle que soit la valeur de p [70,74].

> **Étude de cas : ce qui manque lorsqu'un rapport de cotes n'a pas d'intervalle de confiance**

> Cioca et ses collaborateurs ont comparé l'évolution de la COVID-19 chez 486 patients (200 sans diabète, 62 atteints de diabète de type 1 et 224 de diabète de type 2) et ont rapporté, pour le diabète de type 2, un OR de 2,24 pour l'admission en soins intensifs et un OR de 2,46 pour la ventilation mécanique [8]. Dans le résumé, ces valeurs ne sont accompagnées ni d'intervalles de confiance ni de valeurs p. Que ne pouvons-nous pas savoir sans eux ? Si l'OR de 2,24 provient d'un intervalle de 1,9-2,6 ou de 0,8-6,3 ; dans le premier cas, l'association serait certaine et estimée avec précision, dans le second, elle serait compatible aussi avec l'absence de toute association. Avec 62 patients dans le groupe atteint de diabète de type 1 et un nombre probablement faible d'admissions en soins intensifs, l'intervalle est très probablement large. Le lecteur critique ne rejette pas le résultat, mais il le traite comme une estimation de précision inconnue et cherche les intervalles dans le texte intégral. La situation inverse, un intervalle d'une étroitesse suspecte, apparaît dans l'étude d'Albai et de ses collaborateurs sur le score ISARIC-4C chez 159 patients atteints de diabète de type 2 et de COVID-19, avec une mortalité de 25,15% [75]. La sensibilité de 87,80% est rapportée avec un IC à 95% de 87,66-87,93%, soit une largeur inférieure à 0,3 point de pourcentage. Pour une proportion estimée sur environ 40 patients décédés (25,15% de 159), l'erreur standard serait d'environ $\sqrt{\frac{0{,}878\times0{,}122}{40}} = 0{,}052$, donc un intervalle de Wilson d'environ ±10 points de pourcentage (autour de 74-95%). Des intervalles aussi étroits proviennent probablement d'une erreur de calcul ou de présentation, et ils illustrent pourquoi il vaut la peine de vérifier, au moins approximativement, si la largeur d'un intervalle est cohérente avec le nombre de sujets.

4.13 Principes du calcul de la taille de l'échantillon

4.13.1 Pourquoi la calculer à l'avance

"De combien de patients ai-je besoin ?" est probablement la question qu'un statisticien entend le plus souvent de la part des cliniciens. La réponse ne peut être ni "le plus possible" ni "autant que nous aurons le temps d'en réunir avant la soutenance". Une étude trop petite produit des intervalles si larges qu'elle ne répond pas à la question, et elle expose les patients à des procédures sans bénéfice scientifique. Une étude inutilement grande consomme des ressources et expose parfois plus de patients que nécessaire à un traitement inférieur. Le calcul de la taille de l'échantillon se fait avant le recueil des données, il est inscrit dans le protocole et il est exigé par les comités d'éthique, les financeurs et les revues [18,53].

Toutes les formules ci-dessous suivent la même logique : nous exigeons une certaine précision (une marge d'erreur acceptable) ou la capacité de détecter une certaine différence, et nous partons d'une estimation de la variabilité, tirée d'études antérieures ou d'une étude pilote. Si la variabilité est sous-estimée, l'échantillon calculé sera trop petit.

4.13.2 Pour l'estimation d'une moyenne

Si nous voulons que la marge d'erreur de l'IC à 95% (la moitié de sa largeur) soit au plus égale à E, nous partons de $E = \frac{z\times\sigma}{\sqrt{n}}$ et nous la résolvons par rapport à n :

$$n = \left(\frac{z \times \sigma}{E}\right)^2$$

Exemple. Nous voulons estimer la moyenne de la PAS dans une population de patients à risque cardiométabolique avec une marge de ± 3 mmHg(IC 95%). D'après des études antérieures, l'écart-type est d'environ 14 mmHg.

$n = \left(\frac{1{,}96\times14}{3}\right)^2 = (9{,}147)^2 = 83{,}7$, arrondi à l'unité supérieure, soit 84 patients.

L'arrondi se fait toujours vers le haut ; sinon, nous n'obtenons plus la précision exigée. Remarquez la sensibilité de la formule : pour une marge de ±2 mmHg, $n = \left(\frac{1{,}96 \times 14}{2}\right)^2 = 188{,}2$, donc 189 patients ; pour ±5 mmHg, seulement 31. Diviser la marge par deux multiplie l'échantillon par quatre. Pour la glycémie à jeun, avec $\sigma \approx 15{,}76$ mg/dL et une marge de ±3 mg/dL, il faudrait 107 patients.

4.13.3 Pour l'estimation d'une proportion

La formule analogue pour une proportion est :

$$n = \frac{z^2 \times p_i \times (1 - p_i)}{E^2}$$

où p_i est la proportion anticipée et E la marge souhaitée (en fraction, et non en pourcentage).

Exemple. Un médecin veut estimer la prévalence du diabète non diagnostiqué chez les patients à risque de sa patientèle. Il anticipe une prévalence d'environ 20% et souhaite une marge de ±5 points de pourcentage.

$n = \frac{1{,}96^2 \times 0{,}20 \times 0{,}80}{0{,}05^2} = \frac{3{,}8416 \times 0{,}16}{0{,}0025} = 245{,}9$, donc 246 patients.

Si nous n'avons aucune idée de la prévalence, nous utilisons $p_i = 0{,}5$, qui donne la valeur maximale du produit p_i x (1 - p_i) et donc l'échantillon le plus grand, le plus "sûr" : $n = \frac{3{,}8416 \times 0{,}25}{0{,}0025} = 384{,}2$, donc 385. Pour $p_i = 0{,}18$ et une marge de ±3 points de pourcentage, n = 631. La formule repose sur l'approximation de Wald ; pour les proportions très petites (moins de 5%), on utilise des variantes plus exactes, disponibles dans les logiciels.

4.13.4 Pour la comparaison de deux moyennes : la puissance et la taille de l'effet

Lorsque le but est de comparer deux groupes (par exemple un nouveau traitement par rapport au traitement standard), le calcul exige quatre ingrédients :

- le seuil de signification α, généralement 0,05 en bilatéral (z = 1,96) ;
- la puissance souhaitée, 1 - β, généralement 80% (z = 0,84) ou 90% (z = 1,28) : la probabilité que l'étude détecte la différence, si celle-ci existe réellement ;
- la différence minimale cliniquement pertinente Δ, que nous ne voulons pas manquer ;
- l'écart-type σ de la variable.

Les deux derniers se combinent dans la taille d'effet standardisée, le d de Cohen = Δ / σ [76]. Par convention, d = 0,2 est un petit effet, 0,5 un effet moyen, 0,8 un grand effet. La formule approchée du nombre de patients par groupe est :

n par groupe = 2 x (z(1-α/2) + z(1-β))² / d²

Pour d = 0,5, α = 0,05 et une puissance de 80% : $n = \frac{2 \times (1{,}96 + 0{,}84)^2}{0{,}25} = \frac{2 \times 7{,}85}{0{,}25} = 62{,}8$. Les logiciels, qui utilisent la loi t au lieu de l'approximation normale, donnent 64 patients par groupe. Pour d = 0,3, le même calcul donne 174,4, et les logiciels 176 par groupe ; pour d = 0,8, 26 par groupe ; pour d = 0,2, 394 par groupe. Les petits effets exigent de grands échantillons : pour détecter un effet deux fois plus petit, il nous faut quatre fois plus de patients. Si nous exigeons une puissance de 90% au lieu de 80%, pour d = 0,5, le nombre passe de 64 à 86 par groupe.

Un exemple clinique. Nous voulons savoir si une intervention abaisse la PAS d'au moins 5 mmHg par rapport au groupe témoin, avec σ = 14 mmHg. $d = \frac{5}{14} = 0{,}36$, et les logiciels indiquent 125 patients par groupe pour une puissance de 80%. Revenons maintenant à la comparaison de la PAS entre les sexes dans mg2026, avec 74 hommes et 86 femmes : la

puissance de cet échantillon pour détecter une différence réelle de 5 mmHg n'est que d'environ 61%. Le fait que la différence observée n'ait pas été significative (p = 0,81) n'exclut donc pas une différence de 5 mmHg ; l'intervalle de confiance (-4,9 à 3,8 mmHg) dit la même chose, plus directement.

Pour deux proportions, la formule est plus longue, mais la logique reste la même. Pour détecter une augmentation de la proportion de patients ayant une perte de poids ≥ 5% de 11,3% (bras A) à 29,6% (bras C), avec α = 0,05 et une puissance de 80%, il faudrait environ 76 patients par bras. Les bras réels de mg2026 comptent chacun 53 à 54 patients, et leur puissance pour cette différence est d'environ 67%. L'étude a néanmoins trouvé une différence significative (IC à 95% de la différence 3,4-33,2 points de pourcentage), mais avec un intervalle large, exactement ce que l'on attend d'une étude de puissance modérée.

4.13.5 Ajustements pratiques : perdus de vue et petites populations

Les perdus de vue (patients qui abandonnent, qui ne peuvent plus être contactés, qui ont des données manquantes) réduisent l'échantillon analysé. Si nous anticipons un taux de perte r, nous recrutons n / (1 - r) patients. Pour 64 patients par groupe et des pertes estimées à 20% : 64 / 0,8 = 80 par groupe, soit 160 au total. L'erreur fréquente consiste à multiplier par 1,2 (64 x 1,2 = 76,8, donc 77), ce qui sous-estime le besoin : sur 77 patients recrutés, avec 20% de pertes, il n'en reste que 61,6.

La correction pour population finie s'applique lorsque l'échantillon représente une fraction importante (plus de 5 à 10%) d'une population petite et connue, par exemple tous les patients à risque de la patientèle d'un médecin généraliste :

$$n \text{ corrigé } = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$$

Pour n0 = 246 et une patientèle de N = 1000 patients à risque, $n = \frac{246}{1+\frac{245}{1000}} = \frac{246}{1,245} = 197,6$, donc 198. Pour N = 500, 166 patients. Pour les grandes populations (des dizaines de milliers), la correction est négligeable.

4.13.6 Logiciels de calcul de la taille de l'échantillon

En pratique, personne ne fait ces calculs à la main, mais il est important de comprendre ce que vous saisissez dans le logiciel. G*Power est un logiciel gratuit, largement utilisé, qui couvre les tests t, l'ANOVA, le khi-deux, la corrélation et la régression [77]. Pour deux moyennes, on choisit la famille "t tests", le test "Means: Difference between two independent means (two groups)", le type d'analyse "A priori", et on saisit la taille d'effet d, α, la puissance et le rapport entre les groupes ; pour d = 0,5, le résultat est de 64 par groupe. MedCalc dispose d'un menu dédié (Sampling) pour une moyenne, une proportion, la comparaison de deux moyennes ou de deux proportions, les courbes ROC et d'autres cas [50]. Epi Info propose, dans le module StatCalc, les calculateurs "Population survey" (pour les prévalences, avec la correction pour population finie incluse), "Cohort or cross-sectional" et "Unmatched case-control" [51]. Ces calculateurs sont utilisés lors des travaux pratiques 4.

4.13.7 Ce qu'il faut écrire dans le protocole

Un paragraphe correct sur la taille de l'échantillon permet à quiconque de refaire exactement le calcul. Il doit contenir : la variable principale sur laquelle le calcul a été fondé ; la différence minimale cliniquement pertinente ou la marge de précision souhaitée, avec sa justification ; l'estimation de la variabilité (écart-type ou proportion anticipée) et sa source ; α et le caractère bilatéral ou non du test ; la puissance ; le taux de perdus de vue anticipé ; le nombre final de

patients à recruter ; le logiciel utilisé. Par exemple : "Pour détecter une différence de PAS de 5 mmHg entre les groupes (écart-type estimé à 14 mmHg, d'après l'étude X), avec un test t bilatéral, α = 0,05 et une puissance de 80%, 125 patients par groupe sont nécessaires (calcul dans G*Power). En anticipant 15% de perdus de vue, nous recruterons 148 patients par groupe."

Le calcul a posteriori de la puissance (la puissance "post-hoc", calculée à partir de l'effet observé) a une valeur limitée : c'est une fonction directe de la valeur p et elle n'apporte aucune information nouvelle. Pour une étude déjà terminée, l'intervalle de confiance indique plus clairement quels effets sont compatibles ou non avec les données.

4.13.8 Les conséquences des échantillons trop petits

Une étude de faible puissance pose trois problèmes. Le premier, évident : elle peut manquer un effet réel (erreur de type II, chapitre 5). Le deuxième, moins évident : si elle trouve malgré tout un effet significatif, l'estimation tend à être exagérée, car seules les estimations élevées, dues au hasard, franchissent le seuil de signification. Le troisième : les intervalles de confiance sont si larges que le résultat, significatif ou non, est difficilement utilisable en clinique.

> **Étude de cas : une étude avec une puissance de 29,3%**
>
> Popovici et ses collaborateurs ont comparé, dans une cohorte prospective de 100 patients ayant bénéficié d'une CPRE, la sédation guidée par la surveillance de l'entropie et la sédation conventionnelle [78]. Les épisodes hypertensifs ont été plus rares dans le groupe guidé par l'entropie (6% contre 26%, p = 0,007), et les scores de récupération d'Aldrete ont été plus élevés à 5 minutes (p = 0,003) et à 15 minutes (p < 0,001). Les auteurs signalent cependant honnêtement une puissance statistique limitée, de 29,3%. Une puissance de 29,3% signifie que, si la différence recherchée existait réellement, une étude de cette taille n'aurait qu'environ 3 chances sur 10 de la détecter. Un résultat non significatif obtenu dans ces conditions ne peut pas être lu comme "il n'y a pas de différence" ; il est tout simplement non concluant. Le schéma d'étude observationnel, non randomisé, ajoute une seconde limite : les groupes peuvent différer par d'autres caractéristiques que la méthode de surveillance. La conclusion correcte, que formule d'ailleurs le titre ("is associated with"), est que les résultats génèrent des hypothèses qui doivent être confirmées dans un essai randomisé, avec un échantillon calculé à l'avance pour le critère principal.

4.14 Problèmes résolus

Problème résolu 4.1. IC à 95% et IC à 99% de la moyenne de la glycémie

Énoncé. Dans mg2026, la glycémie à jeun initiale a n = 160, $m = 105{,}44$ mg/dL et $s = 15{,}76$ mg/dL. Calculez à la main l'IC à 95% et l'IC à 99% de la moyenne de la population, et interprétez-les.

Solution. La distribution de la glycémie présente une légère asymétrie (0,37), mais pour n = 160 la moyenne est distribuée de façon presque normale (théorème central limite) ; l'intervalle t est donc approprié.

$\text{ES} = \frac{15{,}76}{\sqrt{160}} = \frac{15{,}76}{12{,}649} = 1{,}246$ mg/dL.

IC à 95% : t(159 ; 0,975) = 1,975 ; marge = 1,975 x 1,246 = 2,461 ; intervalle 105,44 ± 2,46, c'est-à-dire 102,98-107,90 mg/dL.

IC à 99% : t(159 ; 0,995) = 2,607 ; marge = 2,607 x 1,246 = 3,248 ; intervalle 105,44 ± 3,25, c'est-à-dire 102,19-108,69 mg/dL.

Vérification dans Google Sheets : =CONFIDENCE.T(0,05; STDEV(S2:S161); 160) donne 2,461, et =CONFIDENCE.T(0,01; STDEV(S2:S161); 160) donne 3,248.

Interprétation statistique. Avec une confiance de 95%, la moyenne de la glycémie à jeun dans la population d'adultes à risque dont provient l'échantillon se situe entre 102,98 et 107,90 mg/dL ; l'intervalle à 99% est plus large, car nous exigeons une plus grande certitude. Interprétation clinique. L'intervalle entier se situe dans la zone de l'hyperglycémie modérée à jeun (100-125 mg/dL), ce qui confirme que la population étudiée présente, en moyenne, un métabolisme glucidique altéré.

Problème résolu 4.2. Valeurs ou moyenne ? L'âge des patients

Énoncé. L'âge des patients de mg2026 a m = 56,29 ans, s = 8,77 ans, n = 160, avec une distribution normale (Shapiro-Wilk p = 0,687). (a) Estimez, avec une confiance de 95,4%, l'intervalle dans lequel se situent les âges des patients. (b) Estimez, avec une confiance de 95,4%, la moyenne de l'âge dans la population. (c) Commentez la différence.

Solution. (a) Pour les valeurs individuelles, nous utilisons l'écart-type : $m \pm 2s = 56{,}29 \pm 17{,}54$, c'est-à-dire 38,75-73,83 ans. Dans l'échantillon, 94,4% des patients ont un âge compris dans cet intervalle.

(b) Pour la moyenne, nous utilisons l'erreur standard : ES $= \frac{8{,}77}{\sqrt{160}} = 0{,}693$ ans ; $m \pm 2\text{ES} = 56{,}29 \pm 1{,}39$, c'est-à-dire 54,90-57,68 ans.

(c) Le premier intervalle a une largeur de 35 ans et décrit la diversité des patients : des jeunes adultes aux personnes âgées. Le second fait moins de 3 ans et décrit la précision avec laquelle nous connaissons l'âge moyen de la population. Le rapport entre les largeurs est $\sqrt{160} \approx 12{,}6$. À l'examen, les mots-clés de l'énoncé sont "les valeurs" (ÉT) par opposition à "la moyenne" (ES).

Problème résolu 4.3. Un petit échantillon

Énoncé. Chez 8 patients atteints de diabète de type 2, l'HbA1c après trois mois de traitement a les valeurs suivantes (données fictives) : 6,1 ; 5,8 ; 6,4 ; 7,0 ; 6,2 ; 5,9 ; 6,6 ; 6,0%. Calculez l'IC à 95% de la moyenne de l'HbA1c.

Solution. Avec seulement 8 valeurs, nous ne pouvons pas vérifier sérieusement la normalité ; l'HbA1c a généralement une distribution approximativement symétrique dans des groupes homogènes, et les données ne montrent pas de valeur extrême évidente ; nous utilisons donc l'intervalle t, avec prudence.

La moyenne : somme = 50,0, donc $m = \frac{50{,}0}{8} = 6{,}25\%$.

Les écarts à la moyenne : -0,15 ; -0,45 ; 0,15 ; 0,75 ; -0,05 ; -0,35 ; 0,35 ; -0,25. Les carrés : 0,0225 ; 0,2025 ; 0,0225 ; 0,5625 ; 0,0025 ; 0,1225 ; 0,1225 ; 0,0625 ; somme = 1,12.

$s^2 = \frac{1{,}12}{7} = 0{,}16$; s = 0,40%.

ES $= \frac{0{,}40}{\sqrt{8}} = \frac{0{,}40}{2{,}828} = 0{,}141\%$.

ddl = 7 ; t(7 ; 0,975) = 2,365 ; marge = 2,365 x 0,141 = 0,334.

IC à 95% : 6,25 ± 0,33, c'est-à-dire 5,92-6,58%.

Interprétation. La moyenne de la population des patients traités de façon similaire se situe, avec une confiance de 95%, entre 5,92% et 6,58%. Si nous avions utilisé à tort z = 1,96, la marge aurait été de 0,28 et l'intervalle de 5,97-6,53%, environ 17% plus étroit que l'intervalle correct. Avec 8 patients, la différence entre t et z n'est pas une subtilité.

Problème résolu 4.4. IC d'une proportion : Wald, Wilson et la règle de trois

Énoncé. (a) Lors d'un audit, 12 patients hypertendus sur 80 ont une PA non contrôlée sous traitement. Calculez l'IC à 95% de la proportion par la méthode de Wald et par la méthode de Wilson. (b) Dans un autre cabinet, chez 40 patients traités par un nouvel antihypertenseur, aucun cas d'angio-œdème n'a été observé. Que pouvez-vous dire du risque réel ?

Solution. $(a)\, p = \frac{12}{80} = 0{,}15$. ES $= \sqrt{\frac{0{,}15 \times 0{,}85}{80}} = \sqrt{0{,}001594} = 0{,}0399$.

Wald : $0{,}15 \pm 1{,}96 \times 0{,}0399 = 0{,}15 \pm 0{,}078$, c'est-à-dire 7,2%-22,8%.

Wilson : $\frac{z^2}{n} = \frac{3{,}8416}{80} = 0{,}0480$; $\frac{z^2}{2n} = 0{,}0240$. Centre = (0,15 + 0,024) / 1,048 = 0,1660. Sous la racine : $0{,}001594 + \frac{3{,}8416}{4 \times 6400} = 0{,}001594 + 0{,}000150 = 0{,}001744$; racine = 0,0418. Marge = $\left(\frac{1{,}96}{1{,}048}\right) \times 0{,}0418 = 1{,}870 \times 0{,}0418 = 0{,}0781$. IC à 95% de Wilson : 0,166 ± 0,078, c'est-à-dire 8,8%-24,4%.

L'intervalle de Wilson est déplacé vers 50% et asymétrique par rapport à 15% (6,2 points en dessous, 9,4 au-dessus), comme il se doit pour une petite proportion dans un échantillon de taille modérée.

(b) Zéro événement sur 40 : la règle de trois donne une limite supérieure de l'IC à 95% d'environ 3 / 40 = 7,5%. Les données sont compatibles avec un risque réel d'angio-œdème allant jusqu'à environ 1 cas pour 13 patients. L'absence de cas chez 40 patients ne prouve pas l'innocuité du médicament vis-à-vis d'une réaction rare.

Problème résolu 4.5. La différence de deux proportions et le rapport de prévalences

Énoncé. Dans mg2026, à l'inclusion, le diabète était présent chez 18 des 54 patients obèses (33,3%) et chez 11 des 106 patients non obèses (10,4%). Calculez l'IC à 95% de la différence des prévalences et du rapport de prévalences.

Solution. La différence : 0,333 - 0,104 = 0,230 (23,0 points de pourcentage).

ES $= \sqrt{\frac{0{,}333 \times 0{,}667}{54} + \frac{0{,}104 \times 0{,}896}{106}} = \sqrt{0{,}004115 + 0{,}000878} = \sqrt{0{,}004993} = 0{,}0707$.

IC à 95% : $0{,}230 \pm 1{,}96 \times 0{,}0707 = 0{,}230 \pm 0{,}139$, c'est-à-dire 9,1-36,8 points de pourcentage.

Le rapport de prévalences : RP $= \frac{0{,}333}{0{,}104} = 3{,}21$. Sur l'échelle logarithmique : ln(3,21) = 1,167 ; ES$(\ln \text{RP}) = \sqrt{\frac{1}{18} - \frac{1}{54} + \frac{1}{11} - \frac{1}{106}} = \sqrt{0{,}0556 - 0{,}0185 + 0{,}0909 - 0{,}0094} = \sqrt{0{,}1185} = 0{,}344$. $\ln(\text{RP}) \pm 1{,}96 \times 0{,}344 = 1{,}167 \pm 0{,}675$, c'est-à-dire de 0,492 à 1,842 ; après retransformation : $e^{0{,}492} = 1{,}64$ et $e^{1{,}842} = 6{,}31$.

IC à 95% du RP : 1,64-6,31.

Interprétation. La prévalence du diabète est environ 3 fois plus élevée chez les obèses, avec un intervalle compatible avec un rapport allant de 1,6 à 6,3 ; en termes absolus, l'excès est de 23 points de pourcentage (9-37). Les deux intervalles excluent la valeur nulle (respectivement 0 et 1) ; l'association est donc statistiquement significative. S'agissant d'une analyse transversale, elle n'établit pas la causalité (chapitre 10).

Problème résolu 4.6. La taille de l'échantillon pour une étude en cabinet

Énoncé. (a) Un médecin généraliste veut estimer la prévalence de l'hypertension non contrôlée parmi les 1500 patients hypertendus de sa patientèle, en anticipant 30%, avec une marge de ±5

points de pourcentage. Il estime que 10% des patients invités ne répondront pas. (b) Un second médecin veut comparer la baisse de la glycémie entre une nouvelle intervention et le conseil standard : la différence minimale pertinente est de 8 mg/dL, et l'écart-type de la baisse, d'après mg2026, est d'environ 11 mg/dL ; α = 0,05, puissance de 80%, pertes estimées à 15%.

Solution. (a) Étape 1 : $n_0 = \frac{1{,}96^2 \times 0{,}30 \times 0{,}70}{0{,}05^2} = \frac{3{,}8416 \times 0{,}21}{0{,}0025} = 322{,}7$, donc 323. Étape 2, la correction pour population finie : $n = \frac{323}{1+\frac{322}{1500}} = \frac{323}{1{,}215} = 265{,}8$, donc 266. Étape 3, la non-réponse : 266 / 0,90 = 295,6, donc 296 patients invités.

(b) $d = \frac{8}{11} = 0{,}73$. n par groupe = $\frac{2 \times (1{,}96+0{,}84)^2}{0{,}73^2} = \frac{15{,}7}{0{,}529} = 29{,}7$, donc 30 (G*Power, avec la loi t, donne 31). Avec 15% de pertes : 31 / 0,85 = 36,5, donc 37 patients par groupe, 74 au total.

Interprétation. La correction pour population finie a réduit l'échantillon de près de 60 patients, car la patientèle du médecin est petite. Pour la seconde étude, l'effet anticipé est grand (d = 0,73) ; l'échantillon est donc modeste. Si la différence pertinente n'était que de 4 mg/dL($d = 0{,}36$), il faudrait environ 120 patients par groupe, quatre fois plus.

4.15 Exercices proposés

1. Le cholestérol total de mg2026 a n = 160, $m = 209{,}39$ mg/dL, $s = 36{,}91$ mg/dL. Calculez l'erreur standard et l'IC à 95% de la moyenne de la population (t = 1,975).
2. Le HDL de mg2026 a $m = 49{,}52$ mg/dL, $s = 11{,}78$ mg/dL, n = 160. Calculez les intervalles à 68,3%, 95,4% et 99,7% pour les valeurs individuelles et pour la moyenne, et présentez-les dans un tableau comme le tableau 4.1.
3. Dans un échantillon de 25 patients diabétiques, l'HbA1c a une moyenne de 7,2% et un écart-type de 1,5%. Calculez l'IC à 95% de la moyenne. Quelle valeur critique avez-vous utilisée et pourquoi ?
4. Dans mg2026, 41 patients sur 160 ont une PAS ≥ 140 mmHg. Calculez l'IC à 95% de la proportion par la méthode de Wald et par la méthode de Wilson.
5. Un nouveau vaccin a été administré à 120 volontaires, sans aucune réaction anaphylactique. Quelle est la limite supérieure approximative de l'IC à 95% du risque d'anaphylaxie ?
6. Deux groupes indépendants : n1 = 40, m1 = 142 mmHg, s1 = 15 mmHg ; n2 = 45, m2 = 134 mmHg, s2 = 14 mmHg. Calculez la différence des moyennes, son erreur standard et l'IC à 95% approximatif (avec z = 1,96). La différence est-elle significative ?
7. Dans une étude cas-témoins, 20 des 50 cas et 10 des 50 témoins ont été exposés. Calculez l'OR et son IC à 95% sur l'échelle logarithmique.
8. Combien de patients sont nécessaires pour estimer : (a) une moyenne avec σ = 10 et une marge de ±2 unités (IC à 95%) ; (b) une proportion totalement inconnue avec une marge de ±4 points de pourcentage ?
9. Pour comparer deux moyennes avec d = 0,4, α = 0,05 en bilatéral et une puissance de 80%, calculez le nombre de patients par groupe avec la formule approchée. Combien faudrait-il en recruter si l'on anticipe 20% de perdus de vue ?
10. Un article rapporte "PAS $131 \pm 1{,}1$ mmHg$(n = 160)$". Que représente, le plus probablement, la valeur 1,1 ? Calculez l'écart-type correspondant. Un autre article rapporte OR = 1,75 (IC à 95% 0,95-3,10) : l'association est-elle significative au seuil de 0,05 ?

11. Un IC à 95% de la moyenne, calculé sur 40 patients, a une largeur de 10 mmHg. Combien de patients faudrait-il approximativement pour obtenir une largeur de 5 mmHg, en supposant le même écart-type ?

Pièges fréquents

- Confondre l'écart-type et l'erreur standard : l'ÉT décrit les patients, l'ES décrit la précision de la moyenne. La présentation "moyenne ± ES" fait paraître les données moins variables.
- Utiliser l'intervalle pour la moyenne lorsque la question porte sur les valeurs individuelles (et inversement), surtout dans les problèmes du type "estimez les valeurs/la moyenne avec une confiance de 95,4%".
- Interpréter l'IC à 95% comme "95% des patients se situent dans l'intervalle" ou comme "la probabilité de 95% que la moyenne se trouve dans cet intervalle précis".
- Utiliser z = 1,96 au lieu de la valeur t pour les petits échantillons.
- Rapporter l'intervalle de Wald pour de petites proportions ou de petits échantillons, avec des limites négatives ou une couverture insuffisante.
- Conclure à "l'absence d'effet" sur la base d'un intervalle large qui inclut la valeur nulle ; la formulation correcte est "les données ne permettent pas de conclure".
- Calculer l'intervalle de l'OR ou du RR directement sur l'échelle d'origine, de façon symétrique, au lieu de l'échelle logarithmique.
- Arrondir la taille de l'échantillon vers le bas ou ajuster pour les perdus de vue en multipliant par (1 + r) au lieu de diviser par (1 - r).
- Calculer la puissance "post-hoc" à partir de l'effet observé pour "expliquer" un résultat non significatif.
- Oublier que l'intervalle de confiance ne mesure que l'erreur aléatoire ; un échantillon biaisé donne des intervalles étroits autour d'une valeur fausse.

À retenir

- L'erreur standard de la moyenne est ES = $\frac{s}{\sqrt{n}}$; pour la PAS de mg2026, elle vaut 1,097 mmHg, contre un ÉT de 13,88 mmHg.
- L'intervalle pour les valeurs individuelles est $m \pm k \times s$, l'intervalle pour la moyenne est m ± k x ES (k = 1, 2, 3 pour 68,3%, 95,4%, 99,7%).
- IC de la moyenne : $m \pm \frac{t \times s}{\sqrt{n}}$, avec ddl = n - 1 ; pour la PAS, IC à 95% 128,96-133,30 mmHg et IC à 99% 128,27-133,99 mmHg.
- Un IC à 95% signifie que la procédure de calcul capture le vrai paramètre dans 95% des études répétées ; sa largeur diminue avec $\sqrt{n}$.
- Pour les proportions, on préfère l'intervalle de Wilson (diabète dans mg2026 : 12,9-24,8%, contre 12,2-24,1% avec Wald) ; pour zéro événement sur n, la limite supérieure est d'environ 3 / n.
- Les intervalles des OR, des RR et de la moyenne géométrique se calculent sur l'échelle logarithmique et sont asymétriques ; l'estimation ponctuelle est la moyenne géométrique des limites.
- Si l'IC à 95% d'une différence exclut 0 (ou celui d'un rapport exclut 1), p < 0,05 ; l'intervalle montre en outre la taille et la précision de l'effet.
- Taille de l'échantillon : $n = \left(\frac{z \times \sigma}{E}\right)^2$ pour une moyenne (84 pour ± 3 mmHg, σ 14), $n = \frac{z^2 \times p_i \times (1-p_i)}{E^2}$ pour une proportion (246 pour 20% ± 5%), environ 2 x (1,96 + 0,84)² / d² par groupe pour deux moyennes (64 pour d = 0,5 ; 176 pour d = 0,3).

Questions d'autoévaluation

1. (QRU) Dans mg2026, la PAS a s = 13,88 mmHg et n = 160. L'erreur standard de la moyenne est d'environ :

a) 0,087 mmHg
b) 1,097 mmHg
c) 3,72 mmHg
d) 13,88 mmHg
e) 27,2 mmHg

2. (QRU) Pour diviser par deux la largeur d'un intervalle de confiance de la moyenne, à variabilité et niveau de confiance égaux, l'échantillon doit être :
   a) doublé
   b) triplé
   c) multiplié par quatre
   d) multiplié par huit
   e) réduit de moitié
3. (QRU) Une étude rapporte OR = 1,40 (IC à 95% 0,90-2,18). L'affirmation correcte est :
   a) l'association est statistiquement significative au seuil de 0,05
   b) l'intervalle inclut la valeur 1, donc $p \geq 0,05$
   c) l'intervalle prouve l'absence de toute association
   d) l'intervalle devrait être symétrique autour de 1,40
   e) la valeur p est certainement inférieure à 0,01
4. (QRU) Dans une étude portant sur 60 patients, aucun événement indésirable grave n'a été observé. La limite supérieure approximative de l'IC à 95% du risque réel est :
   a) 0%
   b) 1,7%
   c) 3%
   d) 5%
   e) 10%
5. (QRU) Combien de patients sont nécessaires pour estimer une moyenne avec une marge de ± 3 mmHg(IC 95%), si σ = 14 mmHg ?
   a) 19
   b) 47
   c) 84
   d) 176
   e) 246
6. (QRM) À propos de l'interprétation d'un IC à 95% de la moyenne de la PAS de 128,96-133,30 mmHg, les affirmations correctes sont :
   a) si l'étude était répétée de nombreuses fois, environ 95% des intervalles ainsi calculés contiendraient la vraie moyenne
   b) 95% des patients ont une PAS comprise entre 128,96 et 133,30 mmHg
   c) l'intervalle exprime la précision de l'estimation de la moyenne
   d) un IC à 99% calculé sur les mêmes données serait plus large
   e) l'intervalle deviendrait plus étroit si l'échantillon était plus grand
7. (QRM) À propos de l'écart-type (ÉT) et de l'erreur standard (ES), les affirmations correctes sont :
   a) l'ÉT décrit la variabilité des valeurs individuelles
   b) l'ES diminue quand la taille de l'échantillon augmente

c) $\mathrm{ES} = \frac{\text{ÉT}}{\sqrt{n}}$

d) l'intervalle de référence pour les valeurs individuelles se construit avec l'ES

e) rapporter "moyenne ± ES" au lieu de "moyenne ± ÉT" fait paraître les données moins variables

8. (QRM) À propos de l'intervalle de confiance d'une proportion, les affirmations correctes sont :

a) la méthode de Wald peut donner des limites inférieures à 0% pour de petites proportions

b) la méthode de Wilson donne des intervalles asymétriques pour des proportions proches de 0

c) la méthode de Wald a une couverture insuffisante dans les petits échantillons

d) l'intervalle de Wilson pour le diabète dans mg2026 est de 12,9-24,8%

e) l'intervalle de Wald est toujours plus large que celui de Wilson

9. (QRM) Lesquels des facteurs suivants augmentent la taille d'échantillon nécessaire pour comparer deux moyennes ?

a) une différence minimale pertinente plus petite

b) un écart-type plus grand

c) une puissance souhaitée plus élevée (90% au lieu de 80%)

d) un seuil α plus élevé (0,10 au lieu de 0,05)

e) un taux anticipé de perdus de vue plus élevé

10. (QRM) À propos des intervalles de confiance des rapports (OR, RR), les affirmations correctes sont :

a) ils se calculent sur l'échelle logarithmique, puis se retransforment

b) ils sont généralement asymétriques autour de l'estimation ponctuelle

c) l'estimation ponctuelle est approximativement la moyenne géométrique des limites

d) un intervalle qui exclut 1 correspond à un résultat significatif au seuil de 5%

e) un intervalle très large indique un effet très important

Chapitre 5. Les bases des tests statistiques : hypothèses, seuil de signification, erreurs de type I et II, décision et interprétation

Objectifs

À la fin de ce chapitre, l'étudiant sera capable de :

- transformer une question clinique en une hypothèse de recherche, puis en un couple d'hypothèses statistiques H0 et H1, écrites en mots et en notation mathématique ;
- choisir entre un test bilatéral et un test unilatéral et justifier ce choix avant de voir les données ;
- définir correctement la valeur p et reconnaître ses interprétations erronées les plus fréquentes ;
- appliquer la règle de décision pour un seuil α donné et formuler la conclusion statistique sans l'expression "on accepte H0" ;
- décrire les erreurs de type I et II, expliquer le lien entre α, β, la puissance 1 - β et le niveau de confiance 1 - α, et énumérer les facteurs qui influencent la puissance ;
- calculer et interpréter une taille d'effet (d de Cohen, dz, r, eta², OR) et distinguer la significativité statistique de l'importance clinique ;
- résoudre à la main un test z et un test t pour une seule moyenne, par trois voies équivalentes : la valeur critique, la valeur p et l'intervalle de confiance ;
- expliquer l'inflation de l'erreur de type I lors des comparaisons multiples et appliquer les corrections de Bonferroni, de Holm et de Benjamini-Hochberg ;
- rapporter un test statistique selon les normes actuelles (valeur p exacte, taille de l'effet, intervalle de confiance) et reconnaître les pratiques de type p-hacking et HARKing.

5.1 De la question clinique à l'hypothèse testable

Les chapitres précédents ont décrit les données (statistique descriptive), ont montré comment se comporte une variable normale et comment se construit un intervalle de confiance pour une moyenne ou une proportion. Ici commence la seconde partie du cours, la statistique inférentielle au sens strict : nous prenons des décisions sur une population à partir d'un seul échantillon. Tous les tests des chapitres 6 à 9 (test t, ANOVA, Mann-Whitney, khi-deux, test de la corrélation) utilisent la même logique, que nous construisons maintenant une fois pour toutes, avec soin. Si cette logique est comprise, chaque nouveau test ne représente qu'une nouvelle formule pour la "statistique de test" ; le reste du raisonnement demeure inchangé [19,63].

5.1.1 La question de recherche et l'hypothèse de recherche

Toute étude part d'une question clinique. Le médecin généraliste qui a suivi pendant six mois les patients de l'étude mg2026 se demande si l'application mobile, ajoutée au conseil, abaisse la glycémie à jeun davantage que le conseil seul. La question est bonne, mais encore vague : que signifie "davantage", chez qui, après combien de temps, mesuré comment ? Le premier travail du chercheur est de préciser la population, l'intervention, le comparateur et le critère de jugement (le schéma PICO, présenté dans le volume d'informatique médicale [1]). Il en résulte une hypothèse de recherche, une affirmation sur le monde réel que l'étude tente d'étayer : "chez les adultes à risque cardiométabolique suivis en cabinet de médecine générale, l'ajout de l'application d'automesure au conseil produit, en six mois, une baisse de la glycémie à jeun plus importante que le conseil seul".

L'hypothèse de recherche est formulée dans le langage du clinicien et exprime généralement ce que le chercheur s'attend à trouver : une différence, une association, un effet. Elle ne peut pas être testée directement, car nous n'avons pas accès à tous les patients à risque cardiométabolique du pays. Nous avons 160 patients. Pour passer de ces 160 patients à la population, l'hypothèse de recherche doit être traduite sous une forme qu'un calcul peut vérifier : un couple d'hypothèses statistiques portant sur les paramètres de la population (moyennes μ, proportions π, coefficients de corrélation ρ).

5.1.2 Les hypothèses statistiques : H0 et H1

L'hypothèse nulle (H0, hypothèse de différence nulle) affirme que, dans la population, la différence, l'association ou l'effet recherché n'existe pas. La différence observée dans l'échantillon ne serait alors que le résultat de la variabilité d'échantillonnage, c'est-à-dire du hasard qui décide quels patients entrent dans l'étude. L'hypothèse alternative (H1, parfois notée Ha) nie l'hypothèse nulle : dans la population, la différence existe. Les deux hypothèses s'excluent mutuellement et couvrent ensemble toutes les possibilités.

Pourquoi partir de H0, c'est-à-dire de l'hypothèse que, le plus souvent, nous espérons réfuter ? Parce que H0 est la seule qui fixe une valeur précise du paramètre (par exemple, une différence entre les moyennes égale à zéro). Ce n'est qu'en partant d'une valeur précise que nous pouvons calculer à quel point nos données seraient probables. L'hypothèse alternative "la différence n'est pas nulle" englobe une infinité de valeurs possibles et ne permet pas un calcul unique. Le raisonnement ressemble à une démonstration par l'absurde : nous supposons que H0 est vraie, nous calculons ce que nous devrions observer dans ce cas et, si les données réelles sont très difficiles à concilier avec cette supposition, nous y renonçons.

Pour l'exemple de l'application, les hypothèses s'écrivent ainsi :

H0 : la baisse moyenne de la glycémie à 6 mois est la même dans la population traitée par conseil + application et dans la population traitée par conseil seul ; mathématiquement, μB = μA.

H1 : la baisse moyenne de la glycémie diffère entre les deux populations ; mathématiquement, μB ≠ μA.

Remarquez que les hypothèses portent sur la population (lettres grecques μ, π, ρ), et non sur l'échantillon. Les moyennes de l'échantillon, nous les connaissons déjà (3,0 mg/dL dans le bras A et 11,1 mg/dL dans le bras B) ; il n'y a aucun sens à formuler des hypothèses à leur sujet. À l'examen, écrire les hypothèses avec m1 et m2 (les moyennes des échantillons) au lieu de μ1 et μ2 est une erreur fréquente, qui fait perdre des points.

Tableau 5.1. Exemples de traduction de questions cliniques en hypothèses statistiques (jeu de données mg2026 et exemples cliniques).

Question clinique	H0 (en mots)	H0 (mathématiquement)	H1 (mathématiquement)
La moyenne de la PAS des patients diffère-t-elle de 125 mmHg ?	la moyenne de la population est de 125 mmHg	$\mu = 125$	μ ≠ 125
La PAS diffère-t-elle entre les femmes et les hommes ?	les moyennes des deux sexes sont égales	$\mu_F = \mu_M$	μF ≠ μM
La glycémie change-t-elle entre l'inclusion et 6 mois (mêmes patients) ?	la moyenne des différences individuelles est nulle	$\mu_D = 0$	μD ≠ 0
La baisse de la glycémie diffère-t-elle entre les trois bras ?	les trois moyennes sont égales	$\mu_A = \mu_B = \mu_C$	au moins une moyenne diffère
La proportion de patients	les proportions sont égales	$\pi_C = \pi_A$	πC ≠ πA

Question clinique	H0 (en mots)	H0 (mathématiquement)	H1 (mathématiquement)
ayant perdu ≥5% de leur poids diffère-t-elle entre les bras C et A ?			
L'IMC et la PAS sont-ils corrélés ?	le coefficient de corrélation dans la population est nul	$\rho = 0$	$\rho \neq 0$

5.1.3 Tests bilatéraux et unilatéraux

L'hypothèse alternative des exemples ci-dessus est bilatérale (en anglais two-sided) : elle admet une différence dans un sens comme dans l'autre, μB > μA ou μB < μA. Un test unilatéral (en anglais one-sided) ne considère qu'un seul sens : H1 : μB > μA. Pour un même jeu de données, la valeur p unilatérale est la moitié de la valeur p bilatérale, si la différence observée va dans le sens annoncé. C'est précisément là que réside la tentation : un résultat avec un p bilatéral de 0,08 devient "significatif" (p = 0,04) si nous décidons, après avoir vu les données, que le test "aurait dû" être unilatéral.

Un test unilatéral se justifie rarement, et seulement si trois conditions sont remplies simultanément. Première condition : le sens est fixé avant le recueil des données, dans le protocole. Deuxième condition : une différence dans le sens opposé serait biologiquement impossible ou, si elle apparaissait, aurait la même conséquence pratique que l'absence de différence. Troisième condition : le lecteur en est explicitement informé. L'exemple classique est l'étude de non-infériorité, dans laquelle la question est seulement de savoir si le nouveau traitement n'est pas moins efficace de plus d'une marge acceptée (section 5.12). Dans les études courantes en médecine générale, un nouveau médicament peut aussi aggraver l'évolution du patient ; c'est pourquoi le test par défaut, exigé par la plupart des revues et utilisé dans ce cours, est le test bilatéral [42,53]. Les logiciels (JASP, MedCalc, Google Sheets) proposent par défaut la version bilatérale ; si vous choisissez autre chose, vous devez pouvoir dire pourquoi.

5.2 La logique du test : la statistique de test et sa distribution sous H0

Supposons que nous voulions savoir si la moyenne de la PAS dans la population dont proviennent les 160 patients diffère de 125 mmHg, une valeur de référence choisie pour l'exemple. La moyenne observée est de 131,13 mmHg. La différence de 6,13 mmHg pourrait être réelle, mais elle pourrait aussi être le fruit du hasard : si la moyenne de la population était vraiment de 125, un échantillon de 160 patients n'aurait pas exactement une moyenne de 125, mais une valeur voisine, plus grande ou plus petite. La bonne question devient : à quel point un écart de 6,13 mmHg est-il inhabituel si H0 est vraie ?

Pour répondre, nous comparons la différence observée (le "signal") à la variabilité que nous attendrions du seul fait du hasard (le "bruit"). La mesure du bruit pour une moyenne est l'erreur standard de la moyenne, $\text{ES} = \frac{s}{\sqrt{n}}$, que vous avez rencontrée au chapitre 4. Le rapport entre le signal et le bruit s'appelle la statistique de test :

$$t = \frac{m - \mu_0}{\text{ES}} = \frac{m - \mu_0}{\frac{s}{\sqrt{n}}}$$

Pour la PAS : $\text{ES} = \frac{13{,}88}{\sqrt{160}} = \frac{13{,}88}{12{,}649} = 1{,}097$ mmHg, donc $t = \frac{131{,}13 - 125}{1{,}097} = \frac{6{,}13}{1{,}097} = 5{,}59$. La moyenne observée se situe à près de 5,6 erreurs standard de la valeur affirmée par H0.

La statistique de test possède une qualité décisive pour le raisonnement : nous savons comment elle se distribue lorsque H0 est vraie. Si nous répétions l'étude un très grand nombre de fois, avec des échantillons de 160 patients tirés d'une population de moyenne exactement égale à 125, les valeurs t obtenues suivraient une loi de Student à n - 1 = 159 degrés de liberté, très

semblable à la loi normale centrée réduite : centrée sur zéro, symétrique, avec environ 95% des valeurs comprises entre -1,975 et +1,975. Une valeur t = 5,59 se situe bien en dehors de cette zone. Sous H0, elle serait extraordinairement rare.

Chaque test a sa propre statistique (z, t, F, χ^2, U, H), mais l'idée est la même : un nombre qui augmente quand les données s'éloignent de H0 et dont la distribution sous H0 est connue, par la théorie ou par des tables. Les valeurs de la statistique de test si extrêmes que, sous H0, elles apparaissent avec une probabilité inférieure à un seuil choisi (généralement 5%) forment la région critique, ou région de rejet. Sa frontière s'appelle la valeur critique.

5.3 La valeur p

5.3.1 La définition correcte

La valeur p est la probabilité d'obtenir, par le seul jeu du hasard, une statistique de test au moins aussi extrême que celle observée, en supposant que l'hypothèse nulle est vraie et que toutes les conditions d'application du test sont respectées. Chaque mot de la définition compte. "Au moins aussi extrême" signifie que nous additionnons la probabilité de la valeur observée et celle de toutes les valeurs plus éloignées de H0 ; dans un test bilatéral, nous additionnons les deux queues de la distribution. "En supposant que H0 est vraie" signifie que p se calcule dans un monde hypothétique où H0 est vraie ; la valeur p ne nous dit rien de la probabilité que ce monde soit le monde réel.

Pour la PAS, la valeur p bilatérale associée à t = 5,59 et 159 degrés de liberté est d'environ 0,0000001 (dans Google Sheets, =T.DIST.2T(5,59;159)). Autrement dit, si la moyenne de la population était de 125 mmHg, seul un échantillon de 160 patients sur dix millions environ aurait une moyenne au moins aussi éloignée de 125 que la nôtre. Une telle valeur se rapporte sous la forme "p < 0,001", car davantage de décimales n'apportent plus d'information utile.

5.3.2 Ce que la valeur p n'est pas

La valeur p est probablement le nombre le plus souvent mal interprété de la littérature médicale. L'American Statistical Association a publié en 2016 une déclaration officielle, en six principes, précisément pour mettre fin à ces confusions [74]. Un article de référence de Greenland et de ses collaborateurs recense plus de vingt interprétations erronées couramment rencontrées dans des articles publiés [70]. Nous retenons celles qui apparaissent aussi à l'examen de licence et au concours de résidanat.

La valeur p n'est pas la probabilité que H0 soit vraie. Un p = 0,03 ne signifie pas "3% de chances que le traitement n'ait pas d'effet". La probabilité de l'hypothèse, compte tenu des éléments observés, exige un autre type de calcul (le raisonnement bayésien, que vous rencontrerez avec les tests diagnostiques au chapitre 11), qui tient compte aussi de la plausibilité de l'hypothèse avant l'étude.

La valeur p n'est pas la probabilité que le résultat soit "dû au hasard". La formulation de nombreux cours anciens ("p est la probabilité que les différences observées soient apparues par hasard") est un raccourci commode, mais trompeur. La formulation correcte est : p est la probabilité de données aussi extrêmes si le seul mécanisme en jeu était le hasard.

La valeur p ne mesure pas la taille de l'effet ni son importance clinique. Un p = 0,0001 peut accompagner une différence de 0,04 dioptrie, sans aucune pertinence pour le patient, si l'échantillon est grand ; un p = 0,20 peut accompagner une réduction de la mortalité d'un tiers, si l'échantillon est petit. Nous y reviendrons avec des exemples à la section 5.6.

Un p > 0,05 ne prouve pas que H0 est vraie. L'absence de différence statistiquement significative ne signifie pas que la différence est absente ; elle signifie que les données n'ont pas suffi à la mettre en évidence. La section 5.4.2 est consacrée à cette idée.

Le seuil de 0,05 n'est pas une frontière naturelle. La différence entre p = 0,049 et p = 0,051 est négligeable en tant que preuve, bien que la convention les place de part et d'autre de la décision. Un résultat ne devient pas "vrai" lorsque p passe sous 0,05 et "faux" lorsqu'il passe au-dessus.

La valeur p ne dit rien de la qualité du schéma d'étude. Une étude avec une sélection biaisée des patients, des mesures erronées ou des facteurs de confusion non maîtrisés peut produire des valeurs p très petites pour des effets qui n'existent pas. Le test statistique vérifie seulement si les données, telles qu'elles sont, se concilient avec H0 ; il ne vérifie pas si les données sont bonnes.

5.3.3 La valeur p exacte et sa présentation

Autrefois, lorsque la valeur p se cherchait dans des tables imprimées, les auteurs écrivaient seulement "p < 0,05" ou "NS" (non significatif). Aujourd'hui, les logiciels calculent p exactement, et les recommandations de présentation exigent la valeur exacte, avec deux ou trois décimales : p = 0,034, p = 0,81, p = 0,0008. Pour les valeurs très petites, on écrit p < 0,001. On n'écrit jamais p = 0,000, comme l'affichent parfois les logiciels lorsqu'ils arrondissent : la probabilité n'est pas nulle, elle est seulement inférieure à 0,0005. L'écriture exacte permet au lecteur de juger par lui-même de la proximité du résultat avec le seuil et rend possible l'inclusion de l'étude dans une méta-analyse.

5.4 Le seuil de signification α et la règle de décision

5.4.1 Le seuil α et la décision

Le seuil de signification α est la probabilité maximale d'erreur de type I (rejet d'une H0 vraie) que le chercheur accepte. Il est fixé avant l'analyse, dans le protocole. Les valeurs conventionnelles sont 0,05, 0,01 et 0,001 ; en médecine clinique, α = 0,05 est la norme, et des seuils plus stricts apparaissent lorsque les conséquences d'un faux positif sont graves (par exemple l'autorisation d'un médicament) ou lorsque l'on effectue un très grand nombre de tests (les études génétiques utilisent des seuils de l'ordre de 0,00000005).

La règle de décision est simple :

- si p < α, on rejette H0 ; on dit que le résultat est statistiquement significatif au seuil α ;
- si p ≥ α, on ne rejette pas H0 ; le résultat n'est pas statistiquement significatif au seuil α.

Pour la PAS comparée à 125 mmHg, p < 0,001 < 0,05 ; nous rejetons donc H0 : la moyenne de la PAS dans la population de patients à risque cardiométabolique dont provient l'échantillon diffère significativement de 125 mmHg (elle est plus élevée). Si nous répétons le test avec la valeur de référence 130 mmHg, nous obtenons $t = \frac{131{,}13 - 130}{1{,}097} = 1{,}03$ et p = 0,304. Cette fois, p > 0,05 : nous ne rejetons pas H0, et les données sont compatibles avec une moyenne de la population de 130 mmHg. Le même échantillon, la même moyenne, deux conclusions différentes, parce que nous avons posé deux questions différentes.

5.4.2 Pourquoi "on ne rejette pas H0" et non "on accepte H0"

La formulation "on accepte H0", encore présente dans certains supports pédagogiques, suggère que nous avons démontré l'absence de différence. Nous n'avons rien démontré de tel. Le test statistique peut apporter des éléments contre H0, mais il ne peut pas en apporter en sa faveur.

Altman et Bland ont résumé l'idée dans un titre devenu proverbial : l'absence de preuve n'est pas la preuve de l'absence [79].

Un exemple tiré du jeu de données mg2026 rend la chose visible. Prenons seulement 10 patients au hasard et testons de nouveau si la moyenne de la PAS diffère de 125 mmHg. Les 10 patients ont les valeurs 150, 154, 131, 130, 137, 102, 110, 103, 147 et 136 mmHg, avec une moyenne de 130,0 et un écart-type de 19,04. L'erreur standard est de 19,04 / √10 = 6,02, et $t = \frac{130{,}0-125}{6{,}02} = 0{,}83$, avec 9 degrés de liberté ; p = 0,43. Nous ne rejetons pas H0. Il serait absurde d'"accepter" que la moyenne est de 125 mmHg, alors que sur l'ensemble des 160 patients nous avons trouvé une différence extrêmement significative. L'intervalle de confiance à 95% de la moyenne des 10 patients, 116,38-143,62 mmHg, montre pourquoi : il inclut 125, mais il inclut tout aussi bien 140. Le petit échantillon est compatible avec presque n'importe quelle valeur plausible. La conclusion correcte est : les données ne permettent pas de rejeter H0 ; la différence, si elle existe, n'a pas pu être mise en évidence avec cet échantillon.

En langage clinique, un résultat non significatif peut signifier deux choses très différentes : un effet absent ou un effet indéterminé. Nous les distinguons en regardant l'intervalle de confiance. Si l'intervalle est étroit et ne contient que des différences sans importance clinique, nous pouvons dire qu'un effet pertinent est peu probable. Si l'intervalle est large et contient à la fois zéro et des différences importantes, le résultat est indéterminé, et l'étude était probablement trop petite.

5.4.3 L'échelle verbale du cours et ses limites

Dans la tradition de la discipline, la décision s'accompagne d'une échelle verbale qui décrit à quel point la valeur p est petite [14,19] :

Tableau 5.2. L'échelle verbale utilisée dans le cours pour α = 0,05.

Valeur p	Décision	Formulation
p ≥ 0,05	on ne rejette pas H0	différence statistiquement non significative
0,01 ≤ p < 0,05	on rejette H0	différence significative
0,001 ≤ p < 0,01	on rejette H0	différence très significative
p < 0,001	on rejette H0	différence extrêmement significative

L'échelle est utile pour les devoirs et l'examen pratique, car elle oblige l'étudiant à lire attentivement la valeur p. Elle a toutefois deux limites que vous devez connaître. Première limite : "extrêmement significatif" décrit la force des éléments contre H0, et non la taille ou l'importance de l'effet. La taille diffère de 13,4 cm entre les femmes et les hommes de mg2026 (p < 0,001), et la PAS des 160 patients diffère de la valeur de 125 mmHg également avec p < 0,001, alors que, dans le second cas, la différence est de 6 mmHg ; l'échelle les place sur le même échelon. Seconde limite : l'échelle ne comporte pas d'échelons sous "non significatif" ; elle ne distingue donc pas un résultat indéterminé d'un résultat qui exclut un effet pertinent. C'est pourquoi, dans les travaux scientifiques, l'étiquette verbale ne remplace jamais la valeur p exacte, la taille de l'effet et l'intervalle de confiance.

Deux études de collègues de la discipline montrent à quoi ressemblent ces échelons en pratique. Dans une enquête menée auprès de 121 étudiants en médecine sur leurs compétences en science des données, la corrélation entre les connaissances mesurées et les connaissances autoévaluées était de R = 0,107, avec p = 0,246, donc non significative, tandis que la corrélation

entre les appréciations portées sur deux cours était de R = 0,869, p < 0,001, extrêmement significative [16]. Le résultat non significatif ne prouve pas que les étudiants sont incapables d'évaluer leurs connaissances ; il dit que, sur 121 répondants, une association faible n'a pas pu être distinguée de zéro. L'intervalle de confiance de R, s'il avait été rapporté, aurait montré l'étendue de la zone d'incertitude. Dans l'étude multicentrique sur les patients hémodialysés de 10 centres roumains, comparant la cohorte de 2015 (n = 1401) à celle de 2019 (n = 1698), la mortalité à un an était de 9,9% contre 10,7% (p = 0,46, non significatif), et la prévalence de l'infection par le virus de l'hépatite C est passée de 16,9% à 10,5% (p < 0,0001, extrêmement significatif) [39]. Sur un échantillon de plus de 3000 patients, une différence non significative de 0,8 point de pourcentage de la mortalité a une interprétation beaucoup plus solide que la même différence sur 100 patients : l'intervalle de confiance est étroit ; une variation importante de la mortalité est donc peu probable.

5.5 Les erreurs de type I et II, la puissance du test

5.5.1 Les quatre situations possibles

La décision statistique est prise à partir d'un échantillon ; elle peut donc être erronée. La vérité sur la population (H0 vraie ou fausse), nous ne la connaissons pas ; notre décision (rejeter ou ne pas rejeter H0), nous la connaissons. En combinant les deux, on obtient quatre situations.

Tableau 5.3. Les quatre résultats possibles d'un test statistique.

Décision prise	**En réalité, H0 est vraie (l'effet n'existe pas)**	**En réalité, H0 est fausse (l'effet existe)**
On rejette H0	erreur de type I (faux positif) ; probabilité α	décision correcte ; probabilité 1 - β (la puissance)
On ne rejette pas H0	décision correcte ; probabilité 1 - α (le niveau de confiance)	erreur de type II (faux négatif) ; probabilité β

L'erreur de type I consiste à déclarer un effet qui n'existe pas : un médicament inefficace paraît efficace, un facteur sans lien avec la maladie paraît être un facteur de risque. Sa probabilité, lorsque H0 est vraie, est exactement α, le seuil que nous avons choisi. L'erreur de type II consiste à manquer un effet qui existe : un traitement utile est abandonné parce que l'étude "n'a pas montré de différence". Sa probabilité est notée β. Le complément de β, 1 - β, est la puissance du test : la probabilité de détecter l'effet lorsqu'il existe réellement et qu'il a une certaine taille. Le complément de α, 1 - α, est le niveau de confiance : la probabilité de ne pas rejeter H0 lorsqu'elle est vraie ; c'est la même quantité qui apparaît dans le nom de l'intervalle de confiance à 95%.

5.5.2 Deux analogies : le verdict au tribunal et le test diagnostique

L'analogie la plus ancienne est celle du procès pénal. L'accusé est présumé innocent (H0) jusqu'à ce que les preuves le contredisent au-delà de tout doute raisonnable. Condamner un innocent est l'erreur de type I ; acquitter un coupable est l'erreur de type II. Le système judiciaire tolère davantage d'acquittements erronés que de condamnations erronées, de même que la statistique médicale fixe strictement α (0,05) et accepte un β plus grand (généralement 0,20). Remarquez encore un détail : le verdict est "non coupable" ou, plus exactement, "la culpabilité n'a pas été prouvée", et non "l'innocence a été prouvée". C'est exactement la même nuance qui sépare "on ne rejette pas H0" de "on accepte H0".

La seconde analogie vous sera familière grâce au chapitre sur les tests diagnostiques. Le test statistique se comporte comme un test de laboratoire appliqué à l'hypothèse : "positif" signifie le rejet de H0, "la maladie" signifie l'existence de l'effet. L'erreur de type I est un résultat faux positif, et α correspond à 1 - spécificité ; l'erreur de type II est un faux négatif, et la puissance 1 - β

correspond à la sensibilité. Comme pour un test diagnostique, les deux erreurs sont liées : si nous voulons moins de résultats faux positifs (α plus petit), nous aurons, pour le même échantillon, plus de résultats faux négatifs (β plus grand). L'analogie va plus loin encore. La valeur prédictive d'un test positif dépend de la prévalence de la maladie ; de même, la probabilité qu'un résultat "significatif" corresponde à un effet réel dépend de la plausibilité des hypothèses testées dans un domaine. Dans un domaine où l'on teste de nombreuses hypothèses peu plausibles, une part importante des résultats significatifs sont des faux positifs.

5.5.3 La représentation graphique de α, de β et de la puissance

La figure 5.1 place les deux mondes possibles sur le même axe. La courbe de gauche est la distribution de la statistique de test lorsque H0 est vraie ; la courbe de droite est sa distribution lorsqu'il existe un effet d'une certaine taille. La ligne pointillée verticale est la valeur critique (1,96 pour un test z bilatéral avec α = 0,05). L'aire sous la courbe H0, à droite de la valeur critique, vaut α/2 (l'autre moitié se trouve dans la queue gauche, non représentée). L'aire sous la courbe H1, à gauche de la valeur critique, vaut β : dans ces situations, l'effet existe, mais la statistique ne dépasse pas le seuil. L'aire sous la courbe H1, à droite de la valeur critique, est la puissance.

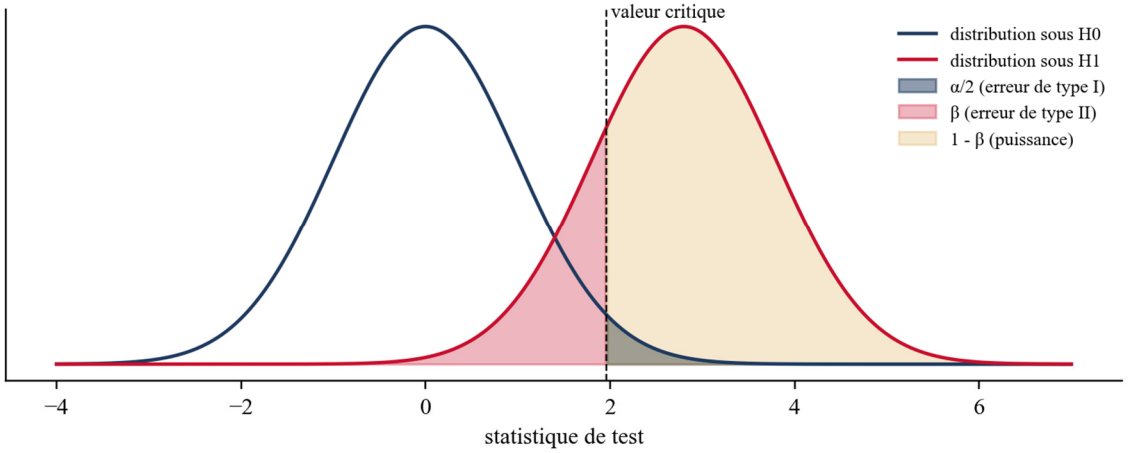


Figure 5.1. Distribution de la statistique de test sous l'hypothèse nulle (à gauche) et sous l'hypothèse alternative (à droite). La valeur critique (ligne pointillée) partage l'axe : l'aire de H0 à sa droite vaut α/2 (l'erreur de type I), l'aire de H1 à sa gauche vaut β (l'erreur de type II), et l'aire de H1 à sa droite est la puissance 1 - β.

La figure explique immédiatement pourquoi α et β varient en sens opposés. Si nous déplaçons la valeur critique vers la droite (α plus petit, par exemple 0,01), l'aire α diminue, mais l'aire β augmente. Les seules façons de réduire les deux erreurs en même temps sont celles qui éloignent les courbes l'une de l'autre ou qui les rétrécissent : un effet plus grand ou une erreur standard plus petite, c'est-à-dire davantage de patients ou des mesures plus précises.

5.5.4 Les facteurs qui influencent la puissance

La puissance d'un test dépend de quatre éléments, qui doivent être fixés à l'étape de planification de l'étude.

La taille de l'effet. Un effet important (par exemple une différence de 20 mmHg entre deux traitements antihypertenseurs) se détecte facilement, même avec peu de patients. Un petit effet exige de grands échantillons.

La variabilité. Plus l'écart-type de la variable est grand, plus le bruit couvre facilement le signal. Les mesures standardisées, les appareils étalonnés et les critères d'inclusion qui homogénéisent l'échantillon augmentent la puissance. Le schéma apparié (les mêmes patients mesurés deux fois) augmente la puissance précisément parce qu'il élimine la variabilité entre patients (chapitre 6).

La taille de l'échantillon. L'erreur standard diminue avec $\sqrt{n}$; les courbes de la figure 5.1 se rétrécissent donc quand n augmente. C'est le facteur que le chercheur contrôle le plus directement.

Le seuil α. Un α plus strict (0,01 au lieu de 0,05) réduit la puissance, et un test unilatéral (justifié) l'augmente, car toute l'aire α est placée dans une seule queue.

La figure 5.2 montre comment la puissance du test t pour deux groupes indépendants augmente avec le nombre de patients de chaque groupe, pour trois tailles d'effet exprimées en d de Cohen (section 5.6). Pour un effet moyen, d = 0,5, il faut 64 patients dans chaque groupe pour une puissance de 80% à α = 0,05 ; pour d = 0,3, il en faut 176 par groupe, pour d = 0,2 près de 394, et pour un grand effet, d = 0,8, 26 suffisent [76]. Si nous voulons une puissance de 90% au lieu de 80% pour d = 0,5, le nombre monte à 86 par groupe ; si nous gardons une puissance de 80% mais abaissons α à 0,01, il en faut 96 par groupe. Les calculs de ce type se font avec des logiciels dédiés, comme G*Power [77], ou avec les modules de calcul de la taille de l'échantillon de MedCalc et de JASP.

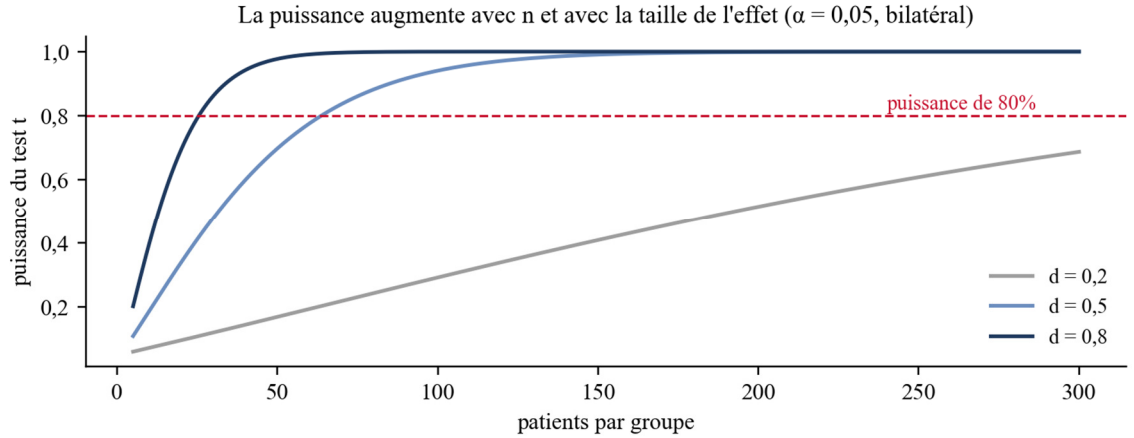


Figure 5.2. Puissance du test t pour deux groupes indépendants en fonction du nombre de patients de chaque groupe, pour trois tailles d'effet (d = 0,2 ; 0,5 ; 0,8), à α = 0,05 en bilatéral. La ligne horizontale marque la puissance de 80%, valeur minimale généralement acceptée lors de la planification d'une étude.

La convention usuelle, β = 0,20 (puissance de 80%), signifie que nous acceptons de manquer un effet réel dans une étude sur cinq. L'asymétrie par rapport à α = 0,05 traduit l'idée que déclarer un effet inexistant est considéré comme plus grave que manquer un effet réel. Dans certaines situations (une maladie grave sans autre traitement), ce rapport pourrait être discuté, mais la convention reste celle qu'utilisent les comités d'éthique et les revues.

Il faut aussi remarquer une conséquence moins intuitive de la faible puissance. Lorsqu'une étude sous-dimensionnée obtient malgré tout un résultat significatif, l'effet estimé est, en moyenne, exagéré : seuls les échantillons dans lesquels le hasard a "gonflé" la différence parviennent à franchir le seuil. C'est pourquoi les résultats spectaculaires de petites études doivent être confirmés dans des études plus grandes, et la présentation de la puissance (ou, mieux, de l'intervalle de confiance) aide le lecteur à juger de la confiance qu'il peut accorder à l'estimation.

Étude de cas : une puissance statistique limitée dans une cohorte observationnelle

Une étude prospective, observationnelle, menée chez 100 patients ayant bénéficié d'une cholangiopancréatographie rétrograde endoscopique (CPRE), a comparé deux stratégies de surveillance de la sédation : la sédation guidée par l'entropie (un indice dérivé de l'électroencéphalogramme) et la surveillance standard [78]. Des épisodes hypertensifs sont survenus chez 6% contre 26% des patients (p = 0,007), et les scores de récupération d'Aldrete ont été plus élevés à 5 minutes (p = 0,003) et à 15 minutes (p < 0,001). Les auteurs rapportent explicitement une puissance statistique limitée, de 29,3%. Qu'est-ce que cela signifie pour le

> lecteur ? Une puissance de 29,3% signifie que, si l'effet supposé dans le calcul existait réellement, une étude de cette taille ne le détecterait que dans environ trois cas sur dix. Pour l'analyse à laquelle se rapporte le calcul, β serait donc d'environ 70%, et un résultat non significatif dirait très peu de chose sur l'absence d'effet. Pour les résultats significatifs, la faible puissance a une autre conséquence : les estimations de l'effet peuvent être exagérées, et une confirmation dans une étude plus grande est nécessaire. Le second point de lecture critique est le schéma d'étude. Les groupes n'ont pas été randomisés ; les différences peuvent donc refléter aussi d'autres caractéristiques des patients ou des équipes. Les auteurs présentent les résultats avec prudence, comme une association ("is associated with"), et non comme un effet causal, ce qui est exactement la formulation correcte pour une étude observationnelle de puissance limitée. Pour vos devoirs, retenez l'ordre : d'abord le schéma d'étude, puis la puissance, puis la valeur p.

5.6 La taille de l'effet et la signification clinique

5.6.1 Les mesures de la taille de l'effet

La valeur p répond à la question "existe-t-il un effet ?", mais pas à celle qui intéresse le médecin : "quelle est la taille de l'effet ?". Pour cette seconde question, nous utilisons la taille de l'effet (en anglais effect size), une mesure qui ne dépend pas de la taille de l'échantillon. Les tailles d'effet peuvent s'exprimer dans les unités d'origine (une différence de 10,19 mg/dL, une différence de risques de 18,3 points de pourcentage) ou sous une forme standardisée, qui permet la comparaison entre des études et des variables différentes.

La différence moyenne standardisée, le d de Cohen, divise la différence entre deux moyennes par l'écart-type commun :

$$d = \frac{m_1 - m_2}{s_{\text{commune}}}$$

Un d = 0,5 signifie que les moyennes sont éloignées d'un demi-écart-type l'une de l'autre. Pour les données appariées, on utilise dz = moyenne des différences / écart-type des différences. Le coefficient de corrélation r est lui-même une taille d'effet, et r² indique la proportion de la variabilité expliquée. En ANOVA (chapitre 6), η^2 = SS intergroupes / SS totale indique quelle proportion de la variabilité totale est due à l'appartenance au groupe. Pour les variables binaires, le rapport de cotes (OR) et le risque relatif (RR) sont des tailles d'effet (chapitre 10) ; il n'existe pas pour eux de seuils universels, et l'interprétation dépend de la fréquence de la maladie et du contexte.

Tableau 5.4. Seuils indicatifs pour l'interprétation de la taille de l'effet (d'après Cohen [76]) ; à utiliser uniquement en l'absence de critère clinique.

Mesure	Petit effet	Effet moyen	Grand effet
d de Cohen, dz	0,2	0,5	0,8
r (corrélation)	0,1	0,3	0,5
η^2 (ANOVA)	0,01	0,06	0,14

Les seuils de Cohen sont des conventions générales, proposées à l'origine pour les sciences du comportement, et ils ne remplacent pas le jugement clinique. Une réduction de la PAS de 0,2 écart-type (environ 3 mmHg) est "petite" selon Cohen, mais, appliquée à l'échelle d'une population entière, elle a des effets réels sur le nombre d'accidents vasculaires.

Quelques exemples tirés de mg2026 fixent les ordres de grandeur. La différence de taille entre les hommes et les femmes a un d = 2,19, un effet énorme, visible à l'œil nu. La baisse de la glycémie entre l'inclusion et 6 mois a un dz = 0,88, un grand effet. La moyenne de la PAS comparée à 125 mmHg correspond à un $d = \frac{131{,}13-125}{13{,}88} = 0{,}44$, un effet petit à moyen. La différence de PAS entre les sexes a un d = -0,04, pratiquement nul. Dans l'ANOVA de la baisse de la glycémie selon les bras, $\eta^2 = 0{,}226$; près de 23% de la variabilité de la baisse de la glycémie est donc expliquée par le bras de traitement, un grand effet.

5.6.2 Signification statistique et signification clinique

Un résultat est statistiquement significatif lorsque les données sont difficiles à concilier avec H0. Il est cliniquement significatif (cliniquement pertinent) lorsque l'effet est assez grand pour changer la décision du médecin ou la vie du patient. Les deux ne coïncident pas automatiquement. Le seuil de pertinence clinique s'appelle la différence minimale cliniquement importante (en anglais minimal clinically important difference, MCID) : la plus petite variation que les patients perçoivent comme bénéfique ou qui justifie un changement de traitement. Elle se fixe à partir de considérations cliniques, d'études menées auprès de patients ou des recommandations, et non à partir de la statistique, et elle devrait être fixée avant l'analyse des données.

En combinant les deux critères, on obtient quatre situations qu'il vaut la peine de savoir reconnaître rapidement :

- statistiquement significatif et cliniquement pertinent : effet certain et important ; l'exemple idéal ;
- statistiquement significatif, mais cliniquement non pertinent : typique des très grands échantillons ; l'effet existe, mais il est trop petit pour compter ;
- statistiquement non significatif, avec un intervalle de confiance qui inclut des effets importants : résultat indéterminé, étude probablement trop petite ;
- statistiquement non significatif, avec un intervalle de confiance étroit autour de zéro : effet pertinent peu probable.

Le premier exemple tiré de mg2026 est la différence de PAS entre les sexes. Les femmes ont une moyenne de 131,37 mmHg (ÉT 14,47 ; n = 86), les hommes de 130,85 mmHg (ÉT 13,25 ; n = 74). Le test t de Welch donne t = -0,24, p = 0,81 ; le résultat est donc non significatif. La différence estimée hommes moins femmes est de -0,5 mmHg, avec un IC à 95% de -4,9 à 3,8 mmHg, et d = -0,04. S'agit-il d'un effet indéterminé ou d'un effet absent ? L'intervalle de confiance exclut des différences supérieures à environ 5 mmHg dans un sens comme dans l'autre. Si nous admettons qu'une différence moyenne de PAS inférieure à 5 mmHg entre les sexes n'a pas de conséquences pratiques dans ce groupe, nous pouvons dire qu'une différence cliniquement pertinente est peu probable. La puissance du test avec 86 et 74 patients était de 88% pour un effet moyen (d = 0,5, soit environ 7 mmHg), mais seulement de 24% pour un petit effet (d = 0,2, environ 3 mmHg). Par conséquent, l'étude pouvait détecter une différence moyenne de 7 mmHg, mais elle ne peut rien nous dire de sûr sur une différence de 2 à 3 mmHg.

Le second exemple est la baisse de la glycémie à jeun entre l'inclusion et 6 mois chez les mêmes 160 patients : de 105,4 à 95,2 mg/dL, avec une différence moyenne de 10,19 mg/dL (ÉT 11,57), IC à 95% 8,38-11,99 mg/dL, t = 11,14, 159 degrés de liberté, p < 0,001. C'est un résultat extrêmement significatif. Est-il aussi cliniquement important ? Ici, il faut faire entrer la médecine dans la discussion. Chez un patient dont la glycémie à jeun se situe dans la zone du prédiabète (100-125 mg/dL, selon les critères de l'ADA [80]), une baisse de 10 mg/dL peut faire passer la valeur sous le seuil de 100 mg/dL ; en tant que moyenne de groupe, l'effet est donc pertinent. Si, par l'absurde, nous avions établi à l'avance que seule une baisse d'au moins 15 mg/dL compte cliniquement, l'intervalle de confiance (8,38-11,99) aurait montré que l'effet moyen se situe à

coup sûr sous ce seuil, bien qu'il soit extrêmement significatif sur le plan statistique. Le même chiffre, deux conclusions cliniques différentes, selon le seuil de pertinence choisi. Il faut encore remarquer qu'une étude sans groupe témoin ne peut pas attribuer la baisse à l'intervention : la glycémie pourrait aussi baisser par régression vers la moyenne ou par l'effet de la participation à l'étude. Dans mg2026, nous avons cependant trois bras randomisés, et la comparaison correcte de l'efficacité est celle qui oppose les bras (chapitre 6).

La littérature des collègues offre des exemples des deux pièges. Dans l'analyse sur quatre ans de la prise en charge du cancer du côlon dans un hôpital universitaire roumain, portant sur 867 patients, les auteurs ont rapporté une corrélation inverse entre l'âge du patient et la durée de l'intervention chirurgicale, r = -0,088, p = 0,01 [55]. Le résultat est statistiquement significatif, mais $r^2 = 0{,}0077$, c'est-à-dire que l'âge explique moins de 1% de la variabilité de la durée de l'opération. Dans un échantillon aussi grand, presque toute association différente de zéro devient significative. Dans l'étude sur les hémodialysés, l'âge moyen différait entre les cohortes (59,7 ± 12,92 contre 61 ± 12,93 ans, p = 0,005) [39] ; une différence de 1,3 an entre deux cohortes de près de 1500 patients chacune est très significative sur le plan statistique, mais n'a pas d'importance clinique en soi. Une étude de concordance entre trois appareils ophtalmologiques, portant sur 170 yeux, a trouvé pour la kératométrie K1 des valeurs de 43,45 ± 1,64 D avec un appareil et de 43,41 ± 1,70 D avec un autre, une différence significative (p = 0,019) [81] ; une différence de 0,04 dioptrie est pourtant bien inférieure à la précision nécessaire pour le calcul d'un implant cristallinien. En sens inverse, l'étude sur les aides auditives intelligentes a rapporté un effet considérable : les patients atteints de surdité neurosensorielle qui portaient des prothèses auditives avaient des scores d'estime de soi de 35,41 ± 5,32 contre 22,99 ± 4,53 chez ceux qui n'en portaient pas (p < 0,001, d de Cohen = 2,515), et l'ANOVA sur trois groupes a donné F = 299,00 avec η^2 = 0,712 [82]. Ici, signification statistique et signification pratique vont de pair, même si, s'agissant d'une étude transversale, le résultat décrit une association et ne démontre pas que la prothèse a augmenté l'estime de soi.

Étude de cas : une corrélation très faible, mais significative

L'analyse rétrospective de 867 patients atteints d'un cancer du côlon, traités avant et pendant la pandémie de COVID-19 dans un hôpital universitaire de Roumanie, a comparé les périodes au moyen d'une série de tests et a constaté une baisse marquée des interventions programmées (p < 0,001), une augmentation des interventions en urgence (p < 0,001) et une augmentation des cas de stade IV au cours de la deuxième année de pandémie (p = 0,045) [55]. Comme résultat secondaire, le résumé rapporte une corrélation inverse entre l'âge et la durée de l'opération, r = -0,088, p = 0,01. L'exercice de lecture critique comporte trois étapes. D'abord, la significativité : avec n = 867, l'erreur standard de r est d'environ $\frac{1}{\sqrt{n}}$, soit 0,034, et tout |r| supérieur à environ 0,067 devient significatif à α = 0,05 ; le seuil de signification est facilement franchi. Ensuite, la taille : $r^2 = 0{,}0077$, moins de 1% de la variabilité, une association négligeable selon n'importe quelle échelle (celle de Colton, utilisée dans le cours, considère comme "absence de corrélation" tout |r| inférieur à 0,25). Enfin, l'interprétation : le résultat ne permet pas de prédire la durée de l'opération à partir de l'âge et ne justifie aucune décision clinique. Les résultats principaux de l'étude (la modification de la structure des interventions) sont des différences de proportions qui ont une réelle pertinence pour l'organisation des services. Qu'est-ce qui aurait ajouté de la valeur à la présentation ? L'intervalle de confiance de r et, pour les comparaisons de proportions, les différences absolues avec leurs intervalles de confiance. La leçon pour l'étudiant est que "significatif" ne veut pas dire "important", surtout lorsque n dépasse quelques centaines.

5.7 Trois voies vers la même décision : la valeur critique, la valeur p et l'intervalle de confiance

La décision d'un test bilatéral au seuil α peut être prise par trois voies, qui conduisent toujours à la même conclusion. Nous les montrons sur l'exemple de la PAS, avec H0 : μ = 125 mmHg, n = 160, m = 131,13 mmHg, s = 13,88 mmHg, ES = 1,097 mmHg.

La voie de la valeur critique. Nous calculons la statistique de test, t = 5,59, et nous la comparons à la valeur critique de la loi de Student pour 159 degrés de liberté et α = 0,05 en bilatéral, t critique = 1,975 (dans Google Sheets, =T.INV.2T(0,05;159)). La règle : on rejette H0 si |t| > t critique. Comme 5,59 > 1,975, nous rejetons H0. C'était la voie utilisée lorsque les calculs se faisaient avec des tables imprimées.

La voie de la valeur p. Nous calculons la probabilité d'une valeur |t| au moins égale à 5,59 sous H0 : p = T.DIST.2T(5,59;159), environ 0,0000001. La règle : on rejette H0 si p < α. Comme p < 0,05, nous rejetons H0. C'est la voie utilisée par tous les logiciels modernes, et elle a l'avantage de montrer la force des preuves, et non seulement de dire si le seuil a été franchi.

La voie de l'intervalle de confiance. Nous construisons l'intervalle de confiance à 1 - α = 95% de la moyenne de la population : $m \pm t$ critique × ES = $131{,}13 \pm 1{,}975 \times 1{,}097 = 131{,}13 \pm 2{,}17$, c'est-à-dire 128,96-133,30 mmHg (les mêmes valeurs qu'au chapitre 4). La règle : on rejette H0 si la valeur affirmée par H0 (ici 125) se situe en dehors de l'intervalle. La valeur 125 est en dehors de l'intervalle ; nous rejetons donc H0. Cette voie est celle qui a le plus grand contenu informatif : elle montre à la fois la décision, le sens et la taille de l'effet, ainsi que la précision de l'estimation.

Cette équivalence n'est pas une coïncidence. La valeur 125 sort de l'intervalle exactement lorsque |m - 125| > t critique × ES, c'est-à-dire exactement lorsque |t| > t critique, c'est-à-dire exactement lorsque p < 0,05. Vérifiez la même chose pour la valeur de référence 130 : t = 1,03 < 1,975 ; p = 0,304 > 0,05 ; 130 se situe dans l'intervalle 128,96-133,30. Les trois voies disent toutes : on ne rejette pas H0.

Tableau 5.5. Les trois voies de décision pour la moyenne de la PAS (n = 160, m = 131,13 mmHg, α = 0,05).

Valeur de H0	t calculé	t critique	Valeur p	L'IC à 95% contient-il μ0 ?	Décision
$\mu_0 = 125$ mmHg	5,59	1,975	< 0,001	non	on rejette H0
$\mu_0 = 130$ mmHg	1,03	1,975	0,304	oui	on ne rejette pas H0
$\mu_0 = 140$ mmHg	-8,08	1,975	< 0,001	non	on rejette H0

La correspondance fonctionne pour un IC à 1 - α et un test bilatéral au seuil α : IC à 95% avec α = 0,05, IC à 99% avec α = 0,01. L'intervalle à 99% de la PAS, 128,27-133,99 mmHg, exclut lui aussi 125 ; le résultat reste donc significatif également à α = 0,01.

5.8 L'algorithme d'un test statistique

Quel que soit le test, la résolution suit les mêmes étapes. Nous les numérotons ici et nous les utiliserons dans tous les chapitres suivants, ainsi que dans le format a-g des devoirs de la discipline.

1. Formulez l'hypothèse de recherche en langage clinique (format des devoirs : point a).
2. Formulez les hypothèses statistiques H0 et H1, en mots et en notation mathématique, avec les paramètres de la population (point b). Précisez si le test est bilatéral ou unilatéral.
3. Fixez le seuil de signification α (par défaut 0,05) avant de regarder les résultats.
4. Identifiez le type des variables et le schéma d'étude (une moyenne comparée à une valeur, deux groupes indépendants, deux mesures chez les mêmes patients, plus de deux groupes, variables qualitatives) et choisissez le test (point c). L'algorithme complet de choix figure au chapitre 8.
5. Vérifiez les conditions d'application du test (normalité, égalité des variances, indépendance des observations, effectifs attendus) ; si elles ne sont pas remplies, choisissez l'alternative appropriée, généralement un test non paramétrique.
6. Calculez la statistique de test, à la main ou avec le logiciel (point d), et obtenez la valeur p (et/ou la valeur critique).
7. Calculez la taille de l'effet et l'intervalle de confiance de l'effet (point e, avec le graphique).
8. Prenez la décision statistique : $p < \alpha$, on rejette H0 ; $p \geq \alpha$, on ne rejette pas H0 ; ajoutez l'étiquette verbale du cours (point f).
9. Formulez la conclusion de recherche en mots, avec les principaux chiffres et l'interprétation clinique, y compris les limites (point g).

Les étapes 3 et 4 se font avant l'étape 6. Changer de test ou de seuil après avoir vu la valeur p transforme l'analyse en une recherche de résultats commodes (section 5.11).

5.9 Les tests z et t pour une seule moyenne

Le test le plus simple compare la moyenne d'un échantillon à une valeur donnée à l'avance : une valeur de référence tirée de la littérature, une cible thérapeutique, la moyenne nationale. Le cours de la discipline distingue deux variantes, le test z et le test t, selon ce que nous savons de la variabilité de la population.

5.9.1 Le test z

Le test z s'utilise lorsque l'écart-type de la population, σ, est connu (par exemple grâce à un grand registre national ou aux spécifications d'un appareil de mesure) et que l'échantillon est suffisamment grand ($n \geq 30$) pour que la moyenne suive une loi normale, même si la variable elle-même n'est pas parfaitement normale (théorème central limite, chapitre 4). La statistique de test est :

$$z = \frac{m - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Sous H0, z suit la loi normale centrée réduite. Les valeurs critiques bilatérales sont de 1,96 pour $\alpha = 0{,}05$, 2,576 pour $\alpha = 0{,}01$ et 3,291 pour $\alpha = 0{,}001$; pour un test unilatéral avec $\alpha = 0{,}05$, la valeur critique est de 1,645.

Supposons, pour l'exemple, qu'une grande étude en population ait établi, pour les adultes du même âge de la région, un cholestérol total moyen de 200 mg/dL, avec $\sigma = 40$ mg/dL. Dans mg2026, la moyenne du cholestérol est de 209,39 mg/dL sur n = 160 patients. L'erreur standard calculée avec σ connu est de 40 / √160 = 40 / 12,649 = 3,162 mg/dL, et $z = \frac{209{,}39 - 200}{3{,}162} = \frac{9{,}39}{3{,}162} =$ 2,97. Comme 2,97 > 1,96 (et même > 2,576), nous rejetons H0 ; la valeur p bilatérale est de 2 ×

P(Z > 2,97) = 0,003 ; la différence est donc très significative. Les patients à risque cardiométabolique de l'étude ont un cholestérol moyen plus élevé que la population générale de référence.

5.9.2 Le test t pour une seule moyenne

En pratique, σ est presque toujours inconnu et nous l'estimons par l'écart-type de l'échantillon, s. Remplacer σ par s ajoute de l'incertitude, surtout dans les petits échantillons, et la statistique

$$t = \frac{m - \mu_0}{\frac{s}{\sqrt{n}}}$$

ne suit plus la loi normale, mais la loi t de Student, décrite en 1908 par William Gosset, chimiste dans une brasserie de Dublin, qui publiait sous le pseudonyme "Student" [83]. La loi t ressemble à la loi normale, mais ses queues sont plus épaisses ; sa forme dépend du nombre de degrés de liberté, ddl = n - 1. Plus n est grand, plus la loi t se rapproche de la loi normale, et au-delà de 120 ddl la différence devient négligeable.

Tableau 5.6. Valeurs critiques de la loi t (bilatéral α = 0,05 et 0,01 ; unilatéral α = 0,05).

ddl	bilatéral 0,05	bilatéral 0,01	unilatéral 0,05
4	2,776	4,604	2,132
9	2,262	3,250	1,833
15	2,131	2,947	1,753
19	2,093	2,861	1,729
29	2,045	2,756	1,699
40	2,021	2,704	1,684
60	2,000	2,660	1,671
120	1,980	2,617	1,658
159	1,975	2,607	1,654
infini (z)	1,960	2,576	1,645

Pour le cholestérol, avec $s = 36{,}91$ mg/dL au lieu de σ = 40, ES $= \frac{36{,}91}{12{,}649} = 2{,}918$ mg/dL et $t = \frac{9{,}39}{2{,}918} = 3{,}22$, avec 159 degrés de liberté ; p = 0,0016. La conclusion ne change pas. Pour n = 160, le choix entre z et t compte peu ; pour n = 10, il aurait beaucoup compté (la valeur critique serait de 2,262 au lieu de 1,96).

La règle du cours (test z pour n ≥ 30 et σ connu, test t pour n < 30 ou σ inconnu) peut être simplifiée pour la pratique actuelle : lorsque σ n'est pas connu, nous utilisons le test t quelle que soit la taille de l'échantillon. C'est exactement ce que font les logiciels statistiques ; JASP, par exemple, propose dans le menu T-Tests > One Sample T-Test le test t de Student et, en option séparée, le test z, pour lequel il demande de saisir la valeur de σ.

5.9.3 Les conditions d'application

Le test t pour une moyenne suppose que les observations sont indépendantes (chaque patient n'intervient qu'une fois) et que la variable a, dans la population, une distribution approximativement normale. La seconde condition compte surtout pour les petits échantillons ; pour n grand, le théorème central limite rend le test robuste à des écarts modérés à la normalité. La normalité se vérifie avec le test de Shapiro-Wilk [66] et le diagramme Q-Q (chapitre 3). Si la variable est fortement asymétrique et que l'échantillon est petit, l'alternative non paramétrique est le test des rangs signés de Wilcoxon pour une seule médiane (chapitre 7) ou une transformation des données (par exemple le passage aux logarithmes des triglycérides).

Pour la glycémie à jeun de mg2026, le test de Shapiro-Wilk donne p = 0,037, soit un léger écart à la normalité (asymétrie 0,37). Avec n = 160, le test t reste utilisable, et le test de Wilcoxon conduit à la même conclusion (problème résolu 5.2). Pour l'IMC, l'asymétrie est importante (1,42) et Shapiro-Wilk p < 0,001 ; ici, la présentation de la médiane et d'un test non paramétrique est plus appropriée, même si n est grand.

5.10 Les comparaisons multiples et l'inflation de l'erreur de type I

5.10.1 Pourquoi le risque de faux positif augmente

Le seuil α = 0,05 contrôle l'erreur de type I pour un seul test. Si nous effectuons plusieurs tests indépendants, chacun à α = 0,05, la probabilité d'obtenir au moins un résultat faux positif augmente rapidement :

$$P(\text{au moins un fals positif}) = 1 - (1-\alpha)^k = 1 - 0{,}95^k$$

Pour k = 3 tests, la probabilité est de 0,143 ; pour k = 5, de 0,226 ; pour k = 10, de 0,401 ; pour k = 20, de 0,642. Autrement dit, un étudiant qui compare les femmes et les hommes sur 20 variables pour lesquelles, en réalité, il n'existe aucune différence a près de deux chances sur trois de "découvrir" au moins une différence significative.

La figure 5.3 illustre le phénomène de deux façons. Dans le panneau A, nous avons simulé 1000 études dans lesquelles deux groupes de 20 patients proviennent de la même population (PAS de moyenne 130 et d'ÉT 14 mmHg) ; H0 est donc vraie à chaque fois. Les valeurs p sont distribuées uniformément entre 0 et 1, et 4,9% d'entre elles tombent sous 0,05 : exactement la proportion de résultats faux positifs promise par α. Le panneau B montre la courbe $1 - 0{,}95^k$ et l'effet de la correction de Bonferroni, qui maintient le risque global proche de 5%. Vous referez la même simulation dans Google Sheets lors des travaux pratiques 5.

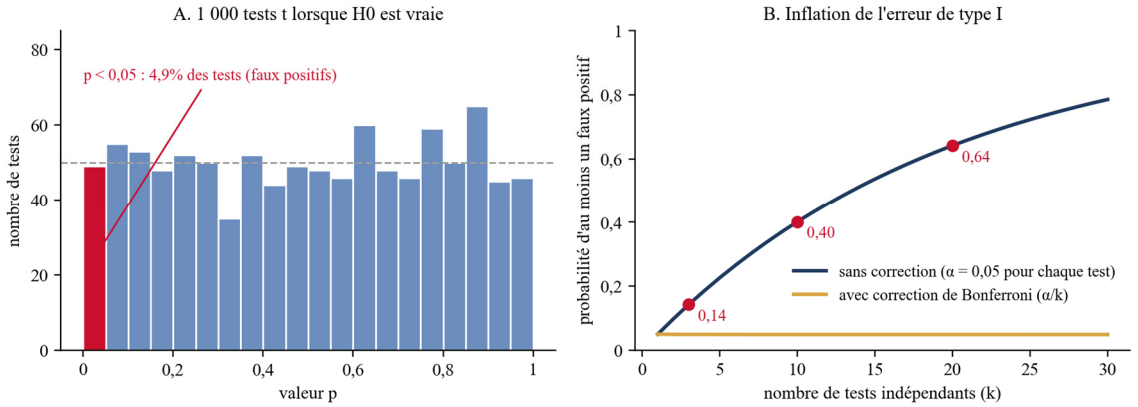


Figure 5.3. A. Distribution des valeurs p dans 1000 tests t simulés lorsque l'hypothèse nulle est vraie (deux groupes de 20 valeurs issues de la même population normale) : les valeurs p sont uniformément distribuées, et environ 5% d'entre elles sont inférieures à 0,05. B. Probabilité d'obtenir au moins un résultat faux positif en fonction du nombre de tests indépendants, sans correction et avec la correction de Bonferroni. Données simulées.

5.10.2 Les corrections de Bonferroni, de Holm et de Benjamini-Hochberg

La correction de Bonferroni divise le seuil par le nombre de tests : chaque test n'est déclaré significatif que si p < α/k. De façon équivalente, les logiciels rapportent des valeurs p ajustées, p × k (plafonnées à 1), que l'on compare à l'α habituel. La méthode est simple et sûre, mais conservatrice : avec de nombreux tests, elle perd beaucoup de puissance.

La méthode de Holm (Bonferroni séquentielle) offre la même garantie, mais elle est moins conservatrice. Nous rangeons les valeurs p par ordre croissant, p(1) ≤ p(2) ≤... ≤ p(k). Nous

comparons p(1) à α/k, p(2) à α/(k - 1), p(3) à α/(k - 2), et ainsi de suite. Nous nous arrêtons au premier test qui ne franchit pas le seuil ; ce test et tous les suivants sont déclarés non significatifs.

La méthode de Benjamini-Hochberg (BH) contrôle une autre quantité : non pas la probabilité d'avoir au moins un faux positif, mais la proportion attendue de faux positifs parmi les résultats déclarés significatifs (le taux de fausses découvertes, en anglais false discovery rate, FDR). Nous rangeons les valeurs p par ordre croissant et comparons p(i) à (i/k) × α. Nous trouvons le plus grand i pour lequel p(i) ≤ (i/k) × α et déclarons significatifs tous les tests de 1 à ce i. La méthode convient lorsque nous effectuons de nombreux tests exploratoires et que nous acceptons qu'une petite partie des "découvertes" soient fausses, à condition de savoir quelle est l'importance de cette partie.

Sur le jeu de données mg2026, nous pouvons faire exactement l'exercice tentant décrit plus haut : comparer les femmes et les hommes sur 14 variables numériques mesurées à l'inclusion, avec le test t de Welch. Les résultats, classés selon la valeur p, figurent dans le tableau 5.7.

Tableau 5.7. Comparaison femmes-hommes sur 14 variables de mg2026, avec les valeurs p non corrigées et ajustées (test t de Welch).

Variable	Moyenne F	Moyenne H	p non corrigé	p Bonferroni	p Holm	p BH
taille (cm)	162,60	175,97	< 0,001	< 0,001	< 0,001	< 0,001
poids (kg)	76,53	90,64	< 0,001	< 0,001	< 0,001	< 0,001
HDL(mg/dL)	53,92	44,41	< 0,001	< 0,001	< 0,001	< 0,001
tour de taille (cm)	86,34	91,09	0,013	0,18	0,14	0,046
triglycérides (mg/dL)	171,21	149,51	0,068	0,95	0,68	0,19
âge (ans)	55,42	57,31	0,17	1	1	0,40
FINDRISC	10,02	10,77	0,27	1	1	0,55
glicemie$_0$(mg/d	104,81	106,16	0,59	1	1	0,79
PAD (mmHg)	78,88	79,62	0,59	1	1	0,79
cholestérol (mg/dL)	208,08	210,91	0,63	1	1	0,79
HbA1c (%)	5,56	5,59	0,68	1	1	0,79
IMC $\left(\frac{\text{kg}}{m^2}\right)$	28,92	29,18	0,68	1	1	0,79
PAS (mmHg)	131,37	130,85	0,81	1	1	0,88
FC (battements/ min)	74,69	74,73	0,98	1	1	0,98

La taille, le poids et le HDL diffèrent entre les sexes quelle que soit la correction ; ce sont des différences biologiques bien connues. Le tour de taille a un p non corrigé de 0,013 : significatif sans correction, non significatif après Bonferroni (le seuil individuel devient 0,05/14 = 0,0036) et après Holm, mais encore significatif après Benjamini-Hochberg (p ajusté 0,046). La différence entre les méthodes n'est pas une contradiction ; elle reflète des questions différentes. Si nous ne voulons aucun faux positif dans l'ensemble du tableau, nous utilisons Bonferroni ou Holm ; si nous acceptons qu'environ 5% des résultats déclarés positifs puissent être faux, BH suffit.

Quand une correction est-elle nécessaire ? Lorsque les tests relèvent d'une même question et que la conclusion se tire de "au moins un résultat significatif" : comparaisons multiples entre groupes après une ANOVA (chapitre 6), nombreux critères de jugement dans une même étude, analyses dans de nombreux sous-groupes, analyses intermédiaires répétées des données. Elle n'est

pas nécessaire pour un critère principal unique, fixé à l'avance dans le protocole. La meilleure protection reste le choix d'un seul critère principal avant le recueil des données ; les critères secondaires sont rapportés comme exploratoires.

Un exemple tiré de la littérature des collègues montre avec quelle facilité les tests s'accumulent. L'étude transversale sur l'association entre la sévérité de l'infection par le VIH et les troubles psychiatriques, portant sur 101 patients, rapporte sept rapports de cotes, avec des valeurs p comprises entre 0,038 et moins de 0,001 (par exemple OR = 5,81, p < 0,001 ; OR = 2,81, p = 0,038) [84]. Chaque résultat pris isolément est significatif. Avec sept tests, le seuil de Bonferroni $\frac{\text{ar fi}\quad 0{,}05}{7} = 0{,}0071$; les résultats avec p = 0,016, 0,024 ou 0,038 ne le franchiraient pas. Ils peuvent être réels, mais doivent être considérés comme des hypothèses à vérifier, et non comme des résultats établis. La présentation des intervalles de confiance de chaque OR aurait aussi aidé le lecteur à juger de la précision des estimations.

Étude de cas : la correction de Benjamini-Hochberg dans une étude menée aux urgences

Une étude rétrospective menée dans un service d'urgences a analysé les différences entre femmes et hommes chez 213 patients (100 femmes, 113 hommes) traités pour une suspicion d'infection bactérienne [9]. Les femmes étaient plus âgées (médiane 77,0 contre 67,0 ans, p < 0,001), et l'origine urinaire de l'infection était plus fréquente chez les femmes (51,0% contre 34,5%, p = 0,015). Pour cette association, le résumé rapporte un OR de 2,14 (IC à 95% 1,20-3,81 ; p = 0,010) et, après correction de Benjamini-Hochberg pour les tests multiples, un p ajusté = 0,039. Les auteurs ont comparé de nombreuses caractéristiques entre les sexes ; la correction pour tests multiples était donc justifiée, et le résultat est resté significatif après celle-ci, ce qui renforce sa crédibilité. Pour les critères cliniques, le tableau est différent : la mortalité hospitalière était de 7,0% chez les femmes (7 sur 100) et de 8,0% chez les hommes (9 sur 113), avec un OR de 0,87 (IC à 95% 0,31-2,43), et un autre critère est survenu chez 14 femmes sur 100 (14,0%) et chez 21 hommes sur 113 (18,6%), OR 0,71 (IC à 95% 0,34-1,49). Le résumé rapporte aussi un OR ajusté de 0,80 (IC à 95% 0,27-2,35 ; p = 0,690). Les intervalles de confiance sont très larges, allant d'une division des cotes par près de trois à leur multiplication par plus de deux. Avec 16 décès au total, l'étude ne peut exclure ni un effet protecteur ni un effet défavorable important ; le résultat est indéterminé, et non "absent". Les leçons sont au nombre de trois : la correction pour comparaisons multiples s'applique lorsque l'on teste de nombreuses hypothèses dans une même étude ; un résultat qui résiste à la correction est plus convaincant ; pour des événements rares, l'intervalle de confiance en dit plus que la valeur p. La comparaison entre l'OR brut et l'OR ajusté (ici, tous deux non significatifs) sera reprise dans les chapitres consacrés à l'analyse du risque et à la régression logistique.

5.11 p-hacking, HARKing et préenregistrement

La valeur p ne garde son sens que si l'analyse a été décidée avant de voir les données. Lorsque le chercheur essaie plusieurs analyses et ne rapporte que celle "qui a marché", le seuil nominal de 5% n'est plus réel. Cette pratique s'appelle le p-hacking (la chasse aux valeurs p) et prend des formes apparemment innocentes : on ajoute quelques patients jusqu'à ce que p descende sous 0,05 ; on exclut les "valeurs aberrantes" seulement après avoir vu que le résultat devient significatif ; on essaie le test t, puis Mann-Whitney, puis la transformation logarithmique ; on répartit les patients en sous-groupes (par sexe, par âge, par milieu de résidence) jusqu'à ce qu'apparaisse un sous-groupe présentant un effet ; on remplace le critère principal par un critère secondaire. Chaque étape paraît raisonnable prise isolément, mais ensemble elles transforment cinq tentatives en une seule présentation "significative", avec un risque réel de faux positif bien supérieur à 5%.

Le HARKing (en anglais hypothesizing after the results are known) consiste à formuler l'hypothèse après avoir vu les résultats et à la présenter comme si elle avait été établie à l'avance. Une association observée par hasard entre le tabagisme et une variable quelconque devient, dans l'introduction de l'article, "notre hypothèse principale". Le lecteur ne peut plus distinguer la confirmation de l'exploration.

L'antidote est le préenregistrement : le protocole de l'étude, avec l'hypothèse principale, le critère principal, la taille de l'échantillon et le plan d'analyse statistique, est publié dans un registre public avant l'inclusion du premier patient. Pour les essais cliniques randomisés, l'enregistrement est obligatoire, et la recommandation CONSORT exige que tout écart au protocole soit rapporté [56]. Pour les études observationnelles, la recommandation STROBE exige la description de toutes les analyses effectuées, et pas seulement de celles qui sont significatives [25]. À votre niveau, la règle pratique est la suivante : écrivez les hypothèses et le test choisi (points a-c du format du devoir) avant d'ouvrir le fichier de données dans JASP.

Toute analyse non planifiée n'est pas fautive. L'exploration des données génère de nouvelles hypothèses ; le problème n'apparaît que lorsque l'exploration est présentée comme une confirmation. La formulation honnête est "dans des analyses exploratoires, non planifiées, nous avons observé..." et elle avertit le lecteur que le résultat doit être vérifié dans une nouvelle étude. Les analyses de sensibilité sont un exemple de bonne pratique : les auteurs répètent l'analyse dans des conditions différentes pour voir si le résultat résiste. Dans l'étude transversale sur les troubles de l'interaction intestin-cerveau, portant sur 136 participants analysés (52 cas et 84 témoins), l'IMC était plus élevé chez les cas (28,15 ± 6,49 contre $24{,}47 \pm \frac{4{,}60\ \text{kg}}{m^2}$, $p = 0{,}001$), et un facteur alimentaire présentait un OR protecteur de 0,36 (IC à 95% 0,18-0,74 ; p = 0,009) ; les auteurs ont répété l'analyse après exclusion des participants dont la protéine C réactive était supérieure à $\frac{10\text{mg}}{L}(n = 98)$ [71]. Un résultat qui se maintient dans l'analyse de sensibilité est plus crédible qu'un résultat qui n'apparaît que dans une seule variante d'analyse.

Étude de cas : la comparaison intragroupe face à la comparaison intergroupes

Un essai clinique contrôlé contre placebo, rapporté selon la recommandation CONSORT, a évalué pendant 3 mois l'effet de l'extrait d'Aloe ferox (2 gélules/jour, 460 mg) associé à un régime, chez 20 personnes obèses, par rapport à 20 témoins appariés, traités par régime et placebo [85]. Le résumé rapporte deux types de résultats : des comparaisons entre les groupes à 3 mois et des comparaisons avant-après à l'intérieur des groupes, avec des valeurs p telles que 0,03 et 0,002 pour l'IMC, 0,01 et 0,039 pour le LDL-cholestérol, 0,018 et moins de 0,001 pour la glycémie à jeun, 0,032 pour le cholestérol total et 0,006 pour la pression diastolique. L'exercice de lecture critique ne vise pas cette étude en particulier, mais une erreur classique de la littérature clinique, que tout lecteur doit savoir éviter. Supposons une étude dans laquelle la glycémie baisse significativement dans le groupe traité (p = 0,01 avant-après) et de façon non significative dans le groupe placebo (p = 0,08). La conclusion "le traitement est efficace, et le placebo ne l'est pas" est fausse : deux résultats situés de part et d'autre du seuil ne démontrent pas que les effets diffèrent entre les groupes. La baisse dans le groupe placebo peut être presque aussi importante, seulement estimée de façon moins précise. Le seul test qui répond à la question de l'efficacité est la comparaison directe entre les groupes : soit la comparaison des baisses (les différences individuelles) entre les groupes avec le test t pour échantillons indépendants, soit la comparaison des valeurs finales ajustées sur la valeur initiale (ANCOVA, chapitre 6). Les comparaisons intragroupes décrivent seulement l'évolution de chaque groupe et sont influencées par la régression vers la moyenne et par l'effet placebo. Lorsque vous lisez une étude contrôlée, cherchez d'abord la comparaison entre les groupes ; dans l'étude sur Aloe ferox, le résumé la rapporte, et le lecteur doit seulement savoir sur laquelle des valeurs p

> appuyer sa conclusion sur l'efficacité. Avec 20 patients par groupe, les intervalles de confiance des différences auraient utilement complété les valeurs p.

5.12 Équivalence et non-infériorité, en bref

Le test classique peut rejeter l'hypothèse "aucune différence", mais il ne peut pas la confirmer. Or c'est parfois exactement ce que nous voulons montrer : qu'un médicament générique a le même effet que le princeps, qu'une méthode de mesure bon marché donne les mêmes résultats qu'une méthode coûteuse, ou qu'un traitement plus commode n'est pas moins efficace que le traitement standard. Pour ce type de questions, les hypothèses s'inversent.

Dans une étude d'équivalence, le chercheur fixe à l'avance une marge d'équivalence Δ (par exemple 5 mmHg pour la PAS), c'est-à-dire la plus grande différence considérée comme sans importance clinique. H0 devient "la différence est au moins égale à Δ en valeur absolue" (les traitements ne sont pas équivalents), et H1 "la différence est inférieure à Δ" (les traitements sont équivalents). En pratique, l'équivalence est démontrée si l'intervalle de confiance entier de la différence se situe entre -Δ et +Δ ; la procédure s'appelle TOST (deux tests unilatéraux). Dans une étude de non-infériorité, un seul sens nous intéresse : le nouveau traitement n'est pas moins efficace que le traitement standard de plus de Δ ; c'est le cas rare où le test unilatéral est justifié.

L'exemple de la PAS selon le sexe est instructif. L'IC à 95% de la différence va de -4,9 à 3,8 mmHg. Si la marge d'équivalence fixée à l'avance avait été de 5 mmHg, l'intervalle se situerait entièrement dans la zone de -5 à +5, et nous pourrions affirmer l'équivalence des deux moyennes. Si la marge avait été de 3 mmHg, nous ne pourrions affirmer ni la différence ni l'équivalence. Pour un calcul rigoureux, on utilise généralement un IC à 90%, correspondant aux deux tests unilatéraux à α = 0,05 ; l'idée de base reste cependant la même : l'équivalence se démontre avec l'intervalle de confiance, jamais avec un p > 0,05.

5.13 Comment rapporter un test statistique

La présentation d'un test dans un article, un mémoire de licence ou un devoir doit permettre au lecteur de refaire le raisonnement. Les normes actuelles, contenues dans les recommandations de l'ICMJE et dans les recommandations CONSORT et STROBE [25,36,56], exigent quelques éléments obligatoires : les statistiques descriptives de chaque groupe (moyenne et ÉT, ou médiane et EIQ, avec n) ; le nom du test, sa variante (par exemple t de Welch, bilatéral) et le seuil α ; la valeur de la statistique de test, avec les degrés de liberté ; la valeur p exacte ; la taille de l'effet, avec son intervalle de confiance à 95% ; l'interprétation clinique, y compris le seuil de pertinence, s'il existe.

Un exemple de présentation complète pour mg2026 : "La glycémie à jeun est passée de 105,4 mg/dL à l'inclusion à 95,2 mg/dL à 6 mois (différence moyenne 10,19 mg/dL, IC à 95% 8,38-11,99 ; test t pour échantillons appariés, t(159) = 11,14, p < 0,001 ; dz = 0,88). La distribution des différences individuelles ne s'écartait pas significativement de la normalité (Shapiro-Wilk p = 0,725)."

Exemple de présentation d'un résultat non significatif : "La PAS ne différait pas significativement entre les femmes ($131{,}37 \pm 14{,}47$ mmHg, $n = 86$) et les hommes ($130{,}85 \pm 13{,}25$ mmHg, $n = 74$) ; la différence hommes moins femmes était de -0,5 mmHg (IC à 95% -4,9 à 3,8 ; test t de Welch, $t(157{,}4) = -0{,}24$, $p = 0{,}81$; $d = -0{,}04$)." Remarquez que la phrase ne dit pas "la PAS est égale chez les femmes et les hommes" ; l'intervalle de confiance laisse le lecteur juger quelles différences sont exclues.

Quelques formulations à éviter : "p = 0,000" (on écrit p < 0,001) ; "NS" sans aucun chiffre ; "tendance à la significativité" pour p = 0,07 (soit nous interprétons l'intervalle de confiance, soit

nous disons que le résultat est non significatif) ; "différence extrêmement significative" sans la taille de la différence ; "les résultats démontrent que..." dans une étude observationnelle.

L'étude rétrospective sur l'évolution de la COVID-19 chez les patients atteints de diabète de type 1 et de type 2, portant sur 486 patients (200 sans diabète, 62 avec un diabète de type 1, 224 avec un diabète de type 2), est un bon exercice de lecture [8]. Le résumé rapporte des proportions claires (formes sévères 33,5% pour le type 2 contre 25,8% pour le type 1 ; mortalité 11,6% contre 8,1%) et des moyennes avec ÉT (durée d'hospitalisation 11,6 ± 7,0 contre 9,1 ± 5,8 jours ; glycémie à l'admission 178,3 ± 34,7 contre 164,8 ± 39,6 mg/dL), mais les rapports de cotes de l'admission en soins intensifs (OR 2,24) et de la ventilation mécanique (OR 2,46) apparaissent sans intervalle de confiance et sans valeur p. Sans intervalle, le lecteur ne peut pas juger si un OR de 2,24 est estimé avec précision (par exemple 1,8-2,8) ou de façon très imprécise (par exemple 0,9-5,6). De plus, les écarts-types importants par rapport à la moyenne de la durée d'hospitalisation suggèrent une distribution asymétrique, pour laquelle la médiane et l'EIQ auraient été plus informatives. Cette observation n'enlève rien à la valeur clinique de l'étude ; elle montre seulement à quel point quelques chiffres de plus dans le résumé comptent pour le lecteur.

5.14 Problèmes résolus

Problème résolu 5.1. De la question aux hypothèses

Énoncé. Pour chacune des questions ci-dessous, formulez l'hypothèse de recherche, les hypothèses statistiques en mots et en notation mathématique, et précisez si le test est bilatéral ou unilatéral. (a) Dans mg2026, la fréquence cardiaque de repos diffère-t-elle entre fumeurs et non-fumeurs ? (b) La proportion de patients ayant un diabète confirmé par HGPO est-elle différente de 10%, une prévalence hypothétique dans la population adulte générale ? (c) Un fabricant affirme qu'une nouvelle bandelette de mesure de la glycémie ne sous-estime pas les valeurs de plus de 3 mg/dL par rapport au laboratoire de référence. (d) La PAS change-t-elle entre l'inclusion et 6 mois chez les mêmes patients ?

Solution. (a) Hypothèse de recherche : la fréquence cardiaque de repos des fumeurs diffère de celle des non-fumeurs. H0 : dans la population, la moyenne de la FC est la même chez les fumeurs et chez les non-fumeurs, $\mu_{\text{fumeurs}} = \mu_{\text{non-fumeurs}}$. H1 : les moyennes diffèrent, $\mu_{\text{fumeurs}} \neq \mu_{\text{non-fumeurs}}$. Test bilatéral : une différence dans un sens comme dans l'autre serait intéressante. (b) Hypothèse de recherche : la prévalence du diabète dans la population à risque cardiométabolique diffère de celle de la population générale. H0 : $\pi = 0{,}10$; H1 : $\pi \neq 0{,}10$. Test bilatéral. (c) Il s'agit d'une question de non-infériorité : H0 : la bandelette sous-estime d'au moins 3 mg/dL (différence moyenne bandelette moins laboratoire ≤ -3) ; H1 : la sous-estimation moyenne est inférieure à 3 mg/dL (différence > -3). Test unilatéral, justifié par la nature de la question, avec une marge fixée à l'avance. (d) Hypothèse de recherche : au cours des 6 mois d'intervention, la PAS change. H0 : la moyenne des différences individuelles (PAS0 - PAS6) est nulle dans la population, $\mu_D = 0$; H1 : $\mu_D \neq 0$. Même si nous nous attendons à une baisse, le test reste bilatéral : une augmentation de la PAS serait un résultat important (par exemple un effet indésirable), qui ne doit pas être ignoré. Dans mg2026, le test apparié montre une baisse moyenne de 4,78 mmHg (IC à 95% 3,36-6,19 ; $t = 6{,}68$; $p < 0{,}001$).

Interprétation. Dans les quatre cas, H0 contient le signe égal (ou, pour la non-infériorité, "égal ou pire"), c'est-à-dire l'affirmation précise qui permet de calculer la distribution de la statistique de test.

Problème résolu 5.2. La glycémie à jeun comparée à 100 mg/dL

Énoncé. Dans mg2026, la glycémie à jeun à l'inclusion a une moyenne de 105,44 mg/dL, un écart-type de 15,76 mg/dL, n = 160. Le seuil de l'ADA pour l'hyperglycémie modérée à jeun est de 100 mg/dL [80]. La moyenne de la population des patients à risque cardiométabolique diffère-t-elle de 100 mg/dL ? Utilisez α = 0,05 et les trois voies de décision.

Solution. Hypothèse de recherche : la moyenne de la glycémie à jeun dans la population à risque cardiométabolique diffère du seuil de 100 mg/dL. H0 : $\mu = 100$ mg/dL ; H1 : μ ≠ 100 mg/dL ; bilatéral, α = 0,05. Le test : une moyenne comparée à une valeur donnée, σ inconnu, donc le test t pour une seule moyenne. Conditions : observations indépendantes ; Shapiro-Wilk p = 0,037 indique un léger écart à la normalité (asymétrie 0,37), mais avec n = 160 la moyenne a une distribution pratiquement normale, et le test t est robuste ; nous vérifierons néanmoins avec le test de Wilcoxon.

Calcul. ES $= \frac{15,76}{\sqrt{160}} = \frac{15,76}{12,649} = 1,246$ mg/dL. $t = \frac{105,44-100}{1,246} = \frac{5,44}{1,246} = 4,36$, avec ddl = 159.

La voie de la valeur critique : t critique (159 ; 0,05 bilatéral) = 1,975 ; 4,36 > 1,975 ; nous rejetons donc H0. La voie de la valeur p : p = T.DIST.2T(4,36;159) = 0,00002, c'est-à-dire p < 0,001. La voie de l'IC : $\text{IC}\,95\% = 105,44 \pm 1,975 \times 1,246 = 105,44 \pm 2,46$, c'est-à-dire 102,98-107,90 mg/dL ; la valeur 100 ne se trouve pas dans l'intervalle ; nous rejetons donc H0. Taille de l'effet : $d = \frac{5,44}{15,76} = 0,35$ (effet petit à moyen). Le test de Wilcoxon pour la médiane comparée à 100 donne p < 0,001, même conclusion.

Conclusion statistique : p < 0,001, on rejette H0 ; la différence est extrêmement significative. Conclusion de recherche : la moyenne de la glycémie à jeun dans la population d'adultes à risque cardiométabolique dont provient l'échantillon est supérieure au seuil de 100 mg/dL, d'environ 3 à 8 mg/dL (IC à 95% de la moyenne 102,98-107,90 mg/dL). Interprétation clinique : le patient "moyen" de l'étude se situe dans la zone de l'hyperglycémie modérée à jeun, ce qui correspond au critère d'inclusion (risque cardiométabolique). Si nous testions la valeur de 105 mg/dL, nous obtiendrions t = 0,35 et p = 0,73 : les données sont compatibles avec une moyenne de 105 mg/dL, ce qui ne signifie pas que la moyenne est exactement de 105.

Problème résolu 5.3. Le test z et le test t pour le cholestérol

Énoncé. Un registre national indique pour les adultes de 35 à 79 ans une moyenne du cholestérol total de 200 mg/dL, avec $\sigma = 40$ mg/dL (valeurs hypothétiques). Dans mg2026, n = 160, $m = 209,39$ mg/dL, $s = 36,91$ mg/dL. (a) Appliquez le test z. (b) Appliquez le test t. (c) Commentez la différence.

Solution. H0 : $\mu = 200$ mg/dL ; H1 : μ ≠ 200 mg/dL ; α = 0,05, bilatéral. (a) Avec σ connu : ES $= \frac{40}{\sqrt{160}} = 3,162$; $z = \frac{9,39}{3,162} = 2,97$. La valeur critique 1,96 est dépassée, et 2,97 > 2,576 dépasse aussi le seuil pour α = 0,01. p = 2 × (1 - Φ(2,97)) = 0,003. On rejette H0 ; la différence est très significative. (b) Avec σ estimé par s : ES $= \frac{36,91}{12,649} = 2,918$; $t = \frac{9,39}{2,918} = 3,22$, ddl = 159 ; p = 0,0016, également très significatif. (c) Les deux tests ne diffèrent que par l'estimation du bruit. Ici, s (36,91) est inférieur au σ du registre (40) ; c'est pourquoi t est un peu plus grand que z. Si le σ du registre décrit correctement notre population également, le test z est le test approprié ; si nos patients sont plus homogènes que la population générale, s reflète mieux leur variabilité et le test t est préférable. Conclusion de recherche : les patients à risque cardiométabolique ont un cholestérol moyen supérieur à celui de la population de référence, d'environ 9 mg/dL (IC à 95% de la moyenne, calculé avec t : 203,6-215,2 mg/dL).

Problème résolu 5.4. Un résultat non significatif sur un petit échantillon

Énoncé. Un étudiant n'a eu accès qu'à 10 patients de mg2026, avec les PAS suivantes : 150, 154, 131, 130, 137, 102, 110, 103, 147, 136 mmHg. Testez si la moyenne diffère de 125 mmHg et interprétez le résultat au regard de l'analyse portant sur l'ensemble des 160 patients.

Solution. H0 : μ = 125 ; H1 : μ ≠ 125 ; α = 0,05. La somme des valeurs est de 1300, donc m = 130,0 mmHg. Les écarts à la moyenne sont 20, 24, 1, 0, 7, -28, -20, -27, 17, 6 ; leurs carrés sont 400, 576, 1, 0, 49, 784, 400, 729, 289, 36, de somme 3264. La variance $s^2 = \frac{3264}{9} = 362{,}67$, donc s = 19,04 mmHg. ES $= \frac{19{,}04}{\sqrt{10}} = \frac{19{,}04}{3{,}162} = 6{,}02$ mmHg. $t = \frac{130{,}0-125}{6{,}02} = 0{,}83$, ddl = 9. La valeur critique pour ddl = 9 est de 2,262 ; 0,83 < 2,262 ; nous ne rejetons donc pas H0. La valeur p = T.DIST.2T(0,83;9) = 0,43. IC $95\% = 130{,}0 \pm 2{,}262 \times 6{,}02 = 130{,}0 \pm 13{,}62$, c'est-à-dire 116,38-143,62 mmHg, un intervalle qui contient 125.

Interprétation. Conclusion statistique : p = 0,43 > 0,05, on ne rejette pas H0 ; la différence est statistiquement non significative. La conclusion de recherche ne peut pas être "la moyenne de la PAS est de 125 mmHg" : l'intervalle de confiance couvre des valeurs de 116 à 144 mmHg, c'est-à-dire d'une pression optimale jusqu'à une hypertension de grade 1. Le résultat est indéterminé, et non négatif. Sur l'ensemble des 160 patients, avec une ES = 1,097 (5,5 fois plus petite), la même question reçoit une réponse claire : t = 5,59, p < 0,001. La différence entre les deux analyses ne vient pas de données contradictoires, mais de la puissance du test. Pour un effet de la taille de celui observé sur l'ensemble de l'échantillon (d ≈ 0,44), la puissance d'un test t sur 10 patients est très faible, de l'ordre de 20 à 25% ; l'échec du rejet de H0 était prévisible.

Problème résolu 5.5. Bonferroni, Holm et Benjamini-Hochberg

Énoncé. Une étude exploratoire compare deux groupes sur six critères et obtient les valeurs p suivantes : 0,012 ; 0,001 ; 0,20 ; 0,041 ; 0,008 ; 0,030. Combien de résultats sont significatifs à α = 0,05 (a) sans correction, (b) avec Bonferroni, (c) avec Holm, (d) avec Benjamini-Hochberg ?

Solution. Nous rangeons les valeurs par ordre croissant : p(1) = 0,001 ; p(2) = 0,008 ; p(3) = 0,012 ; p(4) = 0,030 ; p(5) = 0,041 ; p(6) = 0,20 ; k = 6.

(a) Sans correction, toutes les valeurs inférieures à 0,05 sont significatives : cinq résultats. La probabilité d'avoir au moins un faux positif, si toutes les H0 étaient vraies, ar fi $1 - 0{,}95^6 = 0{,}265$.

(b) Bonferroni : le seuil est de 0,05 / 6 = 0,0083. Seules 0,001 et 0,008 le franchissent : deux résultats. Les valeurs p ajustées (p × 6) sont 0,006 ; 0,048 ; 0,072 ; 0,18 ; 0,246 ; 1.

(c) Holm : nous comparons p(1) = 0,001 à 0,05/6 = 0,0083 (franchit le seuil) ; p(2) = 0,008 à 0,05/5 = 0,010 (franchit) ; p(3) = 0,012 à 0,05/4 = 0,0125 (franchit) ; p(4) = 0,030 à 0,05/3 = 0,0167 (ne franchit pas). Nous nous arrêtons : trois résultats significatifs.

(d) Benjamini-Hochberg : les seuils (i/6) × 0,05 sont 0,0083 ; 0,0167 ; 0,025 ; 0,0333 ; 0,0417 ; 0,05. Nous comparons : 0,001 ≤ 0,0083 ; 0,008 ≤ 0,0167 ; 0,012 ≤ 0,025 ; 0,030 ≤ 0,0333 ; 0,041 ≤ 0,0417 ; 0,20 > 0,05. Le plus grand i pour lequel p(i) ≤ seuil est i = 5 ; cinq résultats sont donc significatifs.

Interprétation. Les méthodes répondent à des questions différentes. Bonferroni et Holm garantissent que la probabilité d'avoir au moins un faux positif dans l'ensemble ne dépasse pas 5% ; Holm est toujours au moins aussi puissant que Bonferroni et devrait être préféré. BH accepte qu'en moyenne au plus 5% des résultats déclarés significatifs soient faux ; il convient aux analyses exploratoires comportant de nombreux tests. Le choix de la méthode se fait avant l'analyse et se rapporte dans la section des méthodes.

Problème résolu 5.6. La puissance et la taille de l'échantillon

Énoncé. Un médecin généraliste veut comparer deux programmes d'éducation sur l'hypertension, au moyen d'un essai randomisé à deux groupes. Il considère comme cliniquement pertinente une différence de PAS moyenne de 5 mmHg ; d'après mg2026, il sait que l'écart-type de la PAS est d'environ 14 mmHg. (a) Quelle taille d'effet cela représente-t-il ? (b) Combien de patients faut-il dans chaque groupe pour une puissance de 80% à α = 0,05 en bilatéral ? (c) Quelle puissance aurait une étude de 53 patients par groupe, comme les bras de mg2026 ?

Solution. $(a)\, d = \frac{\Delta}{\sigma} = \frac{5}{14} = 0{,}36$, un effet petit à moyen. (b) La formule approchée pour la comparaison de deux moyennes est n par groupe = 2 × (z1-α/2 + z1-β)² × σ² / Δ². Pour α = 0,05 en bilatéral, $z_1 - \frac{\alpha}{2} = 1{,}96$; pour une puissance de 80%, z1-β = 0,84. On obtient $n = \frac{2\times(1{,}96+0{,}84)^2\times 14^2}{5^2} = \frac{2\times 7{,}84\times 196}{25} = 122{,}9$, soit 123 patients par groupe ; le calcul exact avec la loi t (G*Power, JASP) donne 125. On y ajoute généralement 10 à 20% pour les perdus de vue. (c) Avec 53 patients par groupe, la puissance pour d = 0,36 n'est que d'environ 45% : l'étude manquerait un effet réel de 5 mmHg plus souvent qu'elle ne le détecterait.

Interprétation. Le calcul de la taille de l'échantillon se fait avant l'étude et part de la différence cliniquement pertinente, et non de la différence que nous espérons trouver. Une étude de 53 patients par bras suffit pour de grands effets (la baisse de la glycémie dans mg2026 diffère entre les bras avec des effets standardisés compris entre environ 0,5 et 1,4), mais elle est trop petite pour une différence de 5 mmHg de la PAS.

Problème résolu 5.7. La valeur p à partir de la statistique t

Énoncé. Sans logiciel statistique, uniquement avec le tableau 5.6, décidez à α = 0,05 : (a) t = 2,10, ddl = 20, bilatéral ; (b) t = 1,80, ddl = 25, bilatéral, puis unilatéral (dans le sens annoncé à l'avance) ; (c) t = 3,10, ddl = 9, bilatéral. Vérifiez ensuite les valeurs p exactes avec Google Sheets.

Solution. (a) La valeur critique pour ddl = 20 est de 2,086 ; 2,10 > 2,086 ; p est donc légèrement inférieur à 0,05 ; on rejette H0, à la limite. Dans Sheets, =T.DIST.2T(2,1;20) donne 0,049. (b) La valeur critique bilatérale pour ddl = 25 est de 2,060 ; 1,80 < 2,060 ; donc p > 0,05 et on ne rejette pas H0 (p bilatéral = T.DIST.2T(1,8;25) = 0,084). En unilatéral, la valeur critique est d'environ 1,708 (=T.INV(0,95;25)) ; 1,80 > 1,708 ; donc p unilatéral = 0,042 < 0,05. Le même résultat est "significatif" en unilatéral et "non significatif" en bilatéral ; c'est précisément pour cette raison que le choix doit être fait et justifié avant l'analyse. (c) Pour ddl = 9, les valeurs critiques sont 2,262 (α = 0,05) et 3,250 (α = 0,01) ; 3,10 se situe entre les deux, donc 0,01 < p < 0,05 ; dans Sheets, p = 0,013. Le résultat est significatif, mais pas très significatif.

5.15 Exercices proposés

1. Pour l'HbA1c à l'inclusion, mg2026 a m = 5,58%, s = 0,45%, n = 160. Testez si la moyenne de la population diffère de 5,7% (seuil inférieur de l'intervalle du prédiabète selon l'ADA). Calculez t, décidez à α = 0,05 et formulez la conclusion de recherche.
2. Le score FINDRISC dans mg2026 a m = 10,37, s = 4,27, n = 160. La moyenne diffère-t-elle de 12 (seuil choisi pour l'exercice) ? Calculez t, l'IC à 95% de la moyenne et d.
3. Fréquence cardiaque de repos à l'inclusion : m = 74,71 battements/min, s = 9,30, n = 160. Testez H0 : μ = 72 battements/min. Calculez t et la taille de l'effet, et commentez la pertinence clinique de la différence.

4. Dans un groupe de 36 enfants d'une communauté rurale, l'hémoglobine moyenne est de 11,2 g/dL. De grandes études ont établi qu'à cet âge $\sigma = 1{,}2$ g/dL, et la moyenne de la population générale est de 11,8 g/dL. Appliquez le test z (α = 0,05, bilatéral) et calculez la valeur p.
5. Classez chaque situation comme erreur de type I, erreur de type II ou décision correcte : (a) une étude conclut qu'un complément alimentaire abaisse le cholestérol, alors que ce complément n'a aucun effet ; (b) une étude avec 15 patients par groupe ne trouve pas de différence significative entre un nouvel antibiotique, en réalité supérieur, et l'antibiotique standard ; (c) une grande étude ne trouve pas de différence entre deux régimes qui ont, en réalité, le même effet.
6. Un étudiant effectue 8 tests indépendants à α = 0,05, et toutes les hypothèses nulles sont vraies. (a) Quelle est la probabilité d'obtenir au moins un résultat faux positif ? (b) Quel seuil individuel imposerait la correction de Bonferroni ? (c) Et pour 4 tests ?
7. Cinq comparaisons ont donné les valeurs p suivantes : 0,002 ; 0,009 ; 0,018 ; 0,035 ; 0,25. Combien sont significatives à α = 0,05 sans correction, avec Bonferroni, avec Holm et avec Benjamini-Hochberg ?
8. À l'aide du tableau 5.6, puis de Google Sheets, trouvez la valeur p bilatérale pour : (a) t = 2,5 avec ddl = 12 ; (b) t = 1,70 avec ddl = 40 (donnez aussi la valeur unilatérale) ; (c) t = -2,9 avec ddl = 159.
9. Une étude rapporte pour la différence de PAS entre deux traitements : différence moyenne de 4 mmHg, IC à 95% de -1 à 9 mmHg. (a) p est-il inférieur ou supérieur à 0,05 ? (b) La différence minimale cliniquement importante est de 5 mmHg. Le résultat correspond-il à un "effet absent" ou à un "effet indéterminé" ? (c) Que recommanderiez-vous pour une étude future ?
10. Pour l'IMC à l'inclusion, $m = \frac{29{,}04\ \text{kg}}{m^2}$, s = 3,91, n = 160, asymétrie 1,42, Shapiro-Wilk p < 0,001. (a) Calculez t pour H0 : $\mu = \frac{30\ \text{kg}}{m^2}$. (b) Quel test préféreriez-vous et pourquoi ? (c) Comment rapporteriez-vous le résultat ?
11. Combien de patients par groupe faut-il pour détecter, avec une puissance de 80% et α = 0,05 en bilatéral, une différence de PAS de 10 mmHg entre deux groupes, si σ = 14 mmHg ? Utilisez la formule approchée du problème résolu 5.6.

Pièges fréquents

- Les hypothèses statistiques s'écrivent avec les paramètres de la population (μ, π, ρ), et non avec les moyennes de l'échantillon (m1, m2).
- "On accepte H0" est une formulation erronée ; un résultat non significatif signifie que les données ne permettent pas de rejeter H0.
- La valeur p n'est ni la probabilité que H0 soit vraie ni la probabilité que le résultat soit dû au hasard.
- "Extrêmement significatif" (p < 0,001) décrit la force des éléments contre H0, et non la taille ou l'importance clinique de l'effet.
- Choisir le test unilatéral après avoir vu le sens de la différence divise artificiellement la valeur p par deux.
- Un effet significatif dans un groupe et non significatif dans un autre ne prouve pas que les effets diffèrent entre les groupes ; il faut comparer directement les groupes.
- Avec de nombreux tests, quelques résultats "significatifs" apparaissent par simple hasard ; sans correction ou sans critère principal fixé à l'avance, ils restent des hypothèses.
- On écrit p < 0,001, jamais p = 0,000.

À retenir

- Le test statistique part de H0 (absence d'effet), calcule à quel point les données seraient inhabituelles si H0 était vraie et rejette H0 lorsque $p < \alpha$.
- α = probabilité de l'erreur de type I (faux positif) ; β = probabilité de l'erreur de type II (faux négatif) ; $1 - \beta$ = puissance ; $1 - \alpha$ = niveau de confiance.
- La puissance augmente avec la taille de l'effet, avec la taille de l'échantillon et avec α, et diminue avec la variabilité ; lors de la planification, on vise généralement une puissance d'au moins 80%.
- La valeur critique, la valeur p et l'intervalle de confiance conduisent à la même décision ; l'intervalle de confiance apporte le plus d'information.
- Le test t pour une moyenne : $t = \frac{m-\mu_0}{\frac{s}{\sqrt{n}}}$, avec n - 1 degrés de liberté ; le test z utilise σ connu.
- Un résultat rapporté de façon complète comprend les statistiques descriptives, le test, la statistique avec ses degrés de liberté, le p exact, la taille de l'effet avec son IC à 95% et l'interprétation clinique.
- $1 - 0{,}95^k$ donne le risque d'au moins un faux positif pour k tests ; Bonferroni (α/k), Holm et Benjamini-Hochberg sont les corrections usuelles.

Questions d'autoévaluation

1. (QRU) La valeur p d'un test statistique représente :
    a) la probabilité que l'hypothèse nulle soit vraie
    b) la probabilité que l'hypothèse alternative soit vraie
    c) la probabilité d'obtenir un résultat au moins aussi extrême que celui observé, si l'hypothèse nulle est vraie
    d) la probabilité de commettre une erreur de type II
    e) la taille d'effet standardisée
2. (QRU) Une étude avec 12 patients par groupe ne trouve pas de différence significative entre un nouveau médicament et le placebo (p = 0,24), alors que le médicament est, en réalité, efficace. L'erreur commise est :
    a) une erreur de type I
    b) une erreur de type II
    c) une erreur d'échantillonnage systématique
    d) une erreur de mesure
    e) ce n'est pas une erreur, puisque p > 0,05
3. (QRU) Pour la moyenne de la PAS de mg2026 (m = 131,13 mmHg, IC à 95% 128,96-133,30 mmHg), le test bilatéral à $\alpha = 0{,}05$ de l'hypothèse H0 : $\mu = 132$ mmHg aura pour résultat :
    a) le rejet de H0, puisque $132 \neq 131{,}13$
    b) le non-rejet de H0, puisque 132 se trouve dans l'IC à 95%
    c) l'acceptation de H0, puisque la moyenne de la population est certainement de 132 mmHg
    d) le rejet de H0, puisque n = 160 est un grand échantillon
    e) la décision ne peut pas être prise sans la valeur p
4. (QRU) La puissance d'un test statistique est :
    a) α
    b) $1 - \alpha$
    c) β
    d) $1 - \beta$
    e) $\alpha + \beta$

5. (QRU) Un chercheur effectue 10 tests indépendants à $\alpha = 0,05$, et toutes les hypothèses nulles sont vraies. La probabilité d'obtenir au moins un résultat significatif est d'environ :
    a) 0,05
    b) 0,10
    c) 0,23
    d) 0,40
    e) 0,50
6. (QRM) À propos des hypothèses statistiques, les affirmations correctes sont :
    a) l'hypothèse nulle affirme généralement l'absence de différence ou d'association
    b) les hypothèses sont formulées sur les paramètres de la population
    c) l'hypothèse alternative bilatérale pour deux moyennes s'écrit $\mu 1 \neq \mu 2$
    d) les hypothèses sont formulées après l'examen des données, afin de choisir le sens du test
    e) H0 et H1 s'excluent mutuellement
7. (QRM) La puissance d'un test t pour deux groupes indépendants augmente si :
    a) le nombre de patients de chaque groupe augmente
    b) la taille réelle de l'effet augmente
    c) l'écart-type de la variable augmente
    d) on passe de $\alpha = 0,01$ à $\alpha = 0,05$
    e) on passe d'un test unilatéral justifié à un test bilatéral
8. (QRM) Dans mg2026, la PAS ne diffère pas significativement entre les sexes : différence de -0,5 mmHg, IC à 95% de -4,9 à 3,8, $p = 0,81$. Les affirmations correctes sont :
    a) on accepte H0, donc la PAS moyenne est identique chez les femmes et chez les hommes
    b) on ne rejette pas H0
    c) l'intervalle de confiance exclut des différences moyennes supérieures à environ 5 mmHg
    d) le résultat démontre que le sexe n'influence la PAS dans aucune population
    e) la taille de l'effet ($d = -0,04$) est pratiquement nulle
9. (QRM) Sont des exemples de pratiques qui augmentent le risque réel de résultats faux positifs :
    a) l'ajout de patients jusqu'à ce que p descende sous 0,05
    b) le choix du test unilatéral après avoir vu le sens de la différence
    c) le préenregistrement du protocole avec le critère principal
    d) la présentation des seuls sous-groupes dans lesquels l'effet a été significatif
    e) l'application de la correction de Holm pour les comparaisons multiples
10. (QRM) À propos de la signification statistique et de la signification clinique, les affirmations correctes sont :

a) dans un très grand échantillon, une différence sans importance clinique peut être extrêmement significative sur le plan statistique

b) une corrélation $r = -0,088$ avec $p = 0,01$ sur 867 patients explique moins de 1% de la variabilité

c) un $p < 0,001$ garantit que l'effet est important pour le patient

d) la différence minimale cliniquement importante est fixée selon des critères cliniques, avant l'analyse

e) la taille de l'effet et l'intervalle de confiance complètent la valeur p dans la présentation des résultats

Chapitre 6. Tests statistiques paramétriques usuels : test t pour échantillons appariés et indépendants, analyse de la variance (ANOVA)

Objectifs

À la fin de ce chapitre, l'étudiant sera capable de :

- reconnaître le schéma d'une étude (une moyenne comparée à une valeur, deux mesures chez les mêmes patients, deux groupes indépendants, trois groupes ou plus) et choisir le test paramétrique adapté ;
- appliquer et calculer à la main le test t pour échantillons appariés, y compris l'intervalle de confiance de la différence moyenne et la taille d'effet dz ;
- appliquer le test t pour échantillons indépendants dans sa version de Student (variance commune) et dans sa version de Welch, interpréter le test de Levene et calculer le d de Cohen ;
- expliquer la logique de l'analyse de la variance (variation entre les groupes comparée à la variation à l'intérieur des groupes) et construire à la main le tableau d'ANOVA (SS, ddl, MS, F) ;
- vérifier les conditions d'application (indépendance, normalité, homogénéité des variances) et choisir l'alternative lorsqu'elles ne sont pas remplies (Welch, Games-Howell, tests non paramétriques) ;
- choisir et interpréter les tests post hoc (Tukey, Bonferroni, Games-Howell, Dunnett) et expliquer pourquoi on ne fait pas de tests t multiples ;
- décrire l'ANOVA à mesures répétées (sphéricité, correction de Greenhouse-Geisser), le principe de l'ANOVA à deux facteurs et le rôle de l'ANCOVA dans l'ajustement sur la valeur initiale ;
- lire de façon critique les résultats des tests t et des ANOVA dans les articles publiés et rapporter correctement ses propres résultats.

6.1 La famille des tests paramétriques

Les tests de ce chapitre sont dits paramétriques parce qu'ils font des hypothèses sur la forme de la distribution dans la population (le plus souvent, la distribution normale) et comparent les paramètres de celle-ci, à savoir les moyennes. En contrepartie de ces hypothèses, ils utilisent toute l'information contenue dans les données et ont, lorsque les conditions sont remplies, la plus grande puissance. Les tests non paramétriques du chapitre 7 (Mann-Whitney, Wilcoxon, Kruskal-Wallis, Friedman) font moins d'hypothèses et travaillent sur les rangs des observations ; ils constituent l'alternative naturelle lorsque la distribution est fortement asymétrique, que l'échantillon est petit ou que la variable est ordinale [19,63].

Le choix du test paramétrique dépend avant tout du schéma de l'étude. Le tableau 6.1 résume les situations que nous allons étudier ; l'algorithme complet de choix du test, y compris pour les variables qualitatives, figure au chapitre 8.

Tableau 6.1. Tests paramétriques pour la comparaison de moyennes et leurs alternatives non paramétriques.

Situation	Exemple dans mg2026	Test paramétrique	Alternative non paramétrique
une moyenne comparée à une valeur donnée	moyenne de la PAS comparée à 125 mmHg	test z ou t pour une seule moyenne	Wilcoxon pour une médiane
deux mesures chez les	glycémie à 0 et à 6 mois	test t pour échantillons	Wilcoxon des rangs signés

Situation	Exemple dans mg2026	Test paramétrique	Alternative non paramétrique
mêmes patients		appariés	
deux groupes indépendants	PAS chez les femmes et chez les hommes	test t pour échantillons indépendants (Student ou Welch)	U de Mann-Whitney
trois groupes indépendants ou plus	baisse de la glycémie dans les bras A, B, C	ANOVA à un facteur	Kruskal-Wallis
trois mesures ou plus chez les mêmes patients	glycémie à 0, 3 et 6 mois	ANOVA à mesures répétées	Friedman

Tous les tests du tableau ont la même structure, celle du chapitre 5 : une hypothèse nulle d'égalité des moyennes, une statistique de test qui rapporte le signal au bruit, une distribution connue de cette statistique sous H0 (t ou F), une valeur p et une décision. Seule change la manière de mesurer le signal et le bruit.

6.2 Comparaison d'une moyenne à une valeur donnée : test z et test t

La première situation du tableau a été traitée en détail au chapitre 5. Nous la reprenons ici brièvement, avec la règle du cours de la discipline. Le test z s'utilise si la taille de l'échantillon est n ≥ 30 et que la dispersion de la population est connue ; la statistique est $z = \frac{m-\mu_0}{\frac{\sigma}{\sqrt{n}}}$, et les valeurs critiques proviennent de la distribution normale centrée réduite (1,96 pour α = 0,05 en bilatéral). Le test t s'utilise si n < 30 ou si la dispersion de la population est inconnue ; la statistique est $t = \frac{m-\mu_0}{\frac{s}{\sqrt{n}}}$, avec n - 1 degrés de liberté. Les deux tests supposent une distribution approximativement normale de la variable dans la population, condition qui perd de son importance à mesure que n augmente.

Dans la pratique actuelle, σ est rarement connu et les logiciels utilisent le test t quel que soit n. Pour la moyenne de la PAS de mg2026 comparée à 125 mmHg, nous avons obtenu t = 5,59 avec 159 degrés de liberté, p < 0,001 ; pour la glycémie à jeun comparée à 100 mg/dL, t = 4,36, p < 0,001 (chapitre 5). Ce qui nous intéresse maintenant, c'est que le test t pour échantillons appariés, premier test nouveau du chapitre, se ramène exactement à ce test pour une seule moyenne.

6.3 Le test t pour échantillons appariés

6.3.1 Quand les données sont appariées

Deux séries de valeurs sont appariées (dépendantes) lorsque chaque valeur d'une série a une correspondante naturelle et unique dans l'autre. Les situations cliniques typiques sont au nombre de quatre. Le même patient mesuré avant et après un traitement (la glycémie à l'inclusion et à 6 mois). Le même patient mesuré par deux méthodes ou deux appareils (la pression artérielle mesurée avec le tensiomètre automatique et avec le tensiomètre manuel). Deux organes pairs du même patient (l'œil droit et l'œil gauche). Des paires appariées (angl. matched pairs) de patients différents, choisis pour se ressembler en âge, en sexe et en sévérité, dont l'un reçoit le traitement et l'autre non, ou encore des jumeaux.

Qu'ont en commun ces situations ? Les valeurs d'une même paire sont corrélées : un patient dont la glycémie est élevée à l'inclusion aura en règle générale une glycémie relativement élevée à 6 mois aussi. Le test apparié met cette corrélation à profit en éliminant de la comparaison les différences entre patients et en ne conservant que la modification à l'intérieur de chaque patient.

6.3.2 La logique : le test travaille sur les différences

L'idée est simple et élégante. Pour chaque patient, nous calculons la différence entre les deux mesures, di = xi,avant - xi,après. Les deux séries deviennent une seule série de n différences. Si le traitement n'a pas d'effet, la moyenne des différences dans la population est nulle. Le test apparié est donc le test t pour une seule moyenne appliqué aux différences, avec la valeur de référence μ0 = 0 :

H0 : μD = 0 (la moyenne des différences dans la population est nulle ; il n'y a pas de modification entre les deux temps)

H1 : μD ≠ 0 (bilatéral)

Statistique de test :

$$t = \frac{m_D}{\text{ESD}} = \frac{m_D}{\frac{s_D}{\sqrt{n}}}, \text{ avec ddl = n - 1}$$

où mD est la moyenne des différences, sD l'écart-type des différences et n le nombre de paires (et non le nombre total de mesures). L'intervalle de confiance à 95% de la différence moyenne est $m_D \pm t$ critique × ESD, et la taille d'effet pour données appariées est dz $= \frac{m_D}{s_D}$.

Les conditions d'application sont les suivantes : les paires sont indépendantes entre elles (des patients différents ne s'influencent pas) et les différences suivent une distribution approximativement normale. Attention à ce détail, qui revient souvent à l'examen : la normalité se vérifie sur les différences, et non sur chaque série séparément. La glycémie à l'inclusion de mg2026 n'est pas parfaitement normale (Shapiro-Wilk p = 0,037), mais les différences glicemie0 - glicemie6 le sont (p = 0,725), et le test apparié est pleinement justifié. Si les différences sont fortement asymétriques ou comportent des valeurs extrêmes, surtout pour un petit n, on utilise le test de Wilcoxon des rangs signés (chapitre 7).

6.3.3 Un exemple calculé à la main

Chez dix patients hypertendus suivis au cabinet de médecine générale, la PAS a été mesurée avant et après quatre semaines de traitement par un inhibiteur de l'enzyme de conversion. Les données et le calcul figurent dans le tableau 6.2 ; la résolution complète, avec toutes les étapes du format a-g, constitue le problème résolu 6.1.

Tableau 6.2. PAS (mmHg) avant et après 4 semaines de traitement chez 10 patients (données fictives).

Patient	Avant	Après	Différence d	d - mD	$(d - m_D)^2$
1	152	140	12	4,2	17,64
2	148	142	6	-1,8	3,24
3	160	148	12	4,2	17,64
4	145	140	5	-2,8	7,84
5	158	150	8	0,2	0,04
6	150	147	3	-4,8	23,04
7	165	150	15	7,2	51,84
8	142	138	4	-3,8	14,44
9	155	146	9	1,2	1,44
10	149	145	4	-3,8	14,44
Total / moyenne	152,4	144,6	78 / 7,8	0	151,60

La moyenne des différences est $m_D = \frac{78}{10} = 7{,}8$ mmHg. L'écart-type des différences : $s_D = \sqrt{\frac{151{,}60}{9}} = \sqrt{16{,}84} = 4{,}10$ mmHg. L'erreur standard : $\text{ESD} = \frac{4{,}10}{\sqrt{10}} = \frac{4{,}10}{3{,}162} = 1{,}30$ mmHg. La statistique de test : $t = \frac{7{,}8}{1{,}30} = 6{,}01$, avec 9 degrés de liberté. La valeur critique pour ddl = 9 et α = 0,05 en bilatéral est 2,262 (tableau 5.6), et pour α = 0,01 elle vaut 3,250 ; t = 6,01 les dépasse toutes les deux. La valeur p exacte est 0,0002. IC à 95% de la baisse moyenne : 7,8 ± 2,262 × 1,30 = 7,8 ± 2,94, soit 4,86-10,74 mmHg. Taille d'effet $\text{dz} = \frac{7{,}8}{4{,}10} = 1{,}90$, un effet très important.

Que se serait-il passé si nous avions ignoré l'appariement et comparé les deux colonnes comme si elles provenaient de patients différents ? Le test t pour échantillons indépendants aurait donné t = 2,96 et p = 0,008 : toujours significatif, mais avec une statistique deux fois plus petite. La raison en est que, dans l'analyse indépendante, le bruit inclut les grandes différences entre patients (la PAS initiale varie entre 142 et 165 mmHg, ÉT 7,11), alors que dans l'analyse appariée le bruit se réduit à la variabilité de la réponse au traitement (ÉT des différences 4,10). Sur des données réelles, avec des effets plus faibles, ignorer l'appariement transforme souvent un résultat significatif en résultat non significatif. L'erreur inverse, traiter des groupes indépendants comme s'ils étaient appariés, est encore plus grave, car elle produit des résultats dénués de sens : l'appariement doit exister dans le schéma de l'étude, il ne s'invente pas au moment de l'analyse.

6.3.4 Le test apparié sur le jeu de données mg2026

Dans mg2026, la glycémie à jeun a été mesurée à l'inclusion et à 6 mois chez l'ensemble des 160 patients. La figure 6.1 montre, dans le panneau A, les trajectoires individuelles de 25 patients (chaque ligne relie les valeurs d'un même patient) et la trajectoire de la moyenne ; dans le panneau B, l'histogramme des 160 différences individuelles.

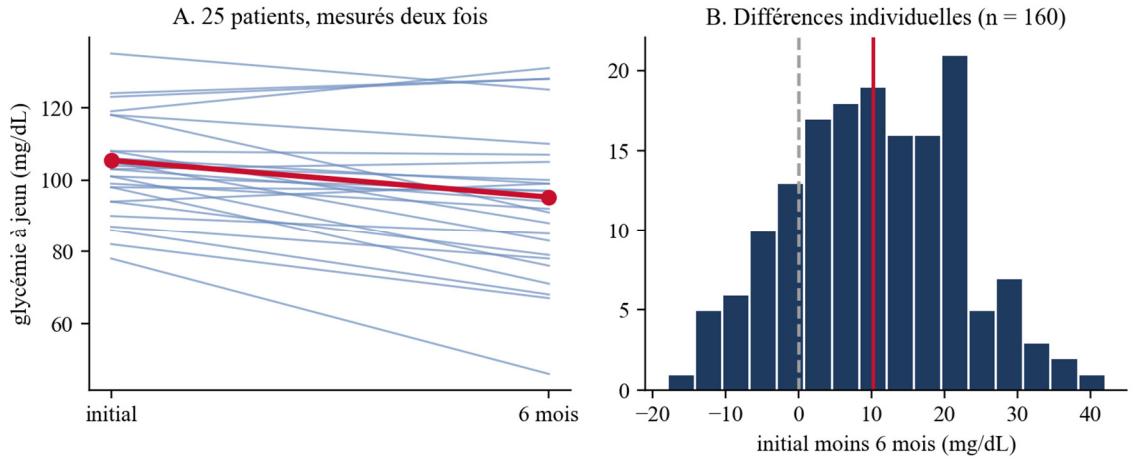


Figure 6.1. Glycémie à jeun à l'inclusion et à 6 mois. A. Trajectoires de 25 patients (lignes fines) et moyenne du groupe (ligne épaisse). B. Histogramme des différences individuelles (valeur initiale moins valeur à 6 mois) chez les 160 patients ; la ligne pointillée marque la différence nulle et la ligne continue la moyenne des différences (10,19 mg/dL). Données synthétiques (jeu de données mg2026).

Le graphique raconte déjà l'histoire avant tout calcul. La plupart des lignes descendent, mais pas toutes : 125 patients ont une glycémie plus basse à 6 mois, 34 l'ont plus élevée et un seul a gardé la même valeur. L'histogramme des différences est à peu près symétrique, centré autour de 10 mg/dL, avec des valeurs comprises entre -18 et +42 mg/dL.

Résultats numériques : moyenne à l'inclusion 105,4 mg/dL, à 6 mois 95,2 mg/dL ; différence moyenne 10,19 mg/dL, avec un écart-type des différences de 11,57 mg/dL. $\text{ESD} = \frac{11{,}57}{\sqrt{160}} =$

$\frac{11,57}{12,649} = 0,915$ mg/dL. $t = \frac{10,19}{0,915} = 11,14$, avec 159 degrés de liberté, p < 0,001. IC à 95% de la différence moyenne : 10,19 ± 1,975 × 0,915 = 10,19 ± 1,81, soit 8,38-11,99 mg/dL. Taille d'effet $\mathrm{dz} = \frac{10,19}{11,57} = 0,88$. Le test de Shapiro-Wilk sur les différences donne p = 0,725, la condition de normalité est donc remplie ; le test de Wilcoxon, à titre de comparaison, donne lui aussi p < 0,001.

L'interprétation se situe à deux niveaux. Sur le plan statistique, la baisse de la glycémie est extrêmement significative, avec une taille d'effet importante. Sur le plan clinique, il faut se rappeler que l'analyse appariée portant sur l'ensemble des patients n'a pas de groupe témoin : elle décrit l'évolution de l'échantillon, et non l'effet d'une intervention particulière. La question de l'efficacité de l'application et de la téléconsultation ne trouve de réponse que par la comparaison des bras (section 6.5). La même logique s'applique sur des périodes plus courtes : entre l'inclusion et 3 mois, la glycémie baisse en moyenne de $6,08$ mg/dL$(\mathrm{IC}\,95\%\,4,61 - 7,54;\ t = 8,22;\ p < 0,001;\ \mathrm{dz} = 0,65)$, et entre 3 et 6 mois, de $4,11$ mg/dL$(\mathrm{IC}\,95\%\,2,10 - 6,13;\ t = 4,03;\ p < 0,001;\ \mathrm{dz} = 0,32)$ supplémentaires.

Un exemple utile pour discuter des conditions d'application apparaît si l'on considère uniquement le bras A (conseil standard, n = 53) et le poids corporel. La perte de poids moyenne entre l'inclusion et 6 mois est de 0,71 kg (ÉT 2,48 kg). Le test t apparié donne t = 2,07, ddl = 52, p = 0,043, soit un résultat tout juste significatif ; le test de Wilcoxon sur les mêmes données donne p = 0,052, tout juste non significatif. Les différences ont une distribution compatible avec la normalité (Shapiro-Wilk p = 0,696), le test t est donc le choix justifié. L'exemple montre combien un résultat situé près du seuil est fragile : deux tests raisonnables aboutissent à des décisions opposées, et une perte de 0,7 kg en 6 mois est de toute façon dépourvue d'importance clinique. Le choix du test se fait d'après les conditions d'application, vérifiées au préalable, et non d'après la valeur p qui arrange.

Étude de cas : les données appariées en pharmacologie et en nutrition

Une étude avant-après chez des patients hypertendus présentant une dysfonction endothéliale a suivi le taux sérique de thrombospondine-1 à l'inclusion (T0) et après un an de traitement par périndopril (T1) : $8061,9 \pm 3684,80$ ng/mL à T0 contre 15380 ± 5887 ng/mL à T1, p < 0,001, avec en outre un groupe témoin de personnes non hypertendues ($9221,03 \pm 6510,21$ ng/mL) [86]. Le second marqueur rapporté a diminué de 0,91 ± 0,51 à $0,50 \pm 0,24$ ng/mL$(p < 0,001)$. Le schéma est classique pour le test apparié (ou pour le test de Wilcoxon si les distributions sont asymétriques ; les écarts-types élevés par rapport aux moyennes suggèrent cette possibilité). L'exercice de lecture critique est le suivant : à partir de la moyenne et de l'ÉT à chaque temps, on ne peut pas reconstituer le test apparié, car ce test utilise l'écart-type des différences individuelles, qui dépend de la corrélation entre les deux mesures. Un compte rendu complet inclurait la différence moyenne (ou la médiane des différences) avec son intervalle de confiance. Un exemple de compte rendu de ce type provient de l'étude prospective portant sur 921 étudiants suivis pendant leur première année d'université : la modification moyenne de l'IMC a été de -0,45 ± 1,07 kg/m² (IC à 95% -0,52 à -0,38 ; p < 0,001), 48,2% des participants ont vu leur IMC diminuer, 38,7% sont restés stables et 13,1% ont vu leur IMC augmenter [5]. Nous pouvons vérifier l'intervalle : $\mathrm{ESD} = \frac{1,07}{\sqrt{921}} = 0,035$, et -0,45 ± 1,96 × 0,035 donne -0,52 à -0,38, exactement les valeurs rapportées. Remarquez aussi la pertinence clinique : une baisse moyenne de 0,45 kg/m² est extrêmement significative pour n = 921, mais faible pour un individu ; les proportions d'étudiants qui prennent ou perdent du poids en disent plus que la moyenne. Un troisième exemple, l'étude rétrospective sur les inhibiteurs du SGLT-2 chez 200 patients atteints de diabète de type 2, rapporte les modifications avant-après sous forme de médianes avec EIQ (par exemple, la glycémie passant de 180,00 à 130,00

> mg%, p < 0,001), ce qui suggère l'utilisation du test de Wilcoxon, abordé au chapitre suivant [59].

6.4 Le test t pour échantillons indépendants

6.4.1 Schéma d'étude et hypothèses

Deux échantillons sont indépendants lorsqu'ils proviennent de patients différents et qu'il n'existe aucune correspondance entre les valeurs d'un groupe et celles de l'autre : femmes et hommes, fumeurs et non-fumeurs, patients traités par le médicament A et patients traités par le médicament B dans un essai randomisé en groupes parallèles. Les groupes peuvent être de tailles différentes. La question est de savoir si les moyennes des deux populations diffèrent :

H0 : µ1 = µ2 (ou µ1 - µ2 = 0)

H1 : µ1 ≠ µ2 (bilatéral)

6.4.2 La version de Student, avec variance commune

Le signal est la différence entre les moyennes des échantillons, m1 - m2. Le bruit est l'erreur standard de cette différence. Dans la version classique, Student suppose que les deux populations ont la même variance et l'estime par une variance commune (angl. pooled variance), moyenne pondérée des deux variances :

$$s^2_{\text{commune}} = \frac{(n_1 - 1) \times {s_1}^2 + (n_2 - 1) \times {s_2}^2}{n_1 + n_2 - 2}$$

$$\text{ES}(m_1 - m_2) = \sqrt{\left[s^2_{\text{commune}} \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]}$$

$t = \frac{m_1 - m_2}{\text{ES}(m_1 - m_2)}$, avec ddl = n1 + n2 - 2

L'intervalle de confiance à 95% de la différence des moyennes est $(m_1 - m_2) \pm t$ critique × ES(m1 - m2), et la taille d'effet est le d de Cohen = (m1 - m2) / scommune.

6.4.3 La version de Welch et le test de Levene

Si les variances des deux populations diffèrent, surtout lorsque les effectifs des groupes sont eux aussi inégaux, la version de Student peut donner des valeurs p erronées. La version de Welch ne suppose plus l'égalité des variances : chaque groupe conserve sa propre variance dans le calcul de l'erreur standard, et les degrés de liberté sont ajustés (la formule de Welch-Satterthwaite donne en général un nombre décimal) :

$$\text{ES}(m_1 - m_2) = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

$$\text{ddl} = \frac{\left(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}\right)^2}{\frac{\left(\frac{{s_1}^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{{s_2}^2}{n_2}\right)^2}{n_2 - 1}}$$

Pour choisir entre les deux versions, les logiciels rapportent le test de Levene d'égalité des variances, avec H0 : $\sigma 1^2 = \sigma 2^2$. Si p Levene ≥ 0,05, nous n'avons pas de raison de rejeter l'égalité des variances et l'on peut utiliser la version de Student ; si p Levene < 0,05, on utilise la version

de Welch. Une règle pratique rapide : si le plus grand écart-type dépasse plus de deux fois le plus petit, l'hypothèse d'une variance commune n'est plus raisonnable.

De plus en plus de statisticiens recommandent d'utiliser par défaut le test de Welch, sans test de Levene préalable : lorsque les variances sont égales, Welch donne pratiquement le même résultat que Student, et lorsqu'elles ne le sont pas, il protège le chercheur d'une erreur. JASP propose les deux versions en cochant simplement les options Student et Welch ; dans Google Sheets, la fonction T.TEST utilise le type 2 pour la variance commune et le type 3 pour des variances inégales. Dans ce cours, rapportez le test de Levene et choisissez la version en fonction de son résultat, mais retenez que choisir Welch n'est jamais une erreur grave.

6.4.4 Groupes de tailles inégales

Le test t n'exige pas des groupes égaux ; dans mg2026, nous avons 86 femmes et 74 hommes, et dans les études observationnelles les déséquilibres peuvent être bien plus marqués (par exemple, 224 patients atteints de diabète de type 2 contre 62 atteints de diabète de type 1 [8]). Trois points doivent néanmoins être retenus. Premièrement : pour un même nombre total de patients, la puissance est maximale lorsque les groupes sont égaux ; une étude avec 20 + 140 patients a beaucoup moins de puissance qu'une étude avec 80 + 80. Deuxièmement : la combinaison de groupes inégaux et de variances inégales est la situation dans laquelle la version de Student se trompe le plus ; si le petit groupe a la plus grande variance, le test de Student devient trop permissif (trop de résultats faussement positifs), et si le petit groupe a la plus petite variance, il devient trop conservateur. L'exercice 3 de ce chapitre présente une telle situation, dans laquelle Student et Welch conduisent à des décisions opposées. Troisièmement : de grandes différences de taille entre les groupes soulèvent des questions de sélection, qu'aucun test ne résout.

6.4.5 Le test pour échantillons indépendants sur le jeu de données mg2026

Le résultat non significatif utilisé au chapitre 5 provient d'un test pour échantillons indépendants : la PAS chez les femmes $131{,}37 \pm 14{,}47$ mmHg$(n = 86)$ et chez les hommes $130{,}85 \pm 13{,}25$ mmHg$(n = 74)$. Le calcul à la main avec la variance commune : $s^2_{\text{commune}} = \frac{85\times14{,}47^2+73\times13{,}25^2}{158} = \frac{85\times209{,}38+73\times175{,}56}{158} = \frac{17797{,}4+12816{,}0}{158} = 193{,}76$, donc scommune = 13,92 mmHg. $\text{ES} = \sqrt{\left[193{,}76 \times \left(\frac{1}{86}+\frac{1}{74}\right)\right]} = \sqrt{193{,}76 \times 0{,}02514} = \sqrt{4{,}871} = 2{,}207$ mmHg. $t = \frac{130{,}85-131{,}37}{2{,}207} = -\frac{0{,}52}{2{,}207} = -0{,}24$, ddl = 158, p = 0,81. Le test de Levene donne p = 0,529, les variances peuvent donc être considérées comme égales ; la version de Welch donne pratiquement le même résultat (t = -0,24, ddl = 157,4, p = 0,81). La différence hommes moins femmes est de -0,5 mmHg (IC à 95% -4,9 à 3,8), et d = -0,04. Le test de Mann-Whitney, à titre de comparaison, donne p = 0,80. On ne rejette pas H0 ; comme nous l'avons vu au chapitre 5, l'intervalle de confiance exclut des différences moyennes supérieures à environ 5 mmHg.

La glycémie à jeun selon le sexe donne un résultat semblable : les hommes ont une moyenne de $106{,}2$ mg/dL$(\text{ÉT } 15{,}19;\ n = 74)$, les femmes de $104{,}8$ mg/dL$(\text{ÉT } 16{,}30;\ n = 86)$. Test de Levene p = 0,426 ; test t de Welch t = 0,54, ddl = 157,0, p = 0,589 ; différence hommes moins femmes 1,35 mg/dL (IC à 95% -3,57 à 6,27), d = 0,09. Non significatif. Pour voir aussi un résultat significatif sur le même type de comparaison, regardez la taille : femmes $162{,}6 \pm 6{,}06$ cm, hommes $175{,}97 \pm 6{,}18$ cm ; Levene p = 0,890 ; test t de Student t = 13,79, ddl = 158, p < 0,001 ; différence 13,37 cm (IC à 95% 11,45-15,28), d = 2,19. Ici, la différence est évidente même sans statistique, mais le test la quantifie.

La littérature publiée par nos collègues contient de nombreuses comparaisons de deux moyennes. Dans l'étude portant sur des patients atteints d'insuffisance cardiaque ischémique à

fraction d'éjection réduite, comparés à 60 témoins sains du même âge, l'ostéoprotégérine était bien plus élevée chez les patients (4,7 ± 0,25 contre $1{,}3 \pm 0{,}67$ ng/mL, $p < 0{,}001$) et la vitamine D3 bien plus basse (20,49 ± 7,31 contre 37,09 ± 4,59 ng/mL, p < 0,001) [87]. Dans l'étude diagnostique par échographie de contraste, le paramètre CEUS-PAT était de 0,83 ± 0,09 chez les 37 patients cirrhotiques et de 0,49 ± 0,11 chez les 27 volontaires sains (p < 0,0001) [2] ; si l'on calcule d à partir des valeurs publiées, on obtient environ 3,4, soit une séparation presque complète des deux groupes. De telles différences énormes sont typiques des schémas cas-témoins avec des témoins parfaitement sains, qui exagèrent la capacité de discrimination d'un test par rapport à la situation réelle, dans laquelle le médecin doit distinguer la cirrhose d'autres hépatopathies (effet de spectre, chapitre 11). Dans l'étude transversale sur les troubles de l'interaction intestin-cerveau, l'IMC était de 28,15 ± 6,49 kg/m² chez les 52 cas et de 24,47 ± 4,60 kg/m² chez les 84 témoins (p = 0,001) [71] ; ici, les groupes sont inégaux et les écarts-types diffèrent (rapport de 1,4), situation dans laquelle la version de Welch est la plus prudente.

Étude de cas : quand l'écart-type se rapproche de la moyenne

Une étude diagnostique menée au service des urgences chez 85 patients faisant l'objet d'une alerte AVC (accident vasculaire cérébral) (69 avec AVC, 16 sans AVC) a comparé deux biomarqueurs sériques [61]. La protéine gliofibrillaire acide (GFAP) était de $142{,}91 \pm 102{,}19$ pg/mL chez les patients avec AVC et de $37{,}76 \pm 19{,}92$ pg/mL chez ceux sans AVC (p < 0,001) ; l'UCH-L1 était de 1307,68 ± 967,54 contre $189{,}81 \pm 92{,}69$ pg/mL($p < 0{,}001$). Les différences sont importantes et cliniquement intéressantes, et pour distinguer l'AVC hémorragique de l'AVC ischémique, la GFAP a atteint une exactitude de 94,2% au seuil de 77,15 pg/mL. Pour un étudiant de première année, ce résumé est un excellent exercice sur les conditions du test t. Premier signal : l'écart-type est presque égal à la moyenne (102,19 contre 142,91 ; 967,54 contre 1307,68). Une variable qui ne peut pas être négative et dont l'ÉT est proche de la moyenne est presque certainement asymétrique à droite ; la moyenne ± 2ÉTar inclut des valeurs négatives, biologiquement impossibles. Deuxième signal : les variances diffèrent énormément entre les groupes (le rapport des écarts-types est d'environ 5 pour la GFAP et de 10 pour l'UCH-L1), et les groupes sont très inégaux (69 contre 16). Dans ces conditions, le test t de Student avec variance commune n'est pas adapté ; les options correctes sont le test de Welch sur les données log-transformées ou le test de Mann-Whitney, et une description par la médiane et l'EIQ rendrait plus fidèlement compte des distributions. La conclusion clinique de l'étude (les biomarqueurs sont bien plus élevés en cas d'AVC) ne change pas, car les différences sont très importantes ; pour des différences plus faibles, le choix du test pourrait modifier la décision. Le résumé ne nomme pas le test utilisé : l'exercice n'est donc pas une critique, mais une liste de points à vérifier avant de choisir le test pour vos propres données.

6.5 Analyse de la variance (ANOVA) à un facteur

6.5.1 Pourquoi on ne fait pas plusieurs tests t

Dans mg2026, nous avons trois bras de traitement. La question "la baisse de la glycémie diffère-t-elle entre les bras ?" pourrait être abordée avec trois tests t : A contre B, A contre C, B contre C. Le problème est celui du chapitre 5 : chaque test comporte un risque de 5% de faux positif, et le risque qu'au moins un des trois soit faussement positif, s'il n'existe aucune différence réelle, monte à environ 1 - 0,95³ = 0,143 (les tests ne sont pas parfaitement indépendants, mais l'ordre de grandeur est correct). Pour 4 groupes, il y a 6 comparaisons, pour 5 groupes 10, pour k groupes k(k - 1)/2, et le risque global augmente rapidement. L'analyse de la variance (ANOVA, angl. analysis of variance), développée par R. A. Fisher dans les années 1920, répond à la question globale par un seul test, en maintenant le risque d'erreur de type I à α [42,53].

H0 : μ1 = μ2 =... = μk (toutes les moyennes des populations sont égales)

H1 : au moins une moyenne diffère des autres

Attention à la formulation de H1 : l'ANOVA ne dit pas quelle moyenne diffère. Pour cela, les tests post hoc sont nécessaires (section 6.6).

6.5.2 La logique : variation entre les groupes comparée à la variation à l'intérieur des groupes

Le nom "analyse de la variance" paraît étrange pour un test qui compare des moyennes. Il vient de la manière dont la comparaison est faite. La variabilité totale des données (l'écart de chaque valeur à la moyenne générale) se décompose en deux parties. La variation entre les groupes indique dans quelle mesure les moyennes des groupes s'écartent de la moyenne générale ; si H0 est vraie, elle est faible, produite uniquement par le hasard. La variation à l'intérieur des groupes indique dans quelle mesure chaque valeur s'écarte de la moyenne de son propre groupe ; elle reflète la variabilité biologique et la variabilité de mesure et ne dépend pas de H0. Le rapport entre les deux, après division de chacune par ses degrés de liberté, est la statistique F :

F = MS inter-groupes / MS intra-groupes = (variation expliquée par le groupe) / (variation non expliquée)

Si H0 est vraie, les deux MS estiment la même variance et F est voisin de 1. Si les moyennes diffèrent, le MS inter-groupes augmente et F devient bien supérieur à 1. La distribution de F sous H0 est la distribution F de Fisher, avec deux nombres de degrés de liberté : k - 1 au numérateur et N - k au dénominateur (N étant le nombre total d'observations).

La figure 6.2 illustre cette logique sur un petit exemple : le score de douleur (0-10) chez 15 patients traités par trois antalgiques, 5 patients par groupe. Les flèches montrent la distance entre chaque moyenne de groupe et la moyenne générale (variation entre les groupes) ; les fines lignes verticales montrent la distance entre chaque patient et la moyenne de son groupe (variation à l'intérieur des groupes). Les moyennes sont bien séparées et les valeurs de chaque groupe sont resserrées autour de la moyenne : F sera grand.

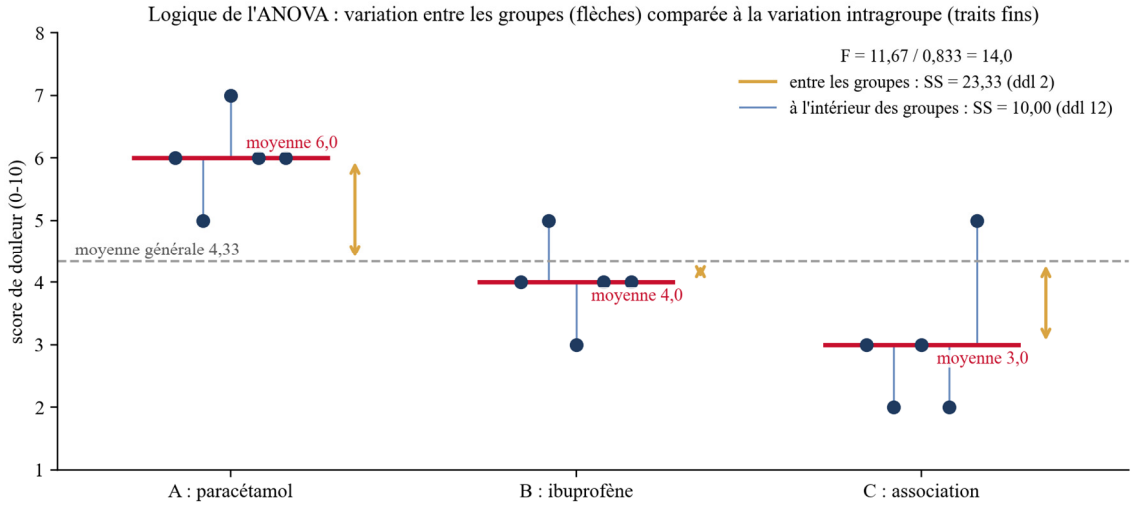


Figure 6.2. La logique de l'ANOVA sur un exemple avec trois antalgiques (A paracétamol, B ibuprofène, C association ; 5 patients chacun, score de douleur 0-10). Les lignes rouges horizontales sont les moyennes des groupes, la ligne pointillée est la moyenne générale (4,33). Les flèches représentent la variation entre les groupes, les lignes fines la variation à l'intérieur des groupes. Données fictives.

6.5.3 Sommes des carrés, degrés de liberté et tableau d'ANOVA

La décomposition se fait à l'aide de sommes de carrés (SS, angl. sum of squares). Notons xij la valeur du patient j du groupe i, mi la moyenne du groupe i, ni le nombre de patients du groupe i, M la moyenne générale et N le nombre total de patients.

SS total $= \sum(x_{ij} - M)^2$, avec ddl = N - 1

SS inter groupes $= \sum n_i \times (m_i - M)^2$, avec ddl = k - 1

SS intra-groupes (résiduelle) = $\sum(x_{ij} - m_i)^2 = \sum(n_i - 1) \times {s_i}^2$, avec ddl = N - k

La propriété fondamentale est l'additivité : SS total = SS inter + SS intra, et ddl total = ddl inter + ddl intra. Les carrés moyens (MS, angl. mean squares) s'obtiennent en divisant chaque SS par ses degrés de liberté : MS inter = SS inter / (k - 1), MS intra = SS intra / (N - k). Ensuite F = MS inter / MS intra, et la valeur p est l'aire sous la distribution F à droite de la valeur calculée. Toutes ces quantités se disposent dans un tableau standard, que produisent tous les logiciels.

Tableau 6.3. Structure du tableau d'ANOVA à un facteur.

Source de variation	SS	ddl	MS	F	p
inter-groupes	$\sum n_i(m_i - M)^2$	k - 1	SS inter / (k - 1)	MS inter / MS intra	d'après la distribution F
intra-groupes	$\sum(n_i - 1)\, {s_i}^2$	N - k	SS intra / (N - k)		
total	$\sum(x_{ij} - M)^2$	N - 1			

La formule SS intra = $\sum(n_i - 1) \times {s_i}^2$ est très utile : elle permet de construire le tableau d'ANOVA à partir des seuls n, moyenne et écart-type de chaque groupe, c'est-à-dire des chiffres rapportés dans n'importe quel article. Nous l'utiliserons sur les données de mg2026 dans le problème résolu 6.6.

Une mesure de la taille d'effet en ANOVA est l'êta carré : η^2 = SS inter / SS total, la proportion de la variabilité totale expliquée par l'appartenance au groupe. Les seuils indicatifs de Cohen sont 0,01 (faible), 0,06 (moyen) et 0,14 (important) [76].

6.5.4 Le tableau d'ANOVA pour l'exemple des antalgiques

Sur les données de la figure 6.2 (A : 6, 5, 7, 6, 6 ; B : 4, 5, 3, 4, 4 ; C : 3, 2, 3, 2, 5), les moyennes des groupes sont 6, 4 et 3, et la moyenne générale est 65/15 = 4,33. SS inter = 5 × [(6 - 4,33)² + (4 - 4,33)² + (3 - 4,33)²] = 5 × (2,778 + 0,111 + 1,778) = 5 × 4,667 = 23,33. SS intra s'obtient en additionnant, dans chaque groupe, les carrés des écarts à la moyenne du groupe : dans le groupe A, 0 + 1 + 1 + 0 + 0 = 2 ; dans le groupe B, 0 + 1 + 1 + 0 + 0 = 2 ; dans le groupe C, 0 + 1 + 0 + 1 + 4 = 6 ; total 10,00. SS total = 33,33 (vérification : 23,33 + 10,00).

Tableau 6.4. Tableau d'ANOVA pour le score de douleur sous trois antalgiques (données fictives).

Source de variation	SS	ddl	MS	F	p
inter-groupes	23,33	2	11,67	14,00	0,0007
intra-groupes	10,00	12	0,833		
total	33,33	14			

La valeur critique F(2 ; 12) pour α = 0,05 est 3,89, et pour α = 0,01 elle vaut 6,93 ; F = 14,00 les dépasse toutes les deux, et p = 0,0007. Nous rejetons H0 : au moins un antalgique diffère des autres quant au score moyen de douleur. $\eta^2 = \frac{23{,}33}{33{,}33} = 0{,}70$: 70% de la variabilité du score est expliquée par le type d'antalgique, un effet très important (ce qui est naturel pour un exemple

construit pour être clair). Pour savoir quels antalgiques diffèrent entre eux, nous appliquerons le test de Tukey dans le problème résolu 6.3.

6.5.5 Les conditions d'application

L'ANOVA à un facteur repose sur trois conditions, semblables à celles du test t pour échantillons indépendants.

Indépendance des observations. Chaque patient n'apparaît qu'une seule fois, dans un seul groupe, et les patients ne s'influencent pas entre eux. C'est la condition la plus importante, et elle est garantie par le schéma de l'étude ; aucun test statistique ne peut la vérifier ni la réparer. Les mesures répétées chez les mêmes patients exigent une ANOVA à mesures répétées (section 6.7).

Normalité dans chaque groupe (ou, de façon équivalente, normalité des résidus, c'est-à-dire des écarts à la moyenne du groupe). Elle se vérifie par le test de Shapiro-Wilk dans chaque groupe ou par le diagramme Q-Q des résidus. Avec des groupes de grande taille et d'effectifs voisins, l'ANOVA est robuste face à des écarts modérés.

Homogénéité des variances (homoscédasticité). Elle se vérifie par le test de Levene (H0 : les variances sont égales) ou par le test de Bartlett, plus sensible aux écarts à la normalité. Règle pratique : l'ANOVA classique reste acceptable si le rapport entre le plus grand et le plus petit écart-type est inférieur à 2 et que les groupes ont des effectifs voisins.

Lorsque les conditions ne sont pas remplies, nous disposons d'alternatives. Pour des variances inégales : l'ANOVA de Welch (F avec degrés de liberté ajustés) et le test post hoc de Games-Howell. Pour des distributions fortement asymétriques ou des données ordinales : le test de Kruskal-Wallis (chapitre 7) ou une transformation logarithmique.

6.5.6 L'ANOVA sur le jeu de données mg2026 : baisse de la glycémie selon le bras

Nous arrivons à la question principale de l'étude mg2026. La variable analysée est la baisse de la glycémie à jeun à 6 mois, calculée pour chaque patient comme glicemie0 - glicemie6 (des valeurs positives signifient une baisse). La figure 6.3 montre sa distribution dans les trois bras.

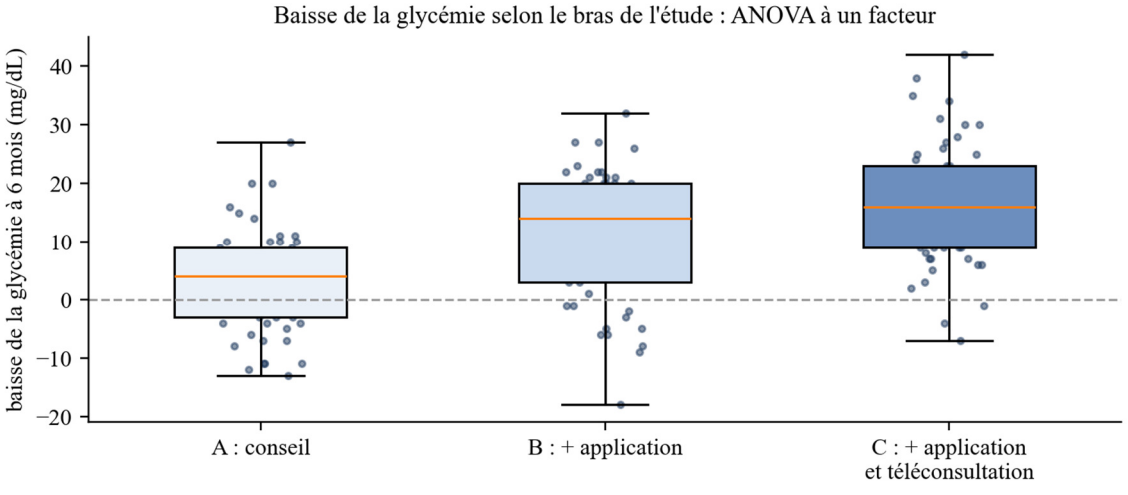


Figure 6.3. Baisse de la glycémie à jeun à 6 mois (glicemie0 - glicemie6) selon le bras de l'étude : A conseil standard, B conseil + application, C conseil + application + téléconsultation. Boîte à moustaches avec superposition des valeurs individuelles ; la ligne pointillée marque une baisse nulle. Données synthétiques (jeu de données mg2026).

Le graphique suggère une baisse de la glycémie de plus en plus marquée du bras A au bras C, avec des dispersions semblables. Statistiques descriptives : A, moyenne 3,0 mg/dL$(\text{ÉT } 8{,}8;\ n = 53)$; B, 11,1 mg/dL$(\text{ÉT } 11{,}2;\ n = 53)$; C, 16,3 mg/dL$(\text{ÉT } 10{,}6;\ n = 54)$; moyenne générale 10,19 mg/dL.

Vérification des conditions. Les patients sont indépendants et chacun n'apparaît que dans un seul bras (randomisation individuelle). Test de Shapiro-Wilk dans chaque bras : A p = 0,575, B p = 0,052, C p = 0,903 ; la normalité est acceptable, le bras B étant à la limite, et les groupes sont grands et égaux. Le rapport entre le plus grand et le plus petit ÉT est de 11,2 / 8,8 = 1,27, inférieur à 2. Le test de Levene donne p = 0,163 dans sa version centrée sur la médiane (Brown-Forsythe) et p = 0,089 dans sa version centrée sur la moyenne ; le test de Bartlett, p = 0,192. Nous n'avons pas de raison de rejeter l'homogénéité des variances.

Tableau 6.5. Tableau d'ANOVA pour la baisse de la glycémie à 6 mois selon le bras (mg2026).

Source de variation	SS	ddl	MS	F	p
inter-bras	4806,9	2	2403,4	22,93	< 0,001
intra-bras	16459,5	157	104,8		
total	21266,4	159			

F(2 ; 157) = 22,93, avec p < 0,001 (la valeur critique pour α = 0,05 est 3,05). $\eta^2 = \frac{4806,9}{21266,4} = 0,226$: près de 23% de la variabilité de la baisse de la glycémie est expliquée par le bras de traitement, un effet important. L'ANOVA de Welch, qui ne suppose pas l'égalité des variances, le confirme : F(2 ; 103,5) = 26,05, p < 0,001. Le test non paramétrique de Kruskal-Wallis donne H = 36,63, p < 0,001.

Conclusion statistique : on rejette H0 ; la baisse moyenne de la glycémie diffère de façon extrêmement significative entre au moins deux bras. La conclusion de recherche exige les tests post hoc, afin de voir quels bras diffèrent.

6.5.7 Le second résultat : la perte de poids

La même analyse pour la perte de poids à 6 mois (greutate0 - greutate6, kg) donne : A 0,71 ± 2,48 kg, B 2,84 ± 2,36 kg, C 3,37 ± 2,39 kg. Les écarts-types sont presque identiques (Levene p = 0,79), et les distributions sont compatibles avec la normalité dans tous les bras (Shapiro-Wilk p entre 0,17 et 0,70). ANOVA : SS inter = 211,9 (ddl 2), SS intra = 911,3 (ddl 157), MS inter = 105,9, MS intra = 5,80, F(2 ; 157) = 18,25, p < 0,001, $\eta^2 = 0,189$. Les bras avec application ont perdu plus de poids que le bras avec conseil standard ; quant à savoir si la téléconsultation apporte un bénéfice supplémentaire par rapport à l'application, ce sont encore les tests post hoc qui nous le diront.

6.6 Les tests post hoc

6.6.1 Quel problème ils résolvent

Un F significatif indique seulement que les moyennes ne sont pas toutes égales. Pour savoir quelles paires diffèrent, on procède à des comparaisons ultérieures, appelées tests post hoc (lat. "après cela"). Ils comparent les paires de groupes, mais ajustent le seuil ou les valeurs p de façon que le risque global d'erreur de type I reste égal à α, exactement l'idée des corrections pour comparaisons multiples du chapitre 5. Il existe de nombreux tests post hoc ; en médecine clinique, quatre sont utilisés.

Le test HSD de Tukey (angl. honestly significant difference) compare toutes les paires possibles et constitue le choix standard lorsque les variances sont homogènes et que les groupes ont des effectifs voisins. Il utilise la distribution de "l'étendue studentisée" q et calcule la différence minimale significative :

HSD = q(k ; N - k ; α) × √(MS intra / n)

où n est le nombre de patients d'un groupe (pour des groupes inégaux, les logiciels utilisent la version de Tukey-Kramer). Deux moyennes diffèrent significativement si leur différence dépasse le HSD.

La correction de Bonferroni appliquée aux tests t deux à deux multiplie chaque valeur p par le nombre de comparaisons. Elle est simple, mais conservatrice lorsque les comparaisons sont nombreuses ; la variante de Holm est préférable.

Le test de Games-Howell ne suppose ni l'égalité des variances ni l'égalité des effectifs. C'est le choix adapté lorsque le test de Levene est significatif ou que les écarts-types diffèrent beaucoup, en général en association avec l'ANOVA de Welch.

Le test de Dunnett compare chaque groupe de traitement uniquement à un groupe témoin, et non toutes les paires entre elles. Comme il y a moins de comparaisons, il a plus de puissance. Il convient lorsque la question est "quels traitements diffèrent du témoin ?" et que les comparaisons entre traitements ne présentent pas d'intérêt.

6.6.2 Le test de Tukey sur le jeu de données mg2026

Pour la baisse de la glycémie selon le bras, le test de Tukey donne les résultats du tableau 6.6.

Tableau 6.6. Comparaisons post hoc pour la baisse de la glycémie à 6 mois (mg2026).

Comparaison	Différence moyenne (mg/dL)	IC à 95% de Tukey	p Tukey	p Games-Howell	p Bonferroni
B - A	8,1	3,4-12,8	0,0002	0,0002	0,0002
C - A	13,3	8,6-18,0	< 0,001	< 0,001	< 0,001
C - B	5,2	0,6-9,9	0,024	0,038	0,044

Les trois paires diffèrent toutes significativement. L'application (bras B) apporte une baisse supplémentaire de 8,1 mg/dL par rapport au conseil standard, et la téléconsultation mensuelle (bras C) 5,2 mg/dL de plus par rapport à l'application. La différence C - B est la plus petite et la plus proche du seuil : elle reste significative avec les trois méthodes, mais son intervalle de confiance commence à 0,6 mg/dL, de sorte que le bénéfice supplémentaire de la téléconsultation pourrait aussi être très faible. Remarquez que les méthodes plus conservatrices (Games-Howell, Bonferroni) donnent des valeurs p légèrement plus élevées pour cette paire ; si le p de Tukey avait été de 0,045, le choix de la méthode aurait pu modifier la décision, c'est pourquoi ce choix est arrêté avant l'analyse.

Pour la perte de poids, le test de Tukey montre que le bras B (différence par rapport à A de 2,13 kg, $p < 0,001$) et le bras C (2,66 kg, $p < 0,001$) diffèrent de A, mais que C ne diffère pas significativement de B (0,53 kg, $p = 0,489$). En combinant les deux résultats, l'application semble apporter l'essentiel du bénéfice sur le poids, tandis que la téléconsultation ajoute un bénéfice supplémentaire surtout sur la glycémie. Nous formulons cela avec prudence, car l'absence de différence significative C - B pour le poids ne prouve pas l'égalité des deux bras.

6.6.3 Autres options d'analyse et ce qu'il ne faut pas faire

Si la question avait été seulement "l'application et la téléconsultation sont-elles meilleures que le conseil standard ?", le test de Dunnett, avec le bras A comme témoin, aurait été le choix naturel. Ce qu'il ne faut pas faire : appliquer des tests t simples, non corrigés, entre toutes les paires ; réaliser des tests post hoc après un F non significatif, à la recherche d'une paire "qui ressort" (à l'exception des comparaisons planifiées à l'avance et justifiées dans le protocole) ; n'interpréter que les paires significatives, en ignorant l'ampleur des différences et les intervalles de confiance.

Un exemple publié d'analyse portant sur plusieurs groupes ordonnés provient de l'échographie hépatique. Le paramètre UGAP a augmenté de façon monotone avec le grade de

stéatose établi par la méthode de référence : S_0 198,3 $\pm \frac{25{,}7 d_B}{m}$, S_1 216,86 $\pm \frac{26{,}3 d_B}{m}$, S_2 237,79 $\pm \frac{26{,}3 d_B}{m}$, S_3 270,8 $\pm \frac{31{,}62 d_B}{m}$ ($p < 0{,}001$), avec des groupes de 48, 56, 14 et 59 sujets [58]. Les écarts-types sont très proches (rapport de 1,2), l'ANOVA classique est donc adaptée ; le groupe S2 est toutefois petit, et les comparaisons qui le concernent ont moins de puissance. Lorsque les groupes sont naturellement ordonnés (stades, grades), on peut tester, en plus de l'ANOVA, la tendance linéaire, qui répond directement à la question "le paramètre augmente-t-il avec le grade ?". Deux autres études de nos collègues comparent trois groupes sans nommer le test dans le résumé : une étude prospective portant sur 150 femmes enceintes réparties en trois groupes égaux de 50 a trouvé des différences significatives de la rigidité artérielle (p < 0,001), de la fréquence cardiaque (p = 0,006), de la pression systolique aortique (p = 0,02), de l'index d'augmentation brachial (p = 0,01) et de la pression pulsée (p = 0,049) [88], et une étude à trois groupes parallèles de patients après un AVC (22, 18 et 30 patients) a rapporté de meilleurs résultats dans les groupes sous thérapie combinée (p < 0,05) ainsi que des améliorations par rapport à l'état initial à 12 mois (p < 0,05) [89]. Dans les deux cas, le lecteur aurait besoin du nom du test (ANOVA ou Kruskal-Wallis), des tests post hoc utilisés et des moyennes avec leurs intervalles de confiance pour juger de l'ampleur des différences ; lorsque les mêmes patientes ou les mêmes patients sont évalués à plusieurs temps, une analyse pour mesures répétées est en outre nécessaire.

Étude de cas : ANOVA, Tukey et êta² dans deux études de nos collègues

L'étude transversale sur les aides auditives intelligentes a comparé l'estime de soi (échelle SES) chez 245 personnes réparties en trois groupes : 73 normo-entendants, 86 porteurs d'aides auditives et 86 patients atteints de surdité neurosensorielle sans appareillage [82]. L'ANOVA à un facteur a donné F = 299,00, p < 0,001, avec η² = 0,712, et le test post hoc HSD de Tukey a montré que les porteurs d'aides auditives ont des scores plus élevés que les personnes sans appareillage (35,41 ± 5,32 contre $22{,}99 \pm 4{,}53$; $p < 0{,}001$; d de Cohen = 2,515). Un η² de 0,712 signifie que 71% de la variabilité de l'estime de soi est associée au groupe, un effet exceptionnellement important en sciences du comportement. Quelques observations de lecture critique : les écarts-types sont proches (5,32 et 4,53), la condition d'homogénéité semble donc remplie et le test de Tukey est le test adapté ; les groupes ont des effectifs voisins ; le schéma est transversal, de sorte que le résultat montre une association forte, mais n'établit pas que l'appareillage a produit l'augmentation de l'estime de soi (les personnes qui choisissent de porter une aide auditive peuvent aussi différer par d'autres caractéristiques). La seconde étude a comparé la déformation longitudinale globale du ventricule gauche (GLS) chez 148 patients répartis en trois groupes : 62 atteints d'insuffisance cardiaque à fraction d'éjection préservée, 46 hypertendus et 40 témoins du même âge [90]. La différence globale était significative (p < 0,001), et les comparaisons par paires ont montré que les témoins (-20,2 ± 1,4%) diffèrent des hypertendus ($-18{,}4 \pm 3{,}0\%$, $p = 0{,}031$) et des patients en insuffisance cardiaque ($-17{,}6 \pm 2{,}3\%$, $p < 0{,}001$). Ici, l'écart-type des hypertendus est plus de deux fois supérieur à celui des témoins ; dans une réanalyse, une ANOVA de Welch avec test de Games-Howell serait une vérification utile. Remarquez aussi que la différence témoins-hypertendus, bien que significative, a un p beaucoup plus proche du seuil que la différence témoins-insuffisance cardiaque, ce qui correspond à une différence moyenne plus faible (1,8 contre 2,6 points de pourcentage).

6.7 ANOVA à mesures répétées

6.7.1 Quand les mêmes patients sont mesurés plusieurs fois

Dans mg2026, la glycémie a été mesurée à l'inclusion, à 3 mois et à 6 mois chez les mêmes patients. Les trois séries ne sont pas indépendantes, l'ANOVA à un facteur n'est donc pas adaptée

; trois tests t appariés non corrigés ne le sont pas non plus (inflation de l'erreur de type I). Le test adapté est l'ANOVA à mesures répétées, généralisation du test t apparié à trois temps ou plus. Comme le test apparié, elle élimine du bruit les différences constantes entre patients et ne compare que les modifications à l'intérieur de chaque patient.

H0 : μ0 mois = μ3 mois = μ6 mois (la glycémie moyenne ne se modifie pas au cours du temps)

H1 : la moyenne diffère à au moins un temps

6.7.2 La sphéricité et la correction de Greenhouse-Geisser

Outre la normalité, l'ANOVA à mesures répétées comporte une condition spécifique, la sphéricité : les variances des différences entre toutes les paires de temps (0-3, 0-6, 3-6) doivent être égales. Elle se vérifie par le test de Mauchly (H0 : la sphéricité est respectée). Si le test de Mauchly est significatif, les degrés de liberté du test F sont multipliés par un facteur de correction ε (epsilon) inférieur à l'unité, le plus souvent estimé par la méthode de Greenhouse-Geisser (plus conservatrice) ou de Huynh-Feldt. La correction ne change pas la valeur de F, mais les degrés de liberté, de sorte que la valeur p augmente légèrement. Pour deux temps, la sphéricité est automatiquement remplie et l'ANOVA à mesures répétées donne exactement le même résultat que le test t apparié ($F = t^2$).

6.7.3 Le résultat sur mg2026

Les moyennes de la glycémie sont de 105,4 / 99,4 / 95,2 mg/dL à 0 / 3 / 6 mois. ANOVA à mesures répétées : SS temps = 8405,5 (ddl 2), SS erreur = 20570,5 (ddl 318), MS temps = 4202,8, MS erreur = 64,69, F(2 ; 318) = 64,97, p < 0,001. Taille d'effet, êta² partiel = 8405,5 / (8405,5 + 20570,5) = 0,29. Le test de Mauchly est significatif (W = 0,873, p < 0,001), la sphéricité n'est donc pas respectée ; la correction de Greenhouse-Geisser a ε = 0,887, les degrés de liberté deviennent 1,77 et 282,2, et p reste < 0,001. Les comparaisons post hoc avec correction de Bonferroni montrent que les trois paires de temps diffèrent toutes (0-3 mois, 0-6 mois et 3-6 mois, tous p < 0,001). Le test non paramétrique équivalent, celui de Friedman, donne χ^2 = 74,67, p < 0,001 (chapitre 7).

Conclusion : la glycémie moyenne diminue de façon significative et progressive au cours des 6 mois, avec une baisse plus importante au cours des 3 premiers mois (6,1 mg/dL) qu'au cours des 3 suivants (4,1 mg/dL). Comme pour le test apparié, l'analyse sur l'ensemble de l'échantillon décrit l'évolution dans le temps, et non l'effet d'une intervention particulière ; pour la question "l'évolution dans le temps diffère-t-elle entre les bras ?", il faut un modèle mixte, avec le facteur temps (intra-patients) et le facteur bras (inter-patients), dont le terme d'interaction temps × bras répond exactement à cette question. JASP permet ce modèle à partir du même menu (Repeated Measures ANOVA, en ajoutant le bras dans "Between Subject Factors").

> **Étude de cas : mesures répétées dans une étude pilote pédiatrique sans groupe témoin**
>
> Une étude pilote prospective à un seul bras a suivi 21 enfants atteints de dermatite atopique et d'une dysbiose intestinale confirmée au laboratoire, traités par une thérapie personnalisée ciblant le microbiote intestinal, avec des évaluations à l'inclusion, à 30 et à 90 jours [91]. Le score POEM est passé de 14,67 ± 5,22 à l'inclusion à 7,10 ± 3,82 à 90 jours, et le score SCORAD de 41,66 ± 15,28 à 17,93 ± 11,68 (les deux p < 0,001) ; environ 76% des enfants ont obtenu une réduction d'au moins 50% du score POEM et 71% du score SCORAD. L'analyse statistique utilisée, l'ANOVA à mesures répétées, est le bon choix pour trois évaluations des mêmes enfants ; un compte rendu complet inclurait le test de sphéricité et, le cas échéant, la correction de Greenhouse-Geisser, ainsi que les comparaisons post hoc entre les

> temps. La leçon principale concerne toutefois le schéma de l'étude. Dans une étude à un seul bras, la baisse des scores ne peut pas être attribuée avec certitude au traitement : la dermatite atopique a une évolution fluctuante, les enfants sont souvent inclus au moment d'une poussée (la régression vers la moyenne est donc probable), et la participation à l'étude modifie la prise en charge. Les auteurs présentent à juste titre leur étude comme une étude pilote ; des résultats statistiquement très significatifs justifient un essai contrôlé randomisé, et non une recommandation thérapeutique. C'est exactement la différence entre le test apparié sur l'ensemble de l'échantillon mg2026 (évolution dans le temps) et la comparaison entre les bras randomisés (effet de l'intervention).

6.8 ANOVA à deux facteurs et ANCOVA, en bref

6.8.1 ANOVA à deux facteurs

Il arrive que nous nous intéressions simultanément à l'effet de deux facteurs sur une variable numérique, par exemple le bras de traitement et le sexe du patient sur la baisse de la glycémie. L'ANOVA à deux facteurs (angl. two-way ANOVA) teste trois hypothèses : l'effet principal du premier facteur (le bras), l'effet principal du second facteur (le sexe) et l'interaction entre eux (l'effet du bras diffère-t-il entre les femmes et les hommes ?). L'interaction est souvent la question clinique la plus intéressante : un traitement qui ne fonctionne que chez les femmes ou que chez les personnes âgées.

Dans mg2026, l'ANOVA brat × sex pour la baisse de la glycémie donne : effet du bras $F(2 ; 154) = 23{,}03$, $p < 0{,}001$; effet du sexe $F(1 ; 154) = 0{,}66$, $p = 0{,}416$; interaction $F(2 ; 154) = 0{,}53$, $p = 0{,}591$. Les moyennes par cellule sont, chez les femmes et chez les hommes : A 2,6 et 3,6 mg/dL ; B 12,7 et 9,6 mg/dL ; C 17,2 et 15,4 mg/dL. Conclusion : l'effet de l'intervention ne diffère pas significativement entre les sexes, nous pouvons donc rapporter l'effet du bras pour l'ensemble de l'échantillon. Attention, comme toujours : les interactions ont une faible puissance, et l'absence d'interaction significative ne prouve pas que l'effet est identique chez les femmes et chez les hommes.

6.8.2 ANCOVA : l'ajustement sur la valeur initiale

L'analyse de la covariance (ANCOVA) combine l'ANOVA et la régression (chapitre 9) : elle compare les moyennes des groupes après ajustement sur une variable numérique, appelée covariable. Dans les essais randomisés, l'application la plus fréquente est l'ajustement sur la valeur initiale du critère de jugement. Même avec une randomisation, les bras peuvent partir de valeurs légèrement différentes : dans mg2026, la glycémie moyenne à l'inclusion était de 107,8 mg/dL dans le bras A, de 106,3 dans le bras B et de 102,3 dans le bras C. L'ANCOVA compare les glycémies à 6 mois comme si tous les patients étaient partis de la même glycémie initiale.

Le résultat sur mg2026 : après ajustement sur la glycémie initiale, la glycémie à 6 mois diffère entre les bras, $F(2 ; 156) = 22{,}26$, $p < 0{,}001$; par rapport au bras A, la glycémie ajustée est inférieure de 8,07 mg/dL dans le bras B (IC à 95% 4,12-12,01) et de 13,29 mg/dL dans le bras C (IC à 95% 9,32-17,25). Les estimations sont presque identiques aux différences entre les baisses moyennes de l'ANOVA (8,1 et 13,3 mg/dL), ce qui montre que les petits déséquilibres à l'inclusion n'ont pas influencé la conclusion. L'ANCOVA est l'analyse recommandée pour les essais randomisés dont le critère de jugement numérique est mesuré avant et après, car elle a en général plus de puissance que la simple comparaison des valeurs finales ou des différences. Les recommandations CONSORT préconisent que les ajustements soient fixés dans le protocole, et non choisis après avoir regardé les données, et déconseillent de tester statistiquement les différences à l'inclusion entre bras randomisés [56] ; si la randomisation a été correcte, toute différence initiale est, par définition, due au hasard.

6.9 Rapporter les tests t et les ANOVA

Un compte rendu complet d'un test t ou d'une ANOVA comprend : les statistiques descriptives par groupe (n, moyenne, ÉT) ou par temps ; la vérification des conditions (normalité, homogénéité des variances, sphéricité, avec le test utilisé) ; la version du test (Student ou Welch ; ANOVA classique ou de Welch) ; la statistique de test avec ses degrés de liberté (t(159) = 11,14 ; F(2 ; 157) = 22,93) ; la valeur p exacte ; la taille d'effet (d, dz, êta²) et l'intervalle de confiance des différences entre moyennes ; le test post hoc et ses résultats ; l'interprétation clinique [48,53].

Exemple pour mg2026 : "La baisse de la glycémie à jeun à 6 mois a différé entre les bras (A: 3,0 ± 8,8; B: 11,1 ± 11,2; C: 16,3 ± 10,6 mg/dL; ANOVA (à un facteur), $F(2;\ 157) = 22{,}93,\ p < 0{,}001,\ \eta^2 = 0{,}23$; test de Levene p = 0,163). Les comparaisons post hoc de Tukey ont montré des baisses plus importantes dans le bras B par rapport au bras A (différence 8,1 mg/dL, IC à 95% 3,4-12,8 ; p < 0,001), dans le bras C par rapport au bras A (13,3 mg/dL, IC à 95% 8,6-18,0 ; p < 0,001) et dans le bras C par rapport au bras B (5,2 mg/dL, IC 95% 0,6 − 9,9; $p = 0{,}024$)."

Pour les graphiques, la boîte à moustaches avec les points individuels (figure 6.3) ou le graphique des moyennes avec IC à 95% sont préférables aux diagrammes en barres avec "antennes" d'ÉT, qui masquent la distribution des données. Pour des données appariées, le graphique à lignes individuelles (figure 6.1 A) montre ce qu'aucun tableau ne peut montrer : combien de patients s'améliorent, combien s'aggravent et dans quelle mesure.

6.10 Problèmes résolus

Problème résolu 6.1. Le test t apparié calculé à la main

Énoncé. Chez 10 patients hypertendus, la PAS a été mesurée avant et après 4 semaines de traitement par un inhibiteur de l'enzyme de conversion (tableau 6.2). La PAS a-t-elle diminué de façon significative ? Résolvez selon le format a-g, avec α = 0,05.

Solution. (a) Hypothèse de recherche : le traitement de 4 semaines modifie la PAS des patients hypertendus. (b) H0 : la moyenne des différences PAS avant - PAS après est nulle dans la population, μD = 0 ; H1 : μD ≠ 0 (bilatéral). (c) Méthode : les mêmes patients mesurés deux fois, donc des données appariées, variable numérique ; test t pour échantillons appariés, à condition que les différences soient approximativement normales. (d) Calcul : les différences sont 12, 6, 12, 5, 8, 3, 15, 4, 9, 4 ; somme 78, moyenne mD = 7,8 mmHg. La somme des carrés des écarts à la moyenne est de 151,60 (tableau 6.2), donc $s_D{}^2 = \frac{151{,}60}{9} = 16{,}84$ et sD = 4,10 mmHg. ESD = $\frac{4{,}10}{\sqrt{10}} = 1{,}30$ mmHg. $t = \frac{7{,}8}{1{,}30} = 6{,}01$, ddl = 9. Test de Shapiro-Wilk sur les différences : p = 0,347, la normalité est acceptable. (e) Résultat : t critique (9 ; 0,05) = 2,262 < 6,01 ; p = 0,0002. IC à 95% de la baisse moyenne : 7,8 ± 2,262 × 1,30 = 4,86-10,74 mmHg. dz = $\frac{7{,}8}{4{,}10} = 1{,}90$. Graphique adapté : lignes individuelles avant-après, comme dans la figure 6.1 A. (f) Conclusion statistique : p < 0,001, on rejette H0 ; la différence est extrêmement significative. (g) Conclusion de recherche : après 4 semaines de traitement, la PAS a diminué en moyenne de 7,8 mmHg (IC à 95% 4,9-10,7 mmHg) ; les 10 patients ont tous eu une baisse, comprise entre 3 et 15 mmHg. Limite principale : l'absence de groupe témoin ; une partie de la baisse peut refléter la régression vers la moyenne (les patients ont été inclus avec des valeurs élevées) ou l'accoutumance à la mesure.

Vérification avec le logiciel : dans Google Sheets, =T.TEST(B2:B11;C2:C11;2;1) donne 0,0002 (type 1 = apparié). Si l'on utilisait à tort le type 2 (échantillons indépendants), le résultat serait p = 0,008 ; l'analyse ignorerait le fait que les valeurs proviennent des mêmes patients.

Problème résolu 6.2. Le test t pour échantillons indépendants calculé à la main

Énoncé. La glycémie à jeun (mg/dL) après 3 mois de traitement a été mesurée chez 8 patients traités par metformine et régime et chez 10 patients traités par régime seul (données fictives). Groupe 1 (metformine + régime) : 112, 118, 105, 120, 110, 115, 108, 124. Groupe 2 (régime) : 125, 130, 118, 135, 128, 122, 140, 126, 132, 119. La glycémie moyenne diffère-t-elle entre les groupes ?

Solution. (a) Hypothèse de recherche : la glycémie à jeun à 3 mois diffère entre les patients traités par metformine et régime et ceux traités par régime seul. (b) H0 : µ1 = µ2 ; H1 : µ1 ≠ µ2. (c) Deux groupes indépendants, variable numérique : test t pour échantillons indépendants ; nous vérifions la normalité dans chaque groupe (Shapiro-Wilk p = 0,965 et p = 0,932) et l'égalité des variances.

(d) Calcul. Groupe 1 : somme 912, m1 = 114,0 ; les écarts à la moyenne sont -2, 4, -9, 6, -4, 1, -6, 10, et la somme de leurs carrés est 4 + 16 + 81 + 36 + 16 + 1 + 36 + 100 = 290 ; $s_1{}^2 = \frac{290}{7} = 41{,}43$; s1 = 6,44. Groupe 2 : somme 1275, m2 = 127,5 ; les écarts sont -2,5 ; 2,5 ; -9,5 ; 7,5 ; 0,5 ; -5,5 ; 12,5 ; -1,5 ; 4,5 ; -8,5, avec une somme des carrés de 440,5 ; $s_2{}^2 = \frac{440{,}5}{9} = 48{,}94$; s2 = 7,00. Le rapport des variances est de 48,94 / 41,43 = 1,18, et le test de Levene donne p = 0,887 : les variances peuvent être considérées comme égales.

Variance commune : $s^2_{\text{commune}} = \frac{290+440{,}5}{8+10-2} = \frac{730{,}5}{16} = 45{,}66$; scommune = 6,76. ES = $\sqrt{\left[45{,}66 \times \left(\frac{1}{8} + \frac{1}{10}\right)\right]} = \sqrt{45{,}66 \times 0{,}225} = \sqrt{10{,}27} = 3{,}205$. $t = \frac{127{,}5-114{,}0}{3{,}205} = \frac{13{,}5}{3{,}205} = 4{,}21$, ddl = 16.

(e) t critique (16 ; 0,05) = 2,120 < 4,21 ; p = 0,0007. IC à 95% de la différence : 13,5 ± 2,120 × 3,205 = 13,5 ± 6,79, soit 6,71-20,29 mg/dL. d de Cohen = 13,5 / 6,76 = 2,00. Version de Welch, à titre de comparaison : ES = $\sqrt{\frac{41{,}43}{8} + \frac{48{,}94}{10}} = \sqrt{5{,}179 + 4{,}894} = \sqrt{10{,}07} = 3{,}174$; t = 4,25 ; ddl = 15,6 ; p = 0,0006, pratiquement le même résultat. (f) p < 0,001, on rejette H0 ; la différence est extrêmement significative. (g) À 3 mois, les patients traités par metformine et régime ont une glycémie à jeun inférieure en moyenne de 13,5 mg/dL à celle des patients traités par régime seul (IC à 95% 6,7-20,3 mg/dL). Dans une étude réelle, il faudrait vérifier si les groupes étaient comparables à l'inclusion (idéalement grâce à la randomisation) ; sinon, la différence peut refléter la sélection des patients.

Problème résolu 6.3. Le tableau d'ANOVA et le test de Tukey à la main

Énoncé. Score de douleur (0-10) 2 heures après l'administration de trois antalgiques, chez 5 patients par groupe (données fictives) : A (paracétamol) 6, 5, 7, 6, 6 ; B (ibuprofène) 4, 5, 3, 4, 4 ; C (association) 3, 2, 3, 2, 5. Le score moyen de douleur diffère-t-il entre les antalgiques ? Si oui, lesquels diffèrent entre eux ?

Solution. (a) Hypothèse de recherche : l'efficacité antalgique (score de douleur à 2 heures) diffère entre les trois traitements. (b) H0 : µA = µB = µC ; H1 : au moins une moyenne diffère. (c) Trois groupes indépendants, variable numérique : ANOVA à un facteur, suivie du test de Tukey si F est significatif. Conditions : l'indépendance est assurée par le schéma de l'étude ; la normalité est difficile à évaluer avec 5 valeurs par groupe (en pratique, avec des groupes aussi petits, on rapporterait aussi le test de Kruskal-Wallis) ; homogénéité des variances : $s^2A = \frac{2}{4} = 0{,}5$, $s^2B = \frac{2}{4} = 0{,}5$, $s^2C = \frac{6}{4} = 1{,}5$; test de Levene p = 0,557.

(d) Calcul. Les moyennes : $m_A = \frac{30}{5} = 6$, $m_B = \frac{20}{5} = 4$, $m_C = \frac{15}{5} = 3$; moyenne générale $M = \frac{65}{15} = 4{,}33$. SS inter = 5 × [(6 - 4,33)² + (4 - 4,33)² + (3 - 4,33)²] = 23,33, ddl = 2, MS

inter = 11,67. SS intra = 2 + 2 + 6 = 10,00 (de façon équivalente : 4 × 0,5 + 4 × 0,5 + 4 × 1,5), ddl = 12, MS intra = 0,833. $F = \frac{11,67}{0,833} = 14,0$. Le tableau complet est le tableau 6.4.

(e) F critique (2 ; 12 ; 0,05) = 3,89 < 14,0 ; p = 0,0007 ; $\eta^2 = 0,70$. Test de Tukey : la valeur de q pour k = 3 groupes et 12 degrés de liberté résiduels, à α = 0,05, est de 3,77 (tables de l'étendue studentisée). HSD $= 3,77 \times \sqrt{\frac{0,833}{5}} = 3,77 \times 0,408 = 1,54$. Les différences : A - B = 2 > 1,54, significative (p Tukey = 0,012) ; A - C = 3 > 1,54, significative (p = 0,0006) ; B - C = 1 < 1,54, non significative (p = 0,23). (f) On rejette H0 globale ; les différences sont extrêmement significatives. (g) L'ibuprofène et l'association réduisent davantage la douleur que le paracétamol (de 2 et de 3 points respectivement sur l'échelle 0-10), et la différence de 1 point entre l'association et l'ibuprofène n'a pas pu être mise en évidence avec 5 patients par groupe. Sur le plan clinique, une différence de 2 points sur une échelle de douleur de 0-10 est habituellement considérée comme perceptible par le patient.

Problème résolu 6.4. Le test apparié à partir des statistiques résumées (mg2026)

Énoncé. Entre l'inclusion et 3 mois, les différences individuelles de glycémie à jeun (glicemie0 - glicemie3) chez les 160 patients de mg2026 ont une moyenne de 6,08 mg/dL et un écart-type de 9,35 mg/dL. Testez si la glycémie s'est modifiée et calculez l'IC à 95% et dz.

Solution. H0 : μD = 0 ; H1 : μD ≠ 0. ESD $= \frac{9,35}{\sqrt{160}} = \frac{9,35}{12,649} = 0,739$ mg/dL. $t = \frac{6,08}{0,739} = 8,22$, ddl = 159 ; t critique 1,975 ; p < 0,001. IC à 95% : 6,08 ± 1,975 × 0,739 = 6,08 ± 1,46, soit 4,62-7,54 mg/dL. dz $= \frac{6,08}{9,35} = 0,65$ (effet moyen à important). Normalité des différences : Shapiro-Wilk p = 0,133. Conclusion : la glycémie a diminué de façon extrêmement significative au cours des 3 premiers mois, en moyenne d'environ 6 mg/dL (IC à 95% 4,6-7,5). Comparez avec la baisse à 6 luni (10,19 mg/dL, dz = 0,88) : environ trois cinquièmes de la baisse totale se sont produits au cours des trois premiers mois. Remarquez que le calcul nécessite l'écart-type des différences ; les écarts-types de la glycémie à chaque temps (15,76 à l'inclusion) ne suffisent pas.

Problème résolu 6.5. Le test pour échantillons indépendants à partir des statistiques résumées (mg2026)

Énoncé. Dans mg2026, l'âge des femmes est de 55,42 ± 8,95 ans (n = 86) et celui des hommes de 57,31 ± 8,50 ans (n = 74). L'âge moyen diffère-t-il entre les sexes ? Pourquoi cela compterait-il pour d'autres comparaisons entre les sexes ?

Solution. H0 : μF = μM ; H1 : μF ≠ μM. Conditions : Shapiro-Wilk p = 0,804 (F) et p = 0,311 (M) ; Levene p = 0,617. Variance commune : $s^2_{\text{commune}} = \frac{85 \times 8,95^2 + 73 \times 8,50^2}{158} = \frac{85 \times 80,10 + 73 \times 72,25}{158} = \frac{6808,5 + 5274,3}{158} = 76,47$; scommune = 8,74 ans. ES $= 8,74 \times \sqrt{\frac{1}{86} + \frac{1}{74}} = 8,74 \times 0,1586 = 1,387$ ans. $t = \frac{57,31 - 55,42}{1,387} = \frac{1,89}{1,387} = 1,36$, ddl = 158 ; t critique 1,975 ; p = 0,17 (version de Welch : t = 1,37, ddl = 156,4, p = 0,173). IC à 95% de la différence : -0,84 à 4,62 ans. $d = \frac{1,89}{8,74} = 0,22$.

Conclusion : on ne rejette pas H0 ; l'âge moyen ne diffère pas significativement entre les sexes. L'intervalle de confiance admet toutefois que les hommes soient plus âgés que les femmes, de 4,6 ans au plus environ. Pourquoi cela compte-t-il ? Parce que la PAS augmente avec l'âge (dans mg2026, d'environ 0,4 mmHg par an, chapitre 9). Une différence d'âge entre les sexes pourrait masquer ou créer une différence de PAS entre les sexes. Ici, la différence d'âge est faible et non significative, mais dans une étude observationnelle la comparabilité des groupes vis-à-vis

des facteurs de confusion doit être vérifiée et, si nécessaire, contrôlée par régression multiple ou par ANCOVA.

Problème résolu 6.6. Le tableau d'ANOVA reconstitué à partir de n, des moyennes et des ÉT (mg2026)

Énoncé. Un article rapporte la baisse de la glycémie à 6 mois dans trois bras : A 3,0 ± 8,8 mg/dL$(n = 53)$, B 11,1 ± 11,2 mg/dL$(n = 53)$, C 16,3 ± 10,6 mg/dL$(n = 54)$. Reconstituez le tableau d'ANOVA.

Solution. N = 160, k = 3. La moyenne générale est la moyenne pondérée : $M = \frac{53\times3,0+53\times11,1+54\times16,3}{160} = \frac{159,0+588,3+880,2}{160} = \frac{1627,5}{160} = 10,17$ mg/dL (avec les moyennes non arrondies, 10,19).

SS inter = 53 × (3,0 - 10,17)² + 53 × (11,1 - 10,17)² + 54 × (16,3 - 10,17)² = 53 × 51,41 + 53 × 0,86 + 54 × 37,58 = 2724,7 + 45,8 + 2029,2 = 4799,7 ; ddl = 2 ; MS inter = 2399,9.

SS intra = 52 × 8,8² + 52 × 11,2² + 53 × 10,6² = 52 × 77,44 + 52 × 125,44 + 53 × 112,36 = 4026,9 + 6522,9 + 5955,1 = 16504,8 ; ddl = 157 ; MS intra = 105,1.

$F = \frac{2399,9}{105,1} = 22,8$; p < 0,001. $\eta^2 = \frac{4799,7}{4799,7+16504,8} = 0,225$.

Interprétation. Le résultat reconstitué ($F = 22,8$, $\eta^2 = 0,225$) est très proche de celui calculé par le logiciel sur les données individuelles ($F = 22,93$, $\eta^2 = 0,226$, tableau 6.5) ; les petites différences proviennent de l'arrondi des moyennes et des écarts-types à une décimale. Cette technique est utile pour la lecture critique des articles : si un résumé rapporte les moyennes, les ÉT et les n par groupe, vous pouvez vérifier l'ordre de grandeur de F et de la valeur p rapportée.

Problème résolu 6.7. La PAS selon le statut tabagique : lire un résultat d'ANOVA

Énoncé. Dans mg2026, la PAS à l'inclusion selon les catégories de tabagisme est la suivante : anciens fumeurs 127,55 ± 14,39 mmHg$(n = 33)$, fumeurs 137,04 ± 14,63 mmHg$(n = 53)$, non-fumeurs 128,50 ± 11,76 mmHg$(n = 74)$. Le logiciel rapporte : SS inter = 2785,6, SS intra = 27846,6, Levene p = 0,114, Shapiro-Wilk p entre 0,23 et 0,86 selon les groupes ; Tukey : fumeurs - anciens fumeurs 9,49 mmHg (p = 0,0045), fumeurs - non-fumeurs 8,54 mmHg (p = 0,0014), non-fumeurs - anciens fumeurs 0,95 mmHg (p = 0,94). Complétez le tableau d'ANOVA et interprétez.

Solution. ddl inter = 3 - 1 = 2 ; ddl intra = 160 - 3 = 157. MS inter = 2785,6 / 2 = 1392,8 ; MS intra = 27846,6 / 157 = 177,4. $F = \frac{1392,8}{177,4} = 7,85$; p = 0,0006 (F critique 3,05). $\eta^2 = \frac{2785,6}{2785,6+27846,6} = \frac{2785,6}{30632,2} = 0,091$, un effet moyen à important. Les conditions sont remplies (Levene p > 0,05 ; normalité acceptable ; rapport $\frac{\text{ÉT } 14,63}{11,76} = 1,24$).

Conclusion statistique : on rejette H0, la PAS moyenne diffère de façon très significative entre les catégories de tabagisme. Le test de Tukey montre que les fumeurs actuels ont une PAS supérieure d'environ 9 mmHg à celle des anciens fumeurs et des non-fumeurs, et que les anciens fumeurs ne diffèrent pas des non-fumeurs. La conclusion de recherche doit être formulée avec prudence : pour cette question, l'étude est une analyse transversale ; les fumeurs peuvent aussi différer par d'autres facteurs (âge, IMC, consommation d'alcool). Au chapitre 9, vous verrez qu'après ajustement sur l'IMC, l'âge et le sexe en régression multiple, la différence fumeurs - anciens fumeurs reste significative (6,74 mmHg, p = 0,012), mais plus faible. Le résultat est compatible avec l'effet presseur connu du tabagisme actif et avec sa réversibilité après l'arrêt, sans pour autant le démontrer.

6.11 Exercices proposés

1. Chez 8 patients atteints de diabète de type 2, l'HbA1c (%) a été mesurée avant et après 3 mois d'éducation thérapeutique : avant 7,8 ; 8,2 ; 7,5 ; 9,0 ; 8,4 ; 7,9 ; 8,8 ; 8,1 ; après 7,2 ; 7,9 ; 7,4 ; 8,1 ; 7,8 ; 7,6 ; 8,0 ; 7,7. Appliquez le test t apparié ($\alpha = 0,05$) : calculez les différences, mD, sD, t, l'IC à 95% et dz.
2. Une étude compare la PAS dans deux groupes indépendants : groupe 1, n = 25, m = 142 mmHg, s = 12 ; groupe 2, n = 30, m = 135 mmHg, s = 14. Calculez le test t avec variance commune, la valeur p, l'IC à 95% et d. Commentez le résultat par rapport au seuil de 0,05.
3. Deux groupes : groupe 1, n = 12, m = 4,8, s = 1,1 ; groupe 2, n = 40, m = 6,1, s = 3,2 (une concentration sérique, en unités arbitraires). Calculez le test t dans la version de Student et dans la version de Welch. Expliquez pourquoi les conclusions diffèrent.
4. Trois groupes de 10 patients chacun ont une PAS moyenne de 120, 128 et 131 mmHg, avec des écarts-types de 10, 12 et 11 mmHg. Construisez le tableau d'ANOVA, calculez F et $\hat{e}ta^2$, et décidez à $\alpha = 0,05$ (F critique $(2 ; 27) = 3,35$).
5. Complétez le tableau d'ANOVA : 4 groupes, N = 40 patients, SS inter = 360, SS intra = 1440. Calculez les ddl, les MS, F et $\hat{e}ta^2$, et décidez à $\alpha = 0,05$ (F critique $(3 ; 36) = 2,87$).
6. Dans l'étude sur les troubles de l'interaction intestin-cerveau [71], l'IMC était de 28,15 ± 6,49 kg/m² chez 52 cas et de 24,47 ± 4,60 kg/m² chez 84 témoins. En utilisant uniquement ces valeurs, calculez le test t de Welch et le d de Cohen. Le résultat est-il compatible avec la valeur p rapportée (p = 0,001) ?
7. Dans l'étude sur la COVID-19 chez les patients diabétiques [8], la durée d'hospitalisation était de 11,6 ± 7,0 jours chez 224 patients atteints de diabète de type 2 et de 9,1 ± 5,8 jours chez 62 patients atteints de diabète de type 1. Calculez, à titre d'exercice, le test t de Welch à partir de ces valeurs. Que vous apprend le rapport entre l'ÉT et la moyenne sur la forme de la distribution, et quel test préféreriez-vous sur les données individuelles ?
8. Dans mg2026, les différences individuelles tas0 - tas6 ont une moyenne de 4,78 mmHg et un écart-type de 9,05 mmHg (n = 160). Calculez t, l'IC à 95% de la différence moyenne et dz. La baisse est-elle cliniquement pertinente ?
9. Pour chaque situation, nommez le test paramétrique adapté et son alternative non paramétrique : (a) la fréquence cardiaque de patients atteints de BPCO avant et après un traitement bronchodilatateur ; (b) la taille des étudiants selon le sexe ; (c) la glycémie à 0, 3 et 6 mois chez les mêmes patients ; (d) la glycémie à 6 mois chez des patients assurés et non assurés ; (e) l'indice de perméabilité bronchique chez des patients traités par trois médicaments différents ; (f) l'IMC moyen d'un groupe comparé à la valeur de 25 kg/m².
10. Pour les données de l'exercice 4, calculez la différence minimale significative de Tukey ($q(3 ; 27 ; 0,05) = 3,51$) et vérifiez si une paire de groupes diffère. Le résultat est-il cohérent avec l'ANOVA ?
11. Un résumé rapporte, pour un marqueur mesuré chez les mêmes patients : $8061,9 \pm 3684,80$ ng/mL avant et 15380 ± 5887 ng/mL après traitement, $p < 0,001$ [86]. Expliquez pourquoi on ne peut pas recalculer le test t apparié à partir de ces chiffres et quelle information supplémentaire serait nécessaire.

Pièges fréquents

- Des données appariées analysées avec le test pour échantillons indépendants perdent de la puissance ; des groupes indépendants analysés comme appariés produisent des résultats dénués de sens.
- Pour le test apparié, la normalité se vérifie sur les différences, et non sur chaque série séparément.

- Dans le test apparié, le nombre de degrés de liberté est n - 1, où n est le nombre de paires, et non le nombre total de mesures.
- Avec des groupes inégaux et des variances inégales, le test de Student peut donner des valeurs p erronées ; on utilise le test de Welch.
- Trois groupes ou plus ne se comparent pas par des tests t multiples non corrigés ; on fait une ANOVA puis un test post hoc.
- Un F significatif ne dit pas quels groupes diffèrent ; on ne fait pas de tests post hoc après un F non significatif, à la recherche d'une paire significative.
- Les mesures répétées chez les mêmes patients exigent une ANOVA à mesures répétées (avec vérification de la sphéricité), et non une ANOVA à un facteur.
- Un ÉT proche de la moyenne pour une variable positive signale une asymétrie ; la moyenne ± ÉT et le test t peuvent alors être inadaptés.

À retenir

- Test t apparié : $t = \frac{m_D}{\frac{s_D}{\sqrt{n}}}$, ddl = n - 1 ; c'est le test t pour une moyenne appliqué aux différences ; dz = $\frac{m_D}{s_D}$.
- Test t de Student pour échantillons indépendants : variance commune $s^2_{\text{commune}} = \frac{(n_1-1)\,{s_1}^2+(n_2-1)\,{s_2}^2}{n_1+n_2-2}$, ddl = n1 + n2 - 2 ; la version de Welch ne suppose pas l'égalité des variances ; le test de Levene vérifie l'égalité des variances.
- L'ANOVA compare la variation entre les groupes à la variation à l'intérieur des groupes : F = MS inter / MS intra, avec ddl k - 1 et N - k ; SS total = SS inter + SS intra.
- η^2 = SS inter / SS total est la taille d'effet en ANOVA ; le d de Cohen est la taille d'effet pour deux moyennes.
- Conditions de l'ANOVA : indépendance, normalité dans les groupes, homogénéité des variances ; alternatives : ANOVA de Welch avec Games-Howell, Kruskal-Wallis.
- Post hoc : Tukey (toutes les paires, variances égales), Bonferroni ou Holm (peu de comparaisons), Games-Howell (variances inégales), Dunnett (chaque groupe comparé au témoin).
- Dans mg2026 : t apparié glycémie 0-6 mois t(159) = 11,14, p < 0,001 ; PAS selon le sexe p = 0,81 ; ANOVA baisse de la glycémie selon le bras F(2 ; 157) = 22,93, p < 0,001, $\eta^2 = 0{,}226$; ANOVA à mesures répétées F = 64,97, p < 0,001.

Questions d'autoévaluation

1. (QRU) La glycémie à jeun a été mesurée à l'inclusion et à 6 mois chez les mêmes 160 patients. Le test paramétrique adapté pour comparer les deux temps est :
    a) le test t pour échantillons indépendants
    b) le test t pour échantillons appariés
    c) l'ANOVA à un facteur
    d) le test du khi-deux
    e) le test z pour une proportion
2. (QRU) Dans un test t apparié portant sur 25 patients, le nombre de degrés de liberté est :
    a) 48
    b) 50
    c) 24
    d) 25
    e) 23

3. (QRU) Dans une ANOVA à un facteur avec 4 groupes totalisant 60 patients, les degrés de liberté du numérateur et du dénominateur de la statistique F sont :
    a) 4 et 60
    b) 3 et 56
    c) 3 et 59
    d) 4 et 56
    e) 59 et 3
4. (QRU) Le test de Levene donne p = 0,003 lors de la comparaison de deux groupes indépendants de 15 et 60 patients. La version adaptée du test t est :
    a) le test t apparié
    b) le test t de Student avec variance commune
    c) le test t de Welch
    d) le test z
    e) le test t pour une seule moyenne
5. (QRU) Dans une ANOVA, SS inter-groupes = 300 et SS total = 1200. L'êta² vaut :
    a) 0,20
    b) 0,25
    c) 0,33
    d) 4,00
    e) il ne peut pas être calculé sans les degrés de liberté
6. (QRM) À propos du test t pour échantillons appariés, les affirmations correctes sont :
    a) il travaille sur les différences individuelles entre les deux mesures
    b) la normalité se vérifie sur les différences
    c) il a en général plus de puissance que le test pour échantillons indépendants appliqué aux mêmes données appariées
    d) il exige que les deux séries aient des variances égales
    e) son alternative non paramétrique est le test de Wilcoxon des rangs signés
7. (QRM) Les conditions d'application de l'ANOVA à un facteur sont :
    a) l'indépendance des observations
    b) une distribution approximativement normale dans chaque groupe
    c) l'homogénéité des variances
    d) des groupes obligatoirement de même effectif
    e) au moins 30 patients dans chaque groupe
8. (QRM) Dans mg2026, l'ANOVA pour la baisse de la glycémie selon le bras donne F(2 ; 157) = 22,93, p < 0,001, $\eta^2 = 0{,}226$, et le test de Tukey montre des différences significatives pour les trois paires. Les affirmations correctes sont :
    a) on rejette l'hypothèse selon laquelle les trois moyennes sont égales
    b) environ 23% de la variabilité de la baisse de la glycémie est expliquée par le bras
    c) l'ANOVA à elle seule montre que le bras C diffère du bras B
    d) le test de Tukey maintient le risque global d'erreur de type I au niveau α
    e) le résultat est dû exclusivement au grand nombre de patients
9. (QRM) Les choix adaptés lorsque les variances des groupes diffèrent beaucoup sont :
    a) le test t de Welch pour deux groupes

b) l'ANOVA de Welch pour trois groupes ou plus

c) le test post hoc de Games-Howell

d) le test t de Student avec variance commune

e) le test de Tukey classique, sans aucune vérification

10. (QRM) À propos de l'ANOVA à mesures répétées, les affirmations correctes sont :

a) elle s'applique lorsque les mêmes patients sont mesurés à trois temps ou plus

b) la sphéricité se vérifie par le test de Mauchly

c) la correction de Greenhouse-Geisser modifie les degrés de liberté du test F

d) son alternative non paramétrique est le test de Kruskal-Wallis

e) pour deux temps, le résultat est équivalent à celui du test t apparié

Chapitre 7. Tests statistiques non paramétriques usuels : U de Mann-Whitney, Wilcoxon, Kruskal-Wallis, Friedman et test du khi-deux

Objectifs

À la fin de ce chapitre, l'étudiant sera capable de :

- reconnaître les situations dans lesquelles les conditions des tests paramétriques ne sont pas remplies (données ordinales, asymétrie marquée, petits échantillons, valeurs extrêmes) et choisir l'alternative non paramétrique adaptée ;
- attribuer des rangs à une série de valeurs, y compris en présence d'ex aequo, et vérifier la somme des rangs ;
- calculer à la main la statistique U de Mann-Whitney, la statistique T de Wilcoxon, la statistique H de Kruskal-Wallis et la statistique de Friedman sur de petits échantillons, et les comparer aux valeurs critiques ;
- interpréter correctement les hypothèses des tests sur les rangs (distributions, médianes) et rapporter une taille d'effet (corrélation bisériale des rangs, epsilon carré, W de Kendall) ;
- construire un tableau de contingence, calculer les effectifs attendus, la statistique du khi-deux et les degrés de liberté, et vérifier les conditions d'application ;
- choisir entre le test du khi-deux, la correction de Yates, le test exact de Fisher, le test de McNemar, le test de tendance et le test d'ajustement ;
- calculer et interpréter le coefficient phi et le V de Cramér, en comprenant que la valeur p ne mesure pas la force de l'association ;
- lire de façon critique la section méthodes et les résultats d'articles publiés par des collègues de la discipline ayant utilisé ces tests.

7.1 Quand on ne peut plus utiliser les tests paramétriques

Au chapitre 6, nous avons travaillé avec le test t et l'ANOVA. Tous deux comparent des moyennes et reposent sur quelques hypothèses : la variable est numérique, la distribution dans chaque groupe est approximativement normale (ou l'échantillon est assez grand pour que la moyenne se comporte de façon normale), et les variances des groupes ne diffèrent pas trop. Dans la pratique médicale, ces hypothèses tombent plus souvent que nous ne le souhaiterions. Le patient ne répond pas à la question "dans quelle mesure êtes-vous satisfait ?" par un nombre mesuré en milligrammes, il coche une case de 1 à 5. La durée d'hospitalisation présente une longue queue à droite, car un seul patient ayant présenté des complications peut rester 40 jours alors que les autres restent 5 jours. Et dans une étude pilote portant sur 12 patients, nous n'avons aucun moyen de vérifier la normalité de façon convaincante.

Les tests non paramétriques ont été conçus précisément pour ces situations. Leur idée commune est simple : au lieu de travailler sur les valeurs elles-mêmes, nous travaillons sur leur position dans un classement, c'est-à-dire sur les rangs. Un patient resté 40 jours à l'hôpital ne tire plus la moyenne vers le haut de tout son poids ; il devient simplement "le dernier du classement". Nous perdons une partie de l'information, mais nous gagnons en robustesse.

7.1.1 Quatre situations typiques

La première situation est celle des **données ordinales**. L'échelle de Likert de satisfaction, le stade NYHA de l'insuffisance cardiaque, le stade de fibrose hépatique F0-F4, le score mMRC de la dyspnée ou le niveau d'études ont des catégories ordonnées, mais la distance entre les catégories n'est pas nécessairement égale. Passer de "insatisfait" à "neutre" n'équivaut pas obligatoirement à passer de "satisfait" à "très satisfait". La moyenne de tels codes a un sens discutable, et le test t suppose justement que les différences entre valeurs ont un sens arithmétique. Pour les données ordinales, les tests sur les rangs sont le choix naturel.

La deuxième situation est l'**asymétrie marquée** d'une variable numérique. Dans le jeu de données mg2026, les triglycérides ont une asymétrie de 0,95 et le test de Shapiro-Wilk donne $p < 0,001$, et l'indice de masse corporelle (IMC) a une asymétrie de 1,42 et un excès d'aplatissement de 3,68. Lorsque les groupes sont grands, le test t résiste assez bien à l'asymétrie, grâce au théorème central limite. Lorsque les groupes sont petits ou que l'asymétrie est forte, la moyenne devient un résumé peu représentatif, et le test sur les rangs est plus sûr. Une alternative abordée au chapitre 3 est la transformation logarithmique, qui ramène l'asymétrie des triglycérides à -0,28 avec SW $p = 0,200$.

La troisième situation est le **petit échantillon**. Avec 6 à 8 patients par groupe, le test de normalité a une puissance très faible : un résultat "$p > 0,05$" au test de Shapiro-Wilk ne prouve pas la normalité, il montre seulement que nous n'avons pas assez de données pour l'infirmer. Dans ce type d'études, de nombreux auteurs choisissent d'emblée un test non paramétrique, et ce choix est raisonnable s'il a été prévu dans le protocole.

La quatrième situation est la présence de **valeurs extrêmes** (angl. outliers) qui ne sont pas des erreurs de mesure. Un patient dont les triglycérides sont à 442 mg/dL est réel et ne doit pas être supprimé simplement parce qu'il fausse la moyenne. Les tests sur les rangs le conservent, mais limitent son influence : que sa valeur soit de 442 ou de 4420, son rang reste le même.

7.1.2 Ce que signifie "non paramétrique" et ce que cela ne signifie pas

Le terme "non paramétrique" vient de ce que ces tests ne supposent pas une forme particulière de la distribution (par exemple, la loi normale de paramètres μ et σ). Cela ne signifie pas qu'ils n'ont aucune condition. Tous supposent des observations indépendantes (ou des paires indépendantes entre elles), et certains supposent que les distributions comparées ont la même forme et ne diffèrent que par leur position, si l'on veut interpréter le résultat comme une différence de médianes. Le test du khi-deux, pour sa part, exige des effectifs attendus suffisamment grands.

Une confusion fréquente à l'examen est l'idée que le test non paramétrique "compare les médianes" en toutes circonstances. Plus exactement, le test de Mann-Whitney vérifie si les valeurs d'un groupe tendent à être plus grandes que celles de l'autre groupe. Si les deux distributions ont la même forme et la même dispersion, cela équivaut à une différence entre les médianes. Si les formes diffèrent beaucoup, le test peut être significatif même avec des médianes égales. C'est pourquoi, dans le compte rendu, nous disons "les distributions diffèrent" ou "les valeurs étaient plus élevées dans le groupe C", et nous donnons les médianes et les écarts interquartiles à titre descriptif. Le tableau 7.1 met en regard les tests paramétriques du chapitre 6 et leurs équivalents non paramétriques.

Tableau 7.1. Correspondance entre les tests paramétriques et les tests non paramétriques.

Situation	Test paramétrique	Test non paramétrique	Ce que l'on compare dans la version non paramétrique
2 groupes indépendants	test t non apparié (Welch)	U de Mann-Whitney (somme des rangs de	les rangs des deux groupes

Situation	Test paramétrique	Test non paramétrique	Ce que l'on compare dans la version non paramétrique
		Wilcoxon)	
2 mesures chez les mêmes sujets	test t apparié	Wilcoxon (signes et rangs) ; test des signes	les rangs des différences, avec leur signe
≥ 3 groupes indépendants	ANOVA à un facteur	Kruskal-Wallis, puis Dunn	les rangs moyens des groupes
≥ 3 mesures chez les mêmes sujets	ANOVA à mesures répétées	Friedman, puis Wilcoxon avec correction	les rangs à l'intérieur de chaque sujet
association de deux variables	corrélation de Pearson	corrélation de Spearman	les rangs des deux variables
deux variables qualitatives	(il n'existe pas de version paramétrique)	khi-deux, Fisher, McNemar	les effectifs observés comparés aux effectifs attendus

7.1.3 L'efficacité relative des tests non paramétriques

Le prix payé pour la robustesse est une perte de puissance lorsque les données sont réellement normales. Cette perte se mesure par l'efficacité relative asymptotique : le rapport entre le nombre de sujets dont a besoin le test paramétrique et le nombre dont a besoin le test non paramétrique pour une même puissance. Pour le test de Mann-Whitney comparé au test t, sur des données parfaitement normales, l'efficacité est de $3/\pi$, soit environ 0,955 (95,5%) [53,63]. Concrètement, ce que le test t fait avec 95 patients, le test de Mann-Whitney le fait avec environ 100. La perte est faible. Pour des distributions à queues épaisses (beaucoup de valeurs extrêmes), l'efficacité du test sur les rangs peut dépasser 1, c'est-à-dire que le test non paramétrique devient plus puissant que le test t, et pour aucune distribution continue elle ne descend en dessous d'environ 0,86. Le test de Wilcoxon pour données appariées a la même efficacité de 95,5% par rapport au test t apparié.

Le test des signes, qui ne conserve que le sens de la différence (plus ou moins) et en ignore l'ampleur, est beaucoup plus coûteux : son efficacité sur des données normales est de $2/\pi$, environ 0,64. Nous verrons à la section 7.4 un exemple tiré de mg2026 dans lequel la même question obtient p = 0,043 au test t apparié, p = 0,052 au test de Wilcoxon et p = 0,092 au test des signes, exactement dans l'ordre de leur efficacité.

La conclusion pratique est rassurante. Si vous avez une raison réelle de choisir un test non paramétrique (données ordinales, forte asymétrie avec un petit n, valeurs extrêmes), vous ne perdez presque rien. Si les données sont numériques et approximativement normales, le test paramétrique reste le premier choix, car il estime directement la différence de moyennes avec un intervalle de confiance, ce qui est plus facile à interpréter cliniquement.

7.2 Les rangs : la matière première des tests non paramétriques

7.2.1 Attribution des rangs en présence d'ex aequo

Le rang d'une valeur est sa position dans la série ordonnée par ordre croissant : la plus petite valeur reçoit le rang 1, la suivante le rang 2, et ainsi de suite. Le problème se pose avec les valeurs égales (angl. ties), très fréquentes pour les scores ordinaux. La règle veut que les valeurs égales reçoivent la **moyenne des rangs** qu'elles auraient occupés si elles avaient été légèrement différentes.

Prenons huit scores de douleur sur l'échelle 0-10, notés par des patients après une extraction dentaire : 3, 5, 5, 2, 7, 5, 9, 3. Ordonnons-les : 2, 3, 3, 5, 5, 5, 7, 9. La valeur 2 occupe la position 1, donc le rang 1. Les deux valeurs 3 occupent les positions 2 et 3, chacune reçoit donc (2 + 3)/2 =

2,5. Les trois valeurs 5 occupent les positions 4, 5 et 6, chacune reçoit donc (4 + 5 + 6)/3 = 5. La valeur 7 reçoit le rang 7, et la valeur 9 le rang 8.

Tableau 7.2. Attribution des rangs en présence d'ex aequo (scores de douleur, n = 8).

Valeur originale, dans l'ordre de recueil	**3**	**5**	**5**	**2**	**7**	**5**	**9**	**3**
Rang	2,5	5	5	1	7	5	8	2,5

7.2.2 La vérification : la somme des rangs

Avant tout calcul, vérifiez toujours la somme des rangs. Pour N observations, la somme des rangs 1, 2,..., N est :

$$\frac{N(N+1)}{2}$$

Pour N = 8, la somme doit être de $\frac{8\times9}{2} = 36$. Additionnons : 2,5 + 5 + 5 + 1 + 7 + 5 + 8 + 2,5 = 36. C'est correct. La règle de la moyenne pour les ex aequo conserve la somme, c'est pourquoi la vérification fonctionne aussi en présence d'égalités. À l'examen, une somme des rangs erronée est le premier signe qu'une valeur a été oubliée ou qu'un rang moyen a été mal calculé, et la vérification prend dix secondes.

Les ex aequo ont aussi un effet sur la variabilité des rangs : ils la réduisent légèrement. C'est pourquoi les formules des tests comportent une "correction pour les ex aequo", que les logiciels appliquent automatiquement. Sur les petits échantillons calculés à la main, nous la montrerons une fois, pour que vous sachiez d'où viennent les petites différences entre votre résultat et celui de JASP.

7.3 Le test U de Mann-Whitney

Le test a été proposé sous sa forme avec la statistique U par Mann et Whitney en 1947 [92], à partir du test de la somme des rangs décrit par Wilcoxon deux ans plus tôt [93]. C'est pourquoi certains logiciels le présentent sous le nom de "Wilcoxon rank-sum test" ; il s'agit du même test. Nous l'utilisons pour comparer deux groupes indépendants lorsque la variable est ordinale, ou numérique sans distribution normale.

7.3.1 Les hypothèses

En langage courant, l'hypothèse nulle énonce que les valeurs des deux populations proviennent de la même distribution : si nous choisissons au hasard un patient du groupe 1 et un patient du groupe 2, la probabilité que le premier ait la valeur la plus élevée est de 0,5. L'hypothèse alternative bilatérale énonce que cette probabilité diffère de 0,5, c'est-à-dire que les valeurs d'un groupe tendent à être systématiquement plus élevées.

H0 : P(X1 > X2) = 0,5 (les distributions sont identiques) H1 : P(X1 > X2) ≠ 0,5

Si nous acceptons l'hypothèse selon laquelle les distributions ont la même forme, les hypothèses peuvent s'écrire à l'aide des médianes : H0 : Me1 = Me2, H1 : Me1 ≠ Me2. C'est la formulation utilisée dans les devoirs de la discipline, et elle est acceptable, à condition de connaître l'hypothèse sous-jacente.

7.3.2 Le calcul de U à la main

Les étapes sont les mêmes pour tout petit échantillon :

1. on place toutes les valeurs des deux groupes dans une seule série et on attribue les rangs, avec la moyenne des rangs pour les ex aequo ;
2. on calcule la somme des rangs de chaque groupe, R1 et R2, et on vérifie $R_1 + R_2 = \frac{N(N+1)}{2}$;
3. on calcule $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$ et $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$; vérification : U1 + U2 = n1 x n2 ;
4. la statistique de test est U = le minimum de U1 et U2 ;
5. on compare U à la valeur critique de la table pour n1, n2 et α ; **on rejette H0 si U est inférieur ou égal à la valeur critique** (attention, c'est l'inverse du test t et du khi-deux).

Un exemple : dans une étude pilote, 5 patients opérés ont reçu une application pour téléphone proposant des exercices respiratoires et des rappels pour les antalgiques, et 6 patients ont reçu les soins habituels. Le score de douleur à 24 heures (0-10) était le suivant : application 2, 3, 3, 5, 6 ; témoins 4, 6, 7, 7, 8, 9. La figure 7.1 montre les valeurs et leurs rangs dans l'échantillon commun.

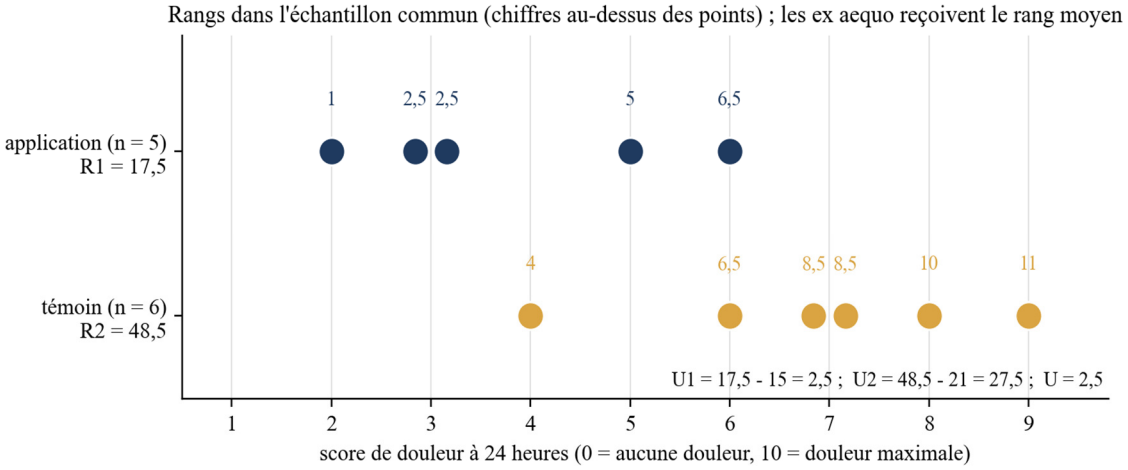


Figure 7.1. Calcul de la statistique U de Mann-Whitney sur un petit exemple : score de douleur à 24 heures chez 5 patients avec application et 6 patients témoins. Les chiffres au-dessus des points sont les rangs dans l'échantillon commun de 11 valeurs ; les ex aequo (3 et 3, 6 et 6, 7 et 7) reçoivent le rang moyen. Données fictives, construites pour le calcul manuel.

La série commune ordonnée est 2, 3, 3, 4, 5, 6, 6, 7, 7, 8, 9 (N = 11). Les rangs : 2 → 1 ; les deux valeurs 3 → 2,5 ; 4 → 4 ; 5 → 5 ; les deux valeurs 6 (une de chaque groupe) → 6,5 ; les deux valeurs 7 → 8,5 ; 8 → 10 ; 9 → 11.

$$R_1(\text{ application }) = 1 + 2{,}5 + 2{,}5 + 5 + 6{,}5 = 17{,}5 R_2(\text{ contrôle }) = 4 + 6{,}5 + 8{,}5 + 8{,}5 + 10 + 11 = 48{,}5$$

Vérification : $17{,}5 + 48{,}5 = 66 = \frac{11 \times 12}{2}$. C'est correct.

$$U_1 = 17{,}5 - \frac{5 \times 6}{2} = 17{,}5 - 15 = 2{,}5 U_2 = 48{,}5 - \frac{6 \times 7}{2} = 48{,}5 - 21 = 27{,}5$$

Vérification : 2,5 + 27,5 = 30 = 5 x 6. La statistique de test est U = 2,5.

Que signifie U, intuitivement ? Si nous formons les 5 x 6 = 30 paires possibles (un patient avec application, un patient témoin), U1 compte combien de fois le patient avec application a le score le plus élevé (les égalités comptent pour moitié). Seules 2,5 des 30 paires "désignent" le groupe avec application comme ayant le plus de douleur. Autrement dit, dans 27,5 des 30 comparaisons, le patient avec application a moins de douleur.

La valeur critique pour n1 = 5, n2 = 6, test bilatéral, α = 0,05, est 3 (tableau 7.15, au début de la section 7.11). Comme U = 2,5 ≤ 3, nous rejetons H0. La valeur p exacte donnée par le logiciel

est de 0,030. Conclusion statistique : la différence entre les distributions du score de douleur est significative au seuil de 0,05. Conclusion de recherche : les patients qui ont utilisé l'application ont rapporté une douleur plus faible à 24 heures (médiane 3 contre 7). S'agissant d'une étude pilote portant sur 11 patients, le résultat doit être confirmé sur un échantillon plus grand.

7.3.3 L'approximation normale

Pour des groupes plus grands (en pratique, les deux groupes au-delà de 10 à 20 sujets), les tables ne sont plus nécessaires. Sous H0, U suit approximativement une distribution normale de moyenne et d'écart-type :

$$\mu_U = \frac{n_1 \times n_2}{2} \sigma_U = \sqrt{\left[\frac{n_1 \times n_2 \times (N+1)}{12}\right]}$$

et la statistique

$$z = \frac{U - \mu_U}{\sigma_U}$$

se compare à 1,96 pour α = 0,05 en bilatéral. S'il existe de nombreux ex aequo, σU est corrigé vers le bas, ce que les logiciels font automatiquement. Sur l'exemple ci-dessus, μU = 15, σU = √(30 x 12/12) = 5,48 (5,44 avec la correction pour les ex aequo), donc $z = \frac{2,5-15}{5,44} = -2,30$, avec un p bilatéral d'environ 0,02. Avec des groupes aussi petits, l'approximation n'est qu'indicative ; c'est la valeur exacte (0,030) qui est correcte.

7.3.4 La taille d'effet : la corrélation bisériale des rangs

La valeur p nous dit à quel point la différence serait surprenante si H0 était vraie, mais pas quelle est son ampleur. Pour le test de Mann-Whitney, la mesure d'effet la plus simple est la corrélation bisériale des rangs (angl. rank-biserial correlation) :

$$r_{\text{rb}} = 1 - \frac{2U}{n_1 \times n_2}$$

où U est le minimum de U1 et U2. Sa valeur va de 0 (aucune différence) à 1 (toutes les valeurs d'un groupe sont supérieures à toutes les valeurs de l'autre). Sur l'exemple de l'application, $r_{\text{rb}} = 1 - \frac{2\times2,5}{30} = 0,83$, un effet très important. Une autre variante, utilisée par certains auteurs, est $r = \frac{z}{\sqrt{N}}$. À titre indicatif, on considère 0,1 comme un effet faible, 0,3 comme moyen et 0,5 comme important [76], mais ces seuils sont des conventions, pas des lois.

7.3.5 Application au jeu de données mg2026 : la satisfaction dans le bras C par rapport au bras A

À 6 mois, les patients de l'étude mg2026 ont noté leur satisfaction à l'égard du programme reçu sur une échelle allant de 1 (pas du tout satisfait) à 5 (très satisfait). Les effectifs par bras figurent dans le tableau 7.3, et la distribution en pourcentage dans la figure 7.2.

Tableau 7.3. Satisfaction à 6 mois selon le bras (nombre de patients), jeu de données mg2026.

Bras	1	2	3	4	5	Total	Médiane (EIQ)
A : conseil	1	13	28	9	2	53	3 (2-3)
B : + application	3	3	14	24	9	53	4 (3-4)

Bras	1	2	3	4	5	Total	Médiane (EIQ)
C : + application et téléconsultation	0	3	12	22	17	54	4 (3-5)
Total	4	19	54	55	28	160	4 (3-4)

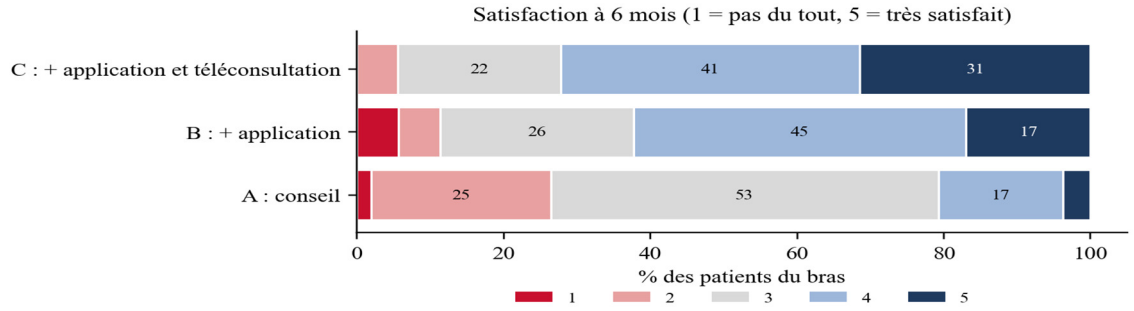


Figure 7.2. Distribution de la satisfaction à 6 mois (1 = pas du tout satisfait, 5 = très satisfait), en pourcentage des patients de chaque bras. Dans le bras C, les notes 4 et 5 prédominent, dans le bras A, la note 3. Données synthétiques (jeu de données mg2026).

La question de recherche : la satisfaction des patients du bras C diffère-t-elle de celle des patients du bras A ? La variable est ordinale, nous utilisons donc le test de Mann-Whitney. Il est inutile de tester la normalité d'une variable à cinq valeurs possibles.

Le résultat donné par le logiciel : les médianes sont de 4 (EIQ 3-5) dans le bras C et de 3 (EIQ 2-3) dans le bras A, U = 2265,5, p < 0,001. Voyons d'où vient ce chiffre. Les sommes des rangs dans l'échantillon commun de 107 patients sont RC = 3750,5 et RA = 2027,5. Alors UC = $3750{,}5 - \frac{54\times55}{2} = 3750{,}5 - 1485 = 2265{,}5$, et UA = 54 x 53 - 2265,5 = 596,5. Les logiciels rapportent tantôt le U d'un groupe, tantôt celui de l'autre ; les deux conduisent à la même valeur p. Approximation normale : $\mu_U = \frac{54\times53}{2} = 1431$, σU = 153,5 (avec la correction pour les nombreux ex aequo), donc $z = \frac{2265{,}5-1431}{153{,}5} = 5{,}44$, bien au-delà de 1,96.

Taille d'effet : $r_{\text{rb}} = 1 - \frac{2\times596{,}5}{2862} = 0{,}58$, un effet important. Interprétation directe : si nous choisissons au hasard un patient du bras C et un patient du bras A, la probabilité que le premier soit plus satisfait (en partageant les égalités par moitié) est de (1 + 0,58)/2 = 0,79.

Conclusion statistique : nous rejetons H0 ; la satisfaction diffère significativement entre les bras C et A (p < 0,001). Conclusion clinique : les patients qui ont bénéficié de l'application et d'une téléconsultation mensuelle ont été plus satisfaits que ceux qui n'ont reçu que le conseil, et la différence est statistiquement significative et, de surcroît, importante. Attention toutefois : la satisfaction n'est pas la même chose que l'efficacité ; elle peut aussi refléter l'attention supplémentaire reçue dans le bras C.

Étude de cas : l'application mobile pour l'asthme, trois tests non paramétriques dans un seul article

Munteanu et ses collaborateurs, parmi lesquels la Dre Frandeș de notre discipline, ont évalué dans une étude pilote contrôlée une application pour téléphone destinée à améliorer l'observance thérapeutique et l'autogestion de la maladie chez 93 patients asthmatiques suivis pendant un an [94]. Le schéma combine des groupes indépendants (traitement standard contre traitement plus application) et des mesures répétées chez les mêmes patients. Les auteurs ont choisi de façon cohérente des tests non paramétriques : le test U de Mann-Whitney pour

> comparer les groupes (p < 0,001), le test de Friedman pour les mesures répétées, avec correction de Bonferroni pour les comparaisons deux à deux (p < 0,001), et le test du khi-deux de Pearson pour un tableau 2 x 2 dans lequel les proportions comparées étaient de 10,30% et 46,30%, avec $\chi^2(1) = 13{,}707$, p < 0,001. La lecture de ce résumé est un bon exercice d'adéquation entre le test et la question : comparaison entre groupes indépendants, comparaison dans le temps chez les mêmes patients, comparaison de proportions. Ce qui aurait apporté une valeur ajoutée : la taille d'effet pour le test de Mann-Whitney (par exemple la corrélation bisériale des rangs) et, pour le tableau 2 x 2, le risque relatif avec son intervalle de confiance, qui montre au clinicien l'ampleur de la différence, en plus de son existence. S'agissant d'une étude pilote, les résultats génèrent des hypothèses pour un essai randomisé de plus grande taille.

7.4 Le test de Wilcoxon pour données appariées et le test des signes

Le test de Wilcoxon des rangs signés (angl. Wilcoxon signed-rank test) [93] est l'alternative non paramétrique du test t apparié. Nous l'utilisons lorsque nous disposons de deux mesures chez les mêmes sujets (avant et après traitement, œil droit et œil gauche, deux méthodes appliquées au même prélèvement) et que les différences ne suivent pas une distribution normale ou que la variable est ordinale.

7.4.1 La logique du test

Pour chaque sujet, nous calculons la différence d = avant - après. Si le traitement n'a pas d'effet, les différences positives et négatives devraient être aussi fréquentes et aussi grandes les unes que les autres. Le test ordonne les valeurs absolues des différences, leur attribue des rangs, puis leur réaffecte leur signe. Sous H0, la somme des rangs positifs (W+) devrait être proche de la somme des rangs négatifs (W-). Les différences nulles sont éliminées (version classique), et n devient le nombre de différences non nulles.

H0 : la distribution des différences est symétrique autour de 0 (médiane des différences = 0)
H1 : médiane des différences ≠ 0

7.4.2 Le calcul à la main

Dix patients en surpoids ont suivi un programme de trois mois de conseil nutritionnel. Le poids (kg) avant et après est donné dans le tableau 7.4.

Tableau 7.4. Calcul du test de Wilcoxon chez 10 patients (données fictives).

Patient	Avant	Après	d	abs(d)	Rang de abs(d)	Rang signé
1	92,0	89,5	2,5	2,5	6	+6
2	85,5	85,5	0,0	0	éliminé	-
3	101,2	97,0	4,2	4,2	8	+8
4	78,4	79,6	-1,2	1,2	2	-2
5	88,0	85,8	2,2	2,2	4,5	+4,5
6	95,3	93,1	2,2	2,2	4,5	+4,5
7	83,1	83,9	-0,8	0,8	1	-1
8	110,4	104,2	6,2	6,2	9	+9
9	90,2	87,0	3,2	3,2	7	+7
10	87,6	86,1	1,5	1,5	3	+3

Le patient 2 a une différence nulle et est éliminé, donc n = 9. Les valeurs absolues ordonnées : 0,8 (rang 1), 1,2 (rang 2), 1,5 (rang 3), 2,2 et 2,2 (rangs 4 et 5, donc 4,5 chacune), 2,5 (rang 6), 3,2 (rang 7), 4,2 (rang 8), 6,2 (rang 9).

$$W^+ = 6 + 8 + 4{,}5 + 4{,}5 + 9 + 7 + 3 = 42\, W^- = 2 + 1 = 3$$

Vérification : $W^+ + W^- = 45 = \frac{9\times10}{2}$. C'est correct. La statistique de test est T = le minimum de W+ et W-, soit T = 3.

La valeur critique pour n = 9, en bilatéral, α = 0,05, est 5 (tableau 7.16). Comme pour le test de Mann-Whitney, **on rejette H0 si T est inférieur ou égal à la valeur critique**. Comme 3 ≤ 5, nous rejetons H0 ; p exact = 0,020. Pour un n plus grand, on utilise l'approximation normale :

$$\mu_T = \frac{n(n+1)}{4} = 22{,}5 \sigma_T = \sqrt{\left[\frac{n(n+1)\,(2n+1)}{24}\right]} = \sqrt{\frac{9\times10\times19}{24}} = 8{,}44 z = \frac{3-22{,}5}{8{,}44} = -2{,}31, \text{ p} \approx 0{,}021$$

Conclusion : le poids a diminué significativement après le programme (médiane des différences 2,2 kg ; 7 des 9 patients ayant une différence non nulle ont perdu du poids).

7.4.3 Le test des signes

Le test des signes est le test le plus simple pour données appariées. Il compte seulement combien de différences sont positives et combien sont négatives, en ignorant leur ampleur. Sous H0, chaque différence non nulle a une probabilité de 0,5 d'être positive, de sorte que le nombre de différences positives suit une distribution binomiale de paramètres n et p = 0,5.

Dans l'exemple ci-dessus, nous avons 7 différences positives sur 9. La probabilité bilatérale d'obtenir 7 ou plus (ou, symétriquement, 2 ou moins) pour une binomiale avec n = 9 et p = 0,5 est de 0,18. Le test des signes ne trouve pas de différence significative là où le test de Wilcoxon trouve p = 0,020. L'explication : les deux patients qui ont grossi n'ont pris que très peu de poids (1,2 et 0,8 kg), tandis que ceux qui ont maigri ont perdu beaucoup plus. Le test de Wilcoxon utilise cette information, le test des signes la jette. Le test des signes reste utile lorsque nous ne disposons que du sens du changement ("mieux", "pareil", "moins bien"), par exemple dans l'évaluation globale par le patient.

7.4.4 Application à mg2026 : le poids dans le bras A, un résultat à la limite

Dans le bras A (conseil seul, n = 53), la médiane de la différence de poids entre l'inclusion et 6 mois est de 1,1 kg. Au total, 32 patients ont perdu du poids, 19 en ont pris et 2 sont restés stables. Le test de Wilcoxon donne W- = 456 et W+ = 870 pour n = 51 différences non nulles, avec p = 0,052.

On voit bien ici pourquoi il n'est pas raisonnable de traiter le seuil de 0,05 comme un mur. Avec p = 0,052, nous ne rejetons pas H0, mais cela ne signifie pas que le programme du bras A n'a aucun effet : cela signifie que les données ne sont pas suffisamment convaincantes au seuil choisi. Nous formulerons la conclusion avec prudence : "à 6 mois, dans le bras avec conseil standard, la perte de poids (médiane 1,1 kg) n'a pas atteint le seuil de signification statistique (Wilcoxon, p = 0,052)". L'absence de preuve n'est pas la preuve de l'absence [79].

Le même exemple illustre l'efficacité relative abordée à la section 7.1.3. Les différences de poids dans le bras A ont en réalité une distribution compatible avec la loi normale (Shapiro-Wilk sur les différences p = 0,696), le test adapté était donc le test t apparié. Celui-ci donne une différence moyenne de 0,71 kg (ÉT 2,48), t = 2,07, ddl 52, p = 0,043. Le test des signes (32 signes plus sur 51) donne p = 0,092. Trois tests, une même question, trois valeurs p ordonnées exactement selon la quantité d'information utilisée : moyenne et variance (t), rangs (Wilcoxon), signes seuls (test des signes). La leçon n'est pas "choisissez le test qui donne un résultat significatif", mais "choisissez le test adapté avant de voir le résultat" ; ici, le test t apparié était justifié par la distribution des différences. Dans les deux autres bras, il n'y a aucune ambiguïté : le test de Wilcoxon donne p < 0,001 aussi bien dans le bras B (médiane 2,3 kg) que dans le bras C (médiane 3,55 kg).

Un exemple tiré des travaux de nos collègues montre à quoi ressemblent des données appariées rapportées sous forme de médianes. Gherbon et ses collaborateurs, avec la Dre Frandeș, ont suivi rétrospectivement 200 patients roumains atteints de diabète de type 2 traités par inhibiteurs du SGLT-2, en comparant les valeurs du début du traitement à celles obtenues après un an [59]. Les données sont appariées (mêmes patients, deux temps), et les résultats sont rapportés sous forme de médiane et d'écart interquartile : la glycémie est passée de 180,00 mg% (EIQ 154,50-207,00) à 130,00 mg% (EIQ 117,50-150,00), p < 0,001 ; l'HbA1c de 8,40% (EIQ 7,98-9,15%) à 7,30% (EIQ 6,90-7,95%), p < 0,001 ; le poids de 90,50 kg (82,00-106,50) à 89,00 kg (77,50-100,00), p = 0,018 ; la pression systolique de 140,00 mmHg (130,00-160,00) à 120,00 mmHg (115,50-130,00), p < 0,001. Le tour de taille apparaît en revanche sous forme de moyenne ± ÉT, de 107,05 (± 16,39) à 102,50 (± 15,11), p = 0,042. Le résumé ne nomme pas le test, mais le choix des médianes suggère un test apparié non paramétrique (Wilcoxon) pour les variables asymétriques et le test t apparié pour le tour de taille. Deux observations de lecture critique : le résumé qualifie ces médianes de "moyennes", une confusion terminologique qu'un étudiant attentif repère immédiatement, et pour des données appariées le chiffre le plus informatif aurait été la médiane des différences individuelles (avec l'EIQ ou un intervalle de confiance), à côté des médianes des deux temps. La différence entre les médianes (par exemple 90,50 - 89,00 = 1,50 kg) n'est pas obligatoirement égale à la médiane des différences.

7.5 Le test de Kruskal-Wallis

Lorsque nous avons trois groupes indépendants ou plus et que la variable est ordinale, ou numérique sans distribution normale, nous utilisons le test de Kruskal-Wallis [95], alternative non paramétrique de l'ANOVA à un facteur. On peut le considérer comme une généralisation du test de Mann-Whitney à k groupes.

7.5.1 Les hypothèses et la statistique H

H0 : les k populations ont la même distribution (si les formes sont similaires : Me1 = Me2 =... = Mek) H1 : au moins une population diffère des autres (au moins une médiane diffère)

Les N valeurs sont toutes ordonnées ensemble et reçoivent des rangs. Pour chaque groupe i, on calcule la somme des rangs Ri. Si H0 est vraie, le rang moyen de chaque groupe devrait être proche du rang moyen général, (N + 1)/2. La statistique de test mesure à quel point les rangs moyens des groupes s'écartent de cette valeur :

$$H = \left[\frac{12}{N(N+1)}\right] \times \sum\left(\frac{{R_i}^2}{n_i}\right) - 3(N+1)$$

Pour des groupes pas trop petits (en règle générale, au moins 5 sujets dans chaque groupe), H suit approximativement une distribution du khi-deux à k - 1 degrés de liberté. On rejette H0 si H dépasse la valeur critique : 5,99 pour 2 degrés de liberté, 7,81 pour 3, 9,49 pour 4 (α = 0,05). En présence d'ex aequo, H est divisé par le facteur de correction

$$C = 1 - \frac{\sum(t^3 - t)}{N^3 - N}$$

où t est le nombre de valeurs d'un groupe d'ex aequo. La correction augmente légèrement H.

7.5.2 Le calcul à la main

Quinze patients opérés ont reçu l'un des trois protocoles d'analgésie : protocole 1 (paracétamol), protocole 2 (paracétamol + anti-inflammatoire), protocole 3 (analgésie multimodale). Le score de douleur à 24 heures (0-10) était le suivant :

- protocole 1 : 6, 7, 5, 8, 7

- protocole 2 : 4, 5, 3, 5, 6
- protocole 3 : 2, 3, 4, 1, 3

Ordonnons les 15 valeurs : 1, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 6, 7, 7, 8. Les rangs : 1 → 1 ; 2 → 2 ; les trois valeurs 3 occupent les positions 3, 4, 5 → rang 4 ; les deux valeurs 4 occupent les positions 6, 7 → 6,5 ; les trois valeurs 5 occupent les positions 8, 9, 10 → 9 ; les deux valeurs 6 → 11,5 ; les deux valeurs 7 → 13,5 ; 8 → 15.

Tableau 7.5. Les rangs pour le test de Kruskal-Wallis (données fictives).

Protocole 1	Rang	Protocole 2	Rang	Protocole 3	Rang
6	11,5	4	6,5	2	2
7	13,5	5	9	3	4
5	9	3	4	4	6,5
8	15	5	9	1	1
7	13,5	6	11,5	3	4
R1	62,5	R2	40	R3	17,5

Vérification : $62{,}5 + 40 + 17{,}5 = 120 = \frac{15\times16}{2}$. C'est correct.

$$H = \left[\frac{12}{15 \times 16}\right] \times \left(\frac{62{,}5^2}{5} + \frac{40^2}{5} + \frac{17{,}5^2}{5}\right) - 3 \times 16 = 0{,}05 \times (781{,}25 + 320 + 61{,}25) - 48$$
$$= 0{,}05 \times 1162{,}5 - 48 = 58{,}125 - 48 = 10{,}125$$

Correction pour les ex aequo : nous avons deux groupes de trois valeurs égales (les valeurs 3 et 5) et trois groupes de deux (les valeurs 4, 6 et 7). $\sum(t^3 - t) = 2 \times 24 + 3 \times 6 = 66$, et $C = 1 - \frac{66}{3375-15} = 0{,}980$. H corrigé $= \frac{10{,}125}{0{,}980} = 10{,}33$.

Avec 2 degrés de liberté, la valeur critique est 5,99. Comme 10,33 > 5,99, nous rejetons H0 (p = 0,006). Les médianes sont de 7, 5 et 3. Conclusion : le score de douleur diffère entre les trois protocoles. Le test ne dit cependant pas quels protocoles diffèrent entre eux ; pour cela, nous avons besoin de comparaisons post hoc.

7.5.3 Les comparaisons post hoc : le test de Dunn

Après un test de Kruskal-Wallis significatif, la procédure la plus utilisée est le test de Dunn. Pour chaque paire de groupes, il compare les rangs moyens (calculés dans le classement commun des N sujets) au moyen d'une statistique z :

z = (rang moyen i - rang moyen j)/√{[N(N + 1)/12] x (1/ni + 1/nj)}

Les valeurs p obtenues sont corrigées en fonction du nombre de comparaisons. La correction la plus simple est celle de Bonferroni : chaque p est multiplié par le nombre de comparaisons (3 pour trois groupes) ou, de façon équivalente, le seuil de 0,05 est divisé par 3. La correction de Holm, appliquée par étapes, est moins conservatrice. Une erreur fréquente consiste à faire trois tests de Mann-Whitney séparés, chacun au seuil de 0,05, sans aucune correction : la probabilité d'obtenir au moins un résultat faussement positif passe alors de 5% à près de 14%.

7.5.4 Application à mg2026 : la baisse de la glycémie et la satisfaction dans les trois bras

Au chapitre 6, nous avons comparé la baisse de la glycémie à jeun à 6 mois entre les trois bras par une ANOVA : F(2 ; 157) = 22,93, p < 0,001. Le test de Kruskal-Wallis sur les mêmes données donne H = 36,63, ddl 2, p < 0,001. Le calcul suit exactement les étapes ci-dessus, mais sur 160 patients. Les sommes des rangs sont de 2696 (A), 4554,5 (B) et 5629,5 (C), soit des rangs moyens de 50,9, 85,9 et 104,3, contre un rang moyen général de (160 + 1)/2 = 80,5.

$H = \left[\frac{12}{160\times161}\right] \times \left(\frac{2696^2}{53} + \frac{4554{,}5^2}{53} + \frac{5629{,}5^2}{54}\right) - 3 \times 161 = 0{,}000466 \times (137\,139{,}9 + 391\,386{,}2 + 586\,875{,}4) - 483 = 36{,}60$, et avec la correction pour les ex aequo 36,63.

Les médianes de la baisse de la glycémie sont de 4 mg/dL (EIQ -3 à 9) dans le bras A, de 14 mg/dL (3 à 20) dans le bras B et de 16 mg/dL (9 à 23) dans le bras C. Le test de Dunn avec correction de Bonferroni montre des différences significatives entre A et B (p = 0,0003) et entre A et C (p < 0,001), mais pas entre B et C (p ajusté = 0,12 ; non ajusté 0,041). Il est intéressant de comparer avec le test de Tukey du chapitre 6, qui avait aussi trouvé la différence C-B significative (p = 0,024). Ce n'est pas une contradiction : les tests sur les rangs sont un peu plus conservateurs lorsque les données sont approximativement normales, et la correction de Bonferroni est plus stricte que celle de Tukey. Pour cette variable numérique, avec une distribution raisonnable dans chaque bras et Levene p = 0,163, l'analyse principale reste l'ANOVA avec Tukey ; le test de Kruskal-Wallis sert d'analyse de sensibilité qui confirme le sens du résultat.

La satisfaction, qui est ordinale, s'analyse directement par le test de Kruskal-Wallis : H = 32,16, ddl 2, p < 0,001. Les rangs moyens sont de 53,5 (A), 86,7 (B) et 100,9 (C). Comparaisons de Dunn avec Bonferroni : A contre B p = 0,0003, A contre C p < 0,001, B contre C p = 0,29. Ainsi, les deux bras avec application ont produit des patients plus satisfaits que le conseil seul, et l'ajout de la téléconsultation n'a pas modifié significativement la satisfaction par rapport à l'application seule. À titre de vérification, le test de Mann-Whitney direct entre B et C donne p = 0,079.

7.5.5 La taille d'effet pour le test de Kruskal-Wallis

L'analogue de l'êta carré de l'ANOVA est l'epsilon carré :

$$\varepsilon^2 = \frac{H}{N-1}$$

Pour la baisse de la glycémie, $\varepsilon^2 = \frac{36{,}63}{159} = 0{,}23$; pour la satisfaction, $\varepsilon^2 = \frac{32{,}16}{159} = 0{,}20$. Ce sont deux effets importants, comparables au $\eta^2 = 0{,}226$ obtenu par l'ANOVA pour la glycémie. Rapporter cette grandeur est une bonne pratique, de plus en plus fréquente dans les articles, comme dans l'étude sur le VIH présentée ci-dessous.

Étude de cas : le test de Kruskal-Wallis dans trois articles de nos collègues

Gherbon et ses collaborateurs, avec la Dre Frandeș, ont analysé de façon transversale 453 adultes atteints de diabète de type 2, parmi lesquels une détresse sévère liée au diabète a été identifiée chez 19 participants (4,2%) [69]. Pour comparer les scores entre trois catégories ordonnées, ils ont utilisé le test de Kruskal-Wallis, avec p = 0,003, puis une régression logistique multivariée dans laquelle l'âge avait un aOR = 1,163 par année (IC à 95% 1,055-1,282 ; p = 0,002). La logique est claire : les scores de détresse sont ordinaux et asymétriques (peu de patients avec des valeurs élevées), un test sur les rangs est donc adapté. Hasan et ses collaborateurs, avec la Dre Tudor, ont étudié 90 patients un mois après une chirurgie thoracique et ont comparé l'anxiété et l'intensité de la douleur entre trois groupes définis par le style de coping, au moyen du test de Kruskal-Wallis : p = 0,028, p = 0,022, p = 0,042 et p = 0,007 pour quatre variables [96]. Les scores de questionnaires validés sont des variables ordinales, et le test est bien choisi ; la question que se pose un lecteur attentif est de savoir combien de comparaisons ont été faites au total et si elles ont été prises en compte, car trois des quatre valeurs p se situent entre 0,02 et 0,05. Marin et ses collaborateurs, avec le Dr R. Moleriu et le Dr Muntean, ont étudié 462 adultes vivant avec le VIH sous traitement antirétroviral [68]. Des réponses immuno-virologiques discordantes sont apparues chez 30,7% des patients (14,5% et 16,2% pour les deux types de discordance), et les comparaisons entre les trois groupes de réponse ont utilisé le test de Kruskal-Wallis avec la taille d'effet : H = 153,62, p <

> 0,001, ε² = 0,33 et H = 115,10, p < 0,001, ε² = 0,25. C'est le modèle de compte rendu complet : la statistique, la valeur p et la taille d'effet. Un ε² de 0,33 signifie qu'environ un tiers de la variabilité des rangs est expliquée par l'appartenance au groupe.

7.6 Le test de Friedman

Le test de Friedman [97] est l'alternative non paramétrique de l'ANOVA à mesures répétées. Nous l'utilisons lorsque les mêmes sujets sont évalués à trois temps ou plus (ou dans trois conditions ou plus, ou avec trois méthodes ou plus) et que la variable est ordinale, ou numérique sans distribution normale.

7.6.1 La logique du test

La différence avec le test de Kruskal-Wallis est que les rangs sont attribués **à l'intérieur de chaque sujet**, et non dans l'échantillon commun. Pour chaque patient, les k mesures reçoivent les rangs 1, 2,..., k. Ainsi, les différences entre patients (l'un a des glycémies autour de 90, l'autre autour de 140) disparaissent, et il ne reste que l'ordre des mesures chez un même patient. S'il n'existe aucun effet du temps, chaque temps devrait recevoir, en moyenne, le rang (k + 1)/2.

H0 : la distribution des valeurs est la même aux k temps H1 : au moins un temps diffère des autres

Avec n sujets et k temps, et Rj la somme des rangs du temps j :

$$\chi^2 F = \left[\frac{12}{n \times k \times (k+1)}\right] \times \sum {R_j}^2 - 3n(k+1)$$

La statistique suit approximativement une distribution du khi-deux à k - 1 degrés de liberté.

7.6.2 Le calcul à la main

Six patients lombalgiques ont noté leur douleur (0-10) au début d'un programme de kinésithérapie, à 2 semaines et à 6 semaines.

Tableau 7.6. Les rangs à l'intérieur de chaque patient pour le test de Friedman (données fictives).

Patient	Début	2 sem.	6 sem.	Rangs (début ; 2 sem. ; 6 sem.)
1	7	5	3	3; 2; 1
2	8	6	4	3; 2; 1
3	6	6	5	2,5; 2,5; 1
4	9	7	4	3; 2; 1
5	5	4	4	3; 1,5; 1,5
6	7	4	2	3; 2; 1
Somme des rangs Rj				17,5; 12; 6,5

Vérification : chaque ligne a une somme des rangs de 1 + 2 + 3 = 6, le total est donc de 6 x 6 = 36 = 17,5 + 12 + 6,5. C'est correct.

$$\chi^2 F = \left[\frac{12}{6 \times 3 \times 4}\right] \times (17{,}5^2 + 12^2 + 6{,}5^2) - 3 \times 6 \times 4$$
$$= \left(\frac{12}{72}\right) \times (306{,}25 + 144 + 42{,}25) - 72 = 0{,}1667 \times 492{,}5 - 72 = 82{,}08 - 72$$
$$= 10{,}08$$

Avec 2 degrés de liberté, la valeur critique est 5,99. Nous rejetons H0. Les logiciels appliquent ici aussi une correction pour les deux égalités (patients 3 et 5) et rapportent 11,00, avec p = 0,004. La douleur a diminué significativement au cours du programme.

7.6.3 Application à mg2026 : la glycémie à 0, 3 et 6 mois

La glycémie à jeun a été mesurée à l'inclusion, à 3 mois et à 6 mois chez l'ensemble des 160 patients. Les médianes sont de 104,5, 99,0 et 92,5 mg/dL (moyennes 105,4, 99,4 et 95,2). La figure 7.3 présente la médiane et l'écart interquartile par bras.

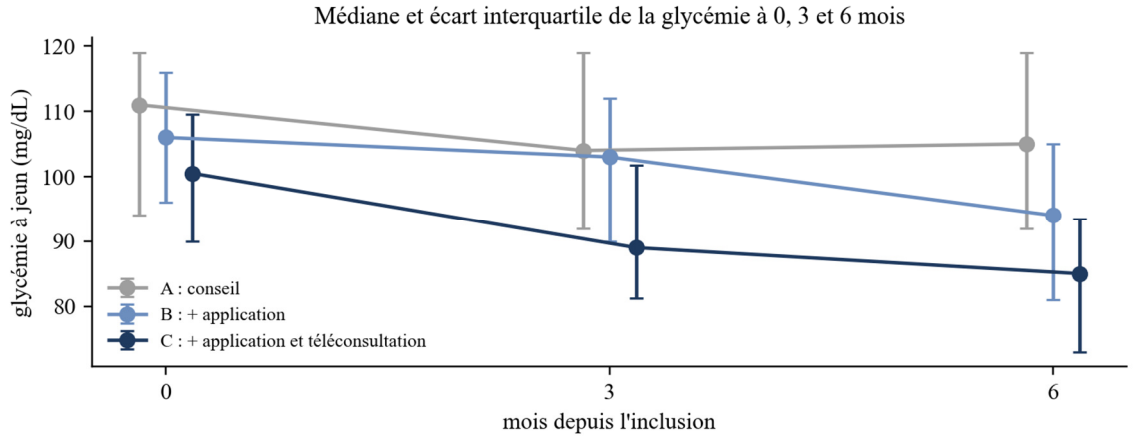


Figure 7.3. Médiane et écart interquartile de la glycémie à jeun à 0, 3 et 6 mois, par bras. Dans le bras C, la baisse est la plus marquée et se poursuit jusqu'à 6 mois ; dans le bras A, la médiane baisse à 3 mois puis stagne. Données synthétiques (jeu de données mg2026).

Test de Friedman sur l'ensemble des patients : les sommes des rangs sont de 406 (0 mois), 295 (3 mois) et 259 (6 mois), soit des rangs moyens de 2,54, 1,84 et 1,62. Sans correction, $\chi^2F = \left[\frac{12}{160\times3\times4}\right] \times (406^2 + 295^2 + 259^2) - 3 \times 160 \times 4 = 73{,}39$; avec la correction pour les ex aequo, 74,67, ddl 2, p < 0,001. À titre de comparaison, l'ANOVA à mesures répétées donne F = 64,97, p < 0,001. Les deux tests disent la même chose.

La taille d'effet pour le test de Friedman est le coefficient de concordance W de Kendall :

$$W = \frac{\chi^2 F}{n(k-1)} = \frac{74{,}67}{160 \times 2} = 0{,}23$$

W va de 0 (aucune tendance commune) à 1 (tous les patients ont exactement le même ordre des mesures). Une valeur de 0,23 indique une tendance commune modérée : de nombreux patients présentent une baisse, mais pas tous dans le même ordre.

Les comparaisons post hoc se font avec le test de Wilcoxon pour chaque paire de temps, avec correction de Bonferroni (3 comparaisons, chaque p est donc multiplié par 3 ou le seuil devient 0,05/3 = 0,0167) :

Tableau 7.7. Les comparaisons post hoc après le test de Friedman (glycémie, n = 160).

Comparaison	Médiane des différences (mg/dL)	p Wilcoxon	p ajusté Bonferroni
0 mois contre 3 mois	7,0	< 0,001	< 0,001
3 mois contre 6 mois	4,0	0,0002	0,0007
0 mois contre 6 mois	9,5	< 0,001	< 0,001

Les trois comparaisons restent toutes significatives après correction. L'analyse par bras est plus instructive : dans le bras C, le test de Friedman donne χ^2F = 64,63, p < 0,001 (médianes

100,5, 89,0 et 85,0 mg/dL), dans le bras B $\chi^2F = 21,07$, $p < 0,001$, et dans le bras A seulement $\chi^2F = 6,04$, $p = 0,049$ (médianes 111, 104 et 105 mg/dL). Dans le bras A, le résultat global franchit de justesse le seuil de 0,05, mais aucune des comparaisons de Wilcoxon ne reste significative après correction de Bonferroni (le plus petit p ajusté est de 0,075, pour 0 contre 6 mois). Cette situation est fréquente en pratique : le test global "voit" une tendance faible, mais la puissance ne suffit pas à localiser la différence après correction.

Le test de Friedman apparaît aussi dans les études de comparaison de méthodes de mesure. Al Barri et ses collaborateurs, avec la Dre Tudor, ont comparé trois appareils de kératométrie et de biométrie oculaire chez des patients candidats à la chirurgie de la cataracte : 170 yeux de 102 patients, âge moyen 69,12 ± 10,26 ans [81]. Chaque œil a été mesuré avec les trois appareils, nous avons donc trois mesures liées, exactement la situation du test de Friedman, suivi de comparaisons de Wilcoxon deux à deux. Pour la kératométrie K1, les moyennes étaient de 43,45 ± 1,64 D (Argos) et de 43,41 ± 1,70 D (Aladdin), avec $p < 0,001$ globalement et $p = 0,019$ pour Argos contre Aladdin. La profondeur de la chambre antérieure était de 3,19 ± 0,42 contre 3,13 ± 0,41mm, $p < 0,001$, et l'épaisseur cornéenne était pratiquement identique entre Aladdin et Sirius (540,27 ± 33,44 contre 540,47 ± 33,78 µm, $p = 0,169$). Les composantes vectorielles de l'astigmatisme J0 ($p = 0,277$) et J45 ($p = 0,084$) n'ont pas différé. L'article illustre très bien la différence entre signification statistique et signification clinique : une différence de 0,04 D pour K1 ou de 0,06 mm pour la profondeur de la chambre antérieure peut être "significative", car les mesures appariées éliminent la variabilité entre les yeux, mais elle reste bien en dessous du seuil qui modifierait le calcul de l'implant cristallinien. La seconde observation concerne l'indépendance : les 170 yeux proviennent de 102 patients, certains patients fournissent donc deux yeux, et les yeux d'un même patient ne sont pas totalement indépendants. Nous reviendrons sur ce thème au chapitre 8, à propos de l'unité d'analyse.

7.7 Le test du khi-deux pour les tableaux de contingence

Jusqu'ici, nous avons travaillé avec des variables numériques ou ordinales. Lorsque les deux variables sont qualitatives (nominales, ou ordinales avec peu de catégories) et que nous voulons savoir si elles sont associées, les données se résument dans un tableau de contingence, et l'outil classique est le test du khi-deux (χ^2) proposé par Karl Pearson en 1900 [98]. C'est probablement le test le plus fréquemment rencontré dans les articles médicaux, car de nombreux critères de jugement cliniques sont de type "oui/non" : guéri ou non, hospitalisé ou non, avec ou sans complication.

7.7.1 Le tableau de contingence et les hypothèses

Un tableau de contingence r x c comporte r lignes (les catégories d'une variable) et c colonnes (les catégories de l'autre). Dans chaque cellule, on inscrit le nombre de sujets qui présentent à la fois la catégorie de la ligne et celle de la colonne. Les totaux des lignes et des colonnes sont appelés totaux marginaux.

L'exemple de travail provient de l'analyse transversale des patients de mg2026 à l'inclusion : existe-t-il une association entre l'obésité (IMC ≥ 30 kg/m²) et le diabète confirmé par l'hyperglycémie provoquée par voie orale (HGPO) ?

Tableau 7.8. Obésité et diabète confirmé par HGPO à l'inclusion, jeu de données mg2026 (effectifs observés).

	Diabète HGPO oui	**Diabète HGPO non**	**Total**

	Diabète HGPO oui	**Diabète HGPO non**	**Total**
Obèses (IMC ≥ 30)	18	36	54
Non obèses (IMC < 30)	11	95	106
Total	29	131	160

La prévalence du diabète est de 18/54 = 33,3% chez les obèses et de 11/106 = 10,4% chez les non-obèses. Les hypothèses :

H0 : l'obésité et le diabète sont indépendants (la prévalence du diabète est la même chez les obèses et chez les non-obèses : p1 = p2) H1 : il existe une association entre l'obésité et le diabète (p1 ≠ p2)

7.7.2 Les effectifs attendus

L'idée du test est de comparer ce que nous avons observé avec ce que nous nous attendrions à observer si H0 était vraie. Si l'obésité n'a aucun lien avec le diabète, la proportion globale de diabétiques, 29/160 = 18,1%, devrait se retrouver à l'identique chez les obèses et chez les non-obèses. Parmi les 54 obèses, nous nous attendrions alors à $\frac{54\times29}{160} = 9{,}79$ diabétiques. La règle générale pour toute cellule :

E = (total de la ligne x total de la colonne)/total général

Tableau 7.9. Les effectifs attendus sous l'hypothèse d'indépendance.

	Diabète HGPO oui	**Diabète HGPO non**	**Total**
Obèses	$\frac{54\times29}{160} = 9{,}79$	$\frac{54\times131}{160} = 44{,}21$	54
Non obèses	$\frac{106\times29}{160} = 19{,}21$	$\frac{106\times131}{160} = 86{,}79$	106
Total	29	131	160

Vérification : les effectifs attendus ont les mêmes totaux marginaux que les effectifs observés (9,79 + 44,21 = 54 ; 9,79 + 19,21 = 29). Si les totaux ne concordent pas, vous avez fait une erreur de calcul. Les effectifs attendus peuvent être décimaux ; on ne les arrondit pas à des nombres entiers.

7.7.3 La statistique, les degrés de liberté, la décision

La statistique du khi-deux additionne, pour toutes les cellules, l'écart quadratique relatif entre l'observé (O) et l'attendu (E) :

$$\chi^2 = \frac{\sum(O-E)^2}{E}$$

Tableau 7.10. Contributions des cellules au χ^2 (obésité x diabète).

Cellule	**O**	**E**	**O - E**	$\frac{(O-E)^2}{E}$
obèse, diabète	18	9,79	8,21	6,891
obèse, sans diabète	36	44,21	-8,21	1,525
non obèse, diabète	11	19,21	-8,21	3,510
non obèse, sans diabète	95	86,79	8,21	0,777
Total	160	160	0	$\chi^2 = 12{,}70$

Remarquez que, dans un tableau 2 x 2, tous les écarts O - E ont la même valeur absolue (ici 8,21). Pour les tableaux 2 x 2, il existe aussi une formule abrégée, utile à l'examen :

$$\chi^2 = \frac{N(\mathrm{ad}-\mathrm{bc})^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)}$$

= 160 x (18 x 95 - 36 x 11)²/(54 x 106 x 29 x 131) = 160 x 1314²/21 745 116 = 12,70

Le nombre de degrés de liberté pour un tableau r x c est :

$$\text{ddl} = (r - 1)\,(c - 1)$$

Pour un tableau 2 x 2, ddl = 1. Le raisonnement : une fois les totaux marginaux fixés, une seule cellule peut être choisie librement, les autres s'en déduisent par soustraction. La valeur critique pour ddl = 1 et α = 0,05 est 3,84 (pour ddl = 2, elle est de 5,99 ; pour ddl = 3, de 7,81 ; pour ddl = 4, de 9,49). Comme 12,70 > 3,84, nous rejetons H0 ; le logiciel donne p = 0,0004 (nous rapportons p < 0,001).

Conclusion statistique : il existe une association significative entre l'obésité et le diabète confirmé par HGPO. Conclusion clinique : dans cette population à risque cardiométabolique, le diabète était environ trois fois plus fréquent chez les obèses (33,3% contre 10,4% ; rapport de prévalences 3,21, IC à 95% 1,64-6,31 ; OR 4,32, IC à 95% 1,86-10,03). S'agissant d'une étude transversale, nous ne pouvons pas dire que l'obésité a "produit" le diabète ; nous disons seulement que les deux surviennent ensemble plus souvent que ne le voudrait le hasard. Les mesures de risque et les schémas d'étude sont détaillés au chapitre 10.

Le khi-deux est toujours un test "bilatéral" par construction : il détecte des écarts dans n'importe quelle direction. Le sens de l'association se lit dans le tableau (pourcentages en ligne), et non dans la valeur du χ^2.

7.7.4 Un tableau 3 x 2 sur mg2026 : le bras et une perte de poids d'au moins 5%

L'objectif secondaire de l'étude mg2026 portait sur la proportion de patients ayant perdu au moins 5% de leur poids initial en 6 mois. La figure 7.4 présente les proportions par bras.

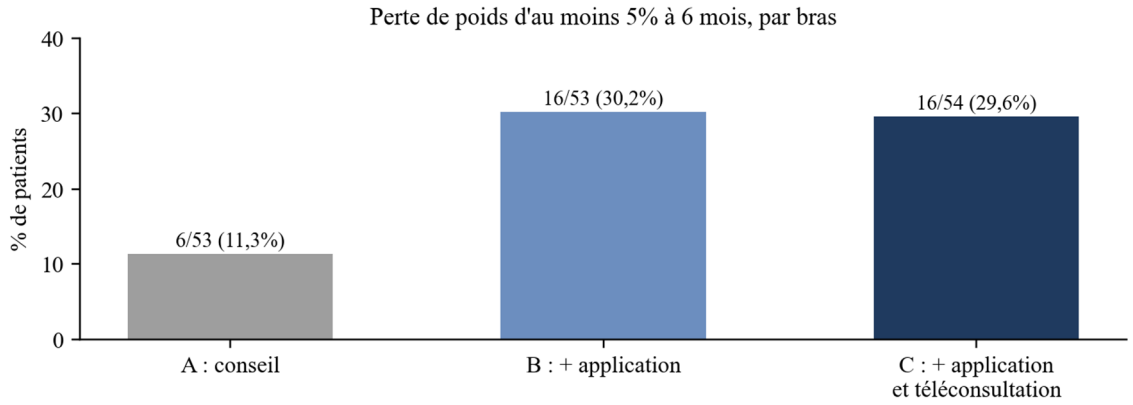


Figure 7.4. Proportion de patients ayant perdu au moins 5% de leur poids à 6 mois, par bras : 6/53 (11,3%) dans le bras A, 16/53 (30,2%) dans le bras B et 16/54 (29,6%) dans le bras C. Test du khi-deux sur le tableau 3 x 2 : χ^2 = 6,77, ddl 2, p = 0,034. Données synthétiques (jeu de données mg2026).

Tableau 7.11. Bras de l'étude et perte de poids ≥ 5% : effectifs observés, attendus (entre parenthèses) et contributions au χ^2.

Bras	Perte ≥ 5%	Sans perte ≥ 5%	Total	% avec perte
A	6 (12,59 ; contribution 3,447)	47 (40,41; 1,074)	53	11,3
B	16 (12,59; 0,925)	37 (40,41; 0,288)	53	30,2
C	16 (12,82; 0,786)	38 (41,18; 0,245)	54	29,6
Total	38	122	160	23,8

Les effectifs attendus se calculent de la même façon : pour le bras A et la colonne "perte ≥ 5%", $E = \frac{53 \times 38}{160} = 12{,}59$. La somme des contributions est χ^2 = 3,447 + 1,074 + 0,925 + 0,288 + 0,786 + 0,245 = 6,77. Degrés de liberté : (3 - 1)(2 - 1) = 2. La valeur critique de 5,99 est dépassée,

p = 0,034. Le plus petit effectif attendu est de 12,6, les conditions d'application sont donc remplies.

Le tableau des contributions indique aussi **où** se trouve l'association : près de la moitié du χ^2 (3,447 sur 6,77) provient d'une seule cellule, le bras A avec perte ≥ 5%, où nous avons observé 6 patients au lieu des 12,6 attendus. Les bras B et C sont très proches l'un de l'autre.

Un χ^2 significatif dans un tableau plus grand que 2 x 2 est, comme pour l'ANOVA, un résultat global. Si nous voulons comparer les bras deux à deux, nous construisons des tableaux 2 x 2 et corrigeons pour les comparaisons multiples. Sans correction : A contre B $\chi^2 = 5,74$, p = 0,017 ; A contre C $\chi^2 = 5,49$, p = 0,019 ; B contre C $\chi^2 = 0,00$, p = 0,95. Avec Bonferroni pour trois comparaisons, le seuil devient 0,0167, et A contre B ainsi que A contre C se retrouvent juste à la limite. Pour la question principale de l'étude, définie dans le protocole comme la comparaison C contre A, nous n'avons pas besoin de correction, et le résultat s'exprime le plus utilement par le risque relatif : 29,6% contre 11,3%, RR 2,62 (IC à 95% 1,11-6,17), différence absolue de 18,3 points de pourcentage (IC à 95% 3,4-33,2), NST d'environ 6.

7.7.5 Les conditions d'application

Le test du khi-deux est un test approché : la distribution de la statistique ne se rapproche de la distribution théorique du χ^2 que lorsque les effectifs attendus sont suffisamment grands. Les règles utilisées en pratique [42,53] :

- les observations sont indépendantes : chaque sujet apparaît dans une seule cellule, une seule fois ;
- on travaille sur des effectifs absolus (nombres de sujets), jamais sur des pourcentages ou des moyennes ;
- pour les tableaux 2 x 2, tous les effectifs attendus doivent être d'au moins 5 ;
- pour les tableaux plus grands, au plus 20% des cellules peuvent avoir un effectif attendu inférieur à 5 (règle des 80%) et aucune cellule ne peut avoir un effectif attendu inférieur à 1.

Attention : la condition porte sur les effectifs **attendus**, et non sur les effectifs observés. Une cellule avec 0 observation peut être parfaitement acceptable si l'effectif attendu y est de 6. Lorsque la condition n'est pas remplie, nous avons trois solutions : le test exact de Fisher (pour les tableaux 2 x 2 et, dans de nombreux logiciels, pour des tableaux plus grands), le regroupement de catégories ayant un sens clinique (par exemple "fost fumator" (ancien fumeur) avec "fumator" (fumeur), si la question le permet) ou l'augmentation de la taille de l'échantillon.

L'inégalité des effectifs des groupes n'est pas en soi un problème pour le test du khi-deux ; seuls comptent les effectifs attendus. Un exemple tiré des travaux de nos collègues : Marin et ses collaborateurs, avec la Dre L. Moleriu, ont comparé rétrospectivement la réponse virologique chez 462 patients infectés par le VIH-1 de sous-type F traités par darunavir potentialisé soit par ritonavir (384 patients), soit par cobicistat (78 patients) [99]. Le résumé rapporte un pourcentage de 84,6% dans le groupe cobicistat, contre 76,8%. Les groupes sont très inégaux (384 et 78), mais avec des dizaines de patients dans chaque cellule, les effectifs attendus sont grands, et la comparaison de deux proportions par le khi-deux (ou par le test de Fisher, pour les cellules avec peu d'événements) est parfaitement valide. Le véritable problème d'un tel schéma est ailleurs : dans une étude observationnelle, le médecin a choisi le traitement en fonction des caractéristiques du patient, de sorte que la différence peut refléter le type de patients traités par chaque schéma (confusion par indication), et pas seulement l'effet du médicament.

7.7.6 La correction de Yates

Pour les tableaux 2 x 2, Frank Yates a proposé une correction de continuité : on soustrait 0,5 de chaque écart absolu avant de l'élever au carré.

$$\chi^2\text{Yates} = \frac{\sum(|O-E|-0{,}5)^2}{E}$$

La correction compense le fait que les effectifs observés sont des nombres entiers, alors que la distribution du χ^2 est continue. Sur l'exemple obésité x diabète, χ^2Yates = 11,20 (p = 0,0008), contre 12,70 sans correction. La correction rend le test plus conservateur. Beaucoup de statisticiens estiment aujourd'hui qu'elle corrige trop et préfèrent, pour les petits tableaux, recourir directement au test de Fisher ; JASP la propose en option ("continuity correction"), et Epi Info l'affiche automatiquement à côté de la version non corrigée. Règle pratique pour l'examen : pour un tableau 2 x 2 avec des effectifs attendus compris entre 5 et 10, rapportez la version avec correction de Yates ou le test de Fisher ; lorsque tous les effectifs attendus sont grands, la différence entre les versions ne change pas la conclusion.

7.7.7 Le test exact de Fisher

Le test exact de Fisher n'utilise aucune approximation. En considérant les totaux marginaux comme fixés, il calcule la probabilité exacte du tableau observé et de tous les tableaux "au moins aussi extrêmes", à l'aide de la distribution hypergéométrique. Pour un tableau de cellules a, b, c, d et N sujets :

$$P = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{N!\ a!\ b!\ c!\ d!}$$

Exemple : dans un audit des infections postopératoires, 8 patients ont reçu un nouveau protocole d'antibioprophylaxie (1 infection) et 10 patients l'ancien protocole (6 infections).

Tableau 7.12. Infection postopératoire selon le protocole de prophylaxie (données fictives).

	Infection	**Sans infection**	**Total**
Nouveau protocole	1	7	8
Ancien protocole	6	4	10
Total	7	11	18

Les effectifs attendus dans la colonne "infection" sont de $\frac{8\times7}{18} = 3{,}11$ et $\frac{10\times7}{18} = 3{,}89$, tous deux inférieurs à 5. Le test du khi-deux n'est pas valide ici, même si le logiciel le calcule sans protester : χ^2 = 4,22, p = 0,040. Le test de Fisher calcule la probabilité de chaque tableau possible ayant les mêmes totaux, selon le nombre a d'infections avec le nouveau protocole (a peut aller de 0 à 7) : P(a = 0) = 0,0038 ; P(a = 1) = 0,0528 (notre tableau) ; P(a = 2) = 0,2217 ; P(a = 3) = 0,3695 ; P(a = 4) = 0,2640 ; P(a = 5) = 0,0792 ; P(a = 6) = 0,0088 ; P(a = 7) = 0,0003. La valeur p bilatérale additionne les probabilités des tableaux qui sont au plus aussi probables que le tableau observé : 0,0038 + 0,0528 + 0,0088 + 0,0003 = 0,066.

La conclusion change : avec le test adapté (Fisher), p = 0,066, nous ne rejetons pas H0. Le khi-deux utilisé à tort aurait donné un résultat "significatif" (0,040), et la correction de Yates (p = 0,117) va dans le même sens que le test de Fisher. Avec seulement 18 patients, nous ne pouvons pas affirmer que le nouveau protocole réduit les infections, même si la différence observée (12,5% contre 60%) mérite une étude plus large. Le calcul manuel du test de Fisher est laborieux ; à l'examen, on vous demande généralement de savoir reconnaître quand il faut l'utiliser et d'interpréter la valeur p donnée par le logiciel.

Étude de cas : petits tableaux aux urgences, chez des patients présentant un accident vasculaire cérébral et un COVID long

Popa et ses collaborateurs, avec la Dre Tudor, ont analysé rétrospectivement 132 patients admis au service d'accueil des urgences pour une alerte AVC (accident vasculaire cérébral), dont 26 présentaient un COVID long et 106 non [100]. Le groupe de 26 est petit et certains symptômes sont rares, de sorte que plusieurs tableaux 2 x 2 ont de faibles effectifs attendus.

> Les auteurs ont utilisé le test exact de Fisher (le résumé emploie la formulation "Fisher's Exact Test chi-squared", p = 0,027, qui mélange les noms de deux tests différents). La céphalée est survenue chez 19,2% contre 5,7% des patients (OR = 3,97, p = 0,040), et l'asymétrie faciale chez 42,3% contre 19,8% (OR = 2,97, p = 0,022). Faisons un calcul approximatif à partir des pourcentages : 19,2% de 26 correspond à environ 5 patients, et 5,7% de 106 à environ 6 patients, soit au total quelque 11 patients avec céphalée. L'effectif attendu de patients avec céphalée dans le groupe COVID long serait alors de 26 x 11/132, soit environ 2,2, bien en dessous de 5 : le test de Fisher était exactement le bon choix. La thrombectomie mécanique a été plus fréquente dans le groupe COVID long (30,8% contre 16,0%), et une autre comparaison a donné 26,9% contre 14,2%, p = 0,143, sans signification statistique. Avec seulement 26 patients dans un groupe, une différence de près de 13 à 15 points de pourcentage peut rester non significative par manque de puissance ; "non significatif" ne veut pas dire ici "sans différence". Ce qui aurait complété le compte rendu : les intervalles de confiance des OR, qui montreraient à quel point les estimations obtenues sur de petits groupes sont imprécises.

7.8 Variantes du test du khi-deux

7.8.1 Le test de McNemar pour proportions appariées

Le test du khi-deux classique suppose des groupes indépendants. Lorsque la même variable dichotomique est mesurée deux fois chez les mêmes sujets (avant et après, ou avec deux tests diagnostiques appliqués au même patient), les données sont appariées et l'on utilise le test de McNemar. C'est l'équivalent, pour les variables binaires, du test t apparié.

Le tableau se construit différemment : les lignes correspondent à l'état au premier temps, les colonnes à l'état au second temps, et chaque patient n'apparaît qu'une seule fois. Les patients dont l'état n'a pas changé (cellules de la diagonale) n'apportent pas d'information sur le changement. Le test n'utilise que les cellules discordantes, b et c :

χ^2McNemar = (b - c)²/(b + c), à 1 degré de liberté ; avec correction de continuité : (|b - c| - 1)²/(b + c)

Application à mg2026 : combien de patients avaient une glycémie à jeun ≥ 100 mg/dL (hyperglycémie modérée à jeun ou plus) à l'inclusion, et combien à 6 mois ?

Tableau 7.13. Glycémie à jeun ≥ 100 mg/dL à l'inclusion et à 6 mois, mêmes 160 patients (jeu de données mg2026).

	À 6 mois ≥ 100	**À 6 mois < 100**	**Total**
À l'inclusion ≥ 100	61	36 (b)	97
À l'inclusion < 100	1 (c)	62	63
Total	62	98	160

La proportion de patients avec une glycémie ≥ 100 est passée de 97/160 = 60,6% à 62/160 = 38,8%. Cellules discordantes : 36 patients sont passés sous 100, un seul patient est passé au-dessus de 100. $\chi^2 = \frac{(36-1)^2}{36+1} = \frac{1225}{37} = 33{,}11$; avec correction, (35 - 1)²/37 = 31,24 ; p < 0,001. Il aurait été erroné d'appliquer ici le test du khi-deux habituel à un tableau "temps x état" comptant 320 observations : le même patient y apparaîtrait deux fois, et l'indépendance serait violée. Dans les travaux de nos collègues, Iovanescu et ses collaborateurs, avec la Dre Frandeș, ont utilisé le test de McNemar dans une étude de dépistage de la néoplasie colorectale dans laquelle plusieurs examens (test immunochimique fécal, sigmoïdoscopie, coloscopie) ont été évalués chez les mêmes personnes (McNemar, P < 0,001) ; 252 personnes ont ensuite eu une coloscopie, et une néoplasie avancée a été trouvée chez 75 d'entre elles (29,76%) [101]. Lorsque deux tests

diagnostiques sont appliqués aux mêmes personnes, la comparaison de leurs proportions de résultats positifs exige un test pour données appariées.

Un exemple moins spectaculaire : la pression systolique ≥ 140 mmHg à l'inclusion et à 6 mois. Au total, 16 patients sont passés sous 140 et 8 au-dessus de 140. χ^2McNemar = (16 - 8)²/24 = 2,67, p = 0,10 (avec correction 2,04, p = 0,15). Bien que la proportion de patients avec une PAS ≥ 140 ait diminué de 41/160 à 33/160, le changement n'est pas statistiquement significatif, car il repose sur seulement 24 patients discordants.

7.8.2 Le test du khi-deux de tendance

Lorsque l'une des variables a des catégories ordonnées (classes d'âge, stades, doses, ou bras d'intensité croissante de l'intervention), le test du khi-deux habituel ignore l'ordre : il donnerait le même résultat si nous mélangions les lignes. Le test de tendance (Cochran-Armitage) pose une question plus précise : la proportion de "oui" augmente-t-elle (ou diminue-t-elle) avec la catégorie ? On attribue à chaque catégorie un score (par exemple 0, 1, 2), et le test n'a qu'un seul degré de liberté, ce qui le rend plus puissant que le test global lorsqu'il existe une tendance monotone.

Dans mg2026, les bras A, B et C sont d'intensité croissante (conseil ; + application ; + application et téléconsultation). Le test de tendance avec les scores 0, 1, 2 donne χ^2trend = 4,92, ddl 1, p = 0,027. Les proportions sont de 11,3%, 30,2% et 29,6% : la tendance existe, mais elle est due presque entièrement au saut de A à B ; de B à C, la proportion n'augmente plus. Le test de tendance se trouve dans Epi Info (StatCalc, "Chi square for trend") et dans MedCalc.

7.8.3 Le test du khi-deux d'ajustement

Le test d'ajustement (angl. goodness of fit) compare la distribution d'une seule variable qualitative à une distribution théorique fixée à l'avance. Les effectifs attendus ne proviennent plus des totaux marginaux, mais de l'hypothèse : E = N x proportion théorique. Les degrés de liberté sont k - 1, où k est le nombre de catégories (si aucun paramètre n'a été estimé à partir des données).

Premier exemple : la randomisation dans mg2026 visait une répartition égale entre les trois bras. Observé 53, 53, 54 ; attendu 53,33 dans chaque bras. χ^2 = 0,0125, ddl 2, p = 0,99 : la répartition respecte très bien le rapport 1:1:1. Deuxième exemple : l'échantillon compte 86 femmes et 74 hommes. Par rapport à un rapport théorique de 50:50 (80 et 80), $\chi^2 = \frac{(86-80)^2}{80} + \frac{(74-80)^2}{80} = 0{,}45 + 0{,}45 = 0{,}90$, ddl 1, p = 0,34. Troisième exemple : supposons que, dans un registre régional, la distribution du niveau d'études chez les adultes de 40 à 70 ans soit de 15% de niveau collège, 35% de niveau lycée ou enseignement professionnel, 35% de niveau universitaire et 15% de niveau postuniversitaire (proportions hypothétiques, utilisées uniquement pour l'exercice). Observé dans mg2026 : 20, 57, 64, 19 ; attendu 24, 56, 56, 24. χ^2 = 0,67 + 0,02 + 1,14 + 1,04 = 2,87, ddl 3, p = 0,41. L'échantillon ne diffère pas significativement de la distribution de référence, ce qui plaide, sans la prouver, en faveur de sa représentativité pour cette caractéristique.

Le test de tendance s'impose d'emblée lorsque les groupes ont un ordre naturel. Piroș et ses collaborateurs, avec la Dre Lupușoru, la Dre L. Moleriu et le Dr Muntean, ont étudié de façon transversale 693 étudiants de 18 à 24 ans [6]. La prévalence du prédiabète était de 21,1% (diabète 1,4%). Les étudiants ont été regroupés par analyse de clusters en trois profils alimentaires : orienté vers la santé (prédiabète 15,4%), mixte (20,0%) et orienté vers la restauration rapide (27,3%), et le test du khi-deux sur le tableau 3 x 2 a donné p = 0,003. Les catégories sont ordonnées selon la qualité de l'alimentation, et les auteurs sont allés plus loin, exactement dans la direction évoquée ci-dessus : un test de tendance (p-trend < 0,001) et une régression logistique dans laquelle chaque catégorie de profil augmentait la cote de prédiabète avec un OR = 1,78 par catégorie (IC à 95% 1,38-2,30 ; p < 0,001). Lecture correcte : il existe une association graduelle entre le profil

alimentaire et le prédiabète, compatible avec une relation de type dose-réponse. S'agissant d'une étude transversale, l'association ne démontre pas la causalité, et les clusters sont construits à partir des mêmes données ; le résultat doit donc être confirmé dans une autre population.

7.9 Au-delà de la valeur p : phi, V de Cramér et autres mesures d'association

La valeur du χ^2 et la valeur p évoluent avec la taille de l'échantillon. Si nous doublions tous les effectifs d'un tableau, les pourcentages resteraient les mêmes, mais le χ^2 doublerait et p diminuerait. C'est pourquoi le χ^2 et p ne disent pas à quel point l'association est **forte**. Pour cela, nous utilisons des mesures qui ne dépendent pas de N.

Pour un tableau 2 x 2, le coefficient phi :

$$\varphi = \sqrt{\frac{\chi^2}{N}}$$

Pour les tableaux plus grands, le coefficient V de Cramér :

$$V = \sqrt{\left[\frac{\chi^2}{N \times (m-1)}\right]}$$, où m est le plus petit de r et c

Tous deux prennent des valeurs comprises entre 0 (indépendance) et 1 (association parfaite). À titre indicatif, pour les tableaux avec m - 1 = 1, 0,1 correspond à une association faible, 0,3 à une association modérée et 0,5 à une association forte [76]. Sur mg2026 :

Tableau 7.14. Test du khi-deux et force de l'association pour quelques tableaux de mg2026.

Tableau	χ^2	ddl	p	φ ou V
obésité x diabète HGPO (2 x 2)	12,70	1	< 0,001	$\varphi = 0{,}28$
bras x perte ≥ 5% (3 x 2)	6,77	2	0,034	$V = 0{,}21$
sexe x statut tabagique (2 x 3)	0,54	2	0,76	$V = 0{,}06$
âge ≥ 50 ans x diabète HGPO (2 x 2)	4,36	1	0,037	$\varphi = 0{,}17$

La dernière ligne mérite un commentaire. Le diabète est survenu chez 2 des 34 patients de moins de 50 ans (5,9%) et chez 27 des 126 patients de 50 ans ou plus (21,4%). Le test sans correction donne p = 0,037, le test de Fisher p = 0,044 et la correction de Yates p = 0,066, car l'un des effectifs attendus n'est que de 6,2. Le résultat est à la limite, et l'association, mesurée par φ = 0,17, est faible. La conclusion honnête : un signal compatible avec une augmentation du risque de diabète avec l'âge, mais fragile dans cet échantillon.

Pour les tableaux 2 x 2 des études cliniques et épidémiologiques, les mesures les plus utiles sur le plan clinique ne sont pas φ ou V, mais celles qui ont des unités interprétables : la différence de proportions (avec IC à 95%), le risque relatif ou le rapport de prévalences, le rapport de cotes (OR) et le nombre de sujets à traiter. Le khi-deux répond à la question "existe-t-il une association ?", et ces mesures répondent à la question "quelle est son ampleur et dans quel sens va-t-elle ?". Un bon article rapporte les deux. Le chapitre 10 leur est consacré.

> **Étude de cas : les comportements alimentaires de commodité et les classes d'IMC chez les étudiants**
>
> Moleriu et ses collaborateurs, avec la Dre L. Moleriu, la Dre Lupuşoru et le Dr Muntean, ont analysé 921 étudiants de 18 à 24 ans [4]. Parmi eux, 24,1% étaient en surpoids ou obèses (19,4% en surpoids, 4,7% obèses). Les méthodes combinent exactement les outils de ce chapitre et du chapitre 10 : le test du khi-deux pour l'association entre les comportements alimentaires et la classe d'IMC, le V de Cramér pour la force de ces associations, puis une

> régression logistique multivariée avec des rapports de cotes ajustés (par exemple aOR 7,20, IC à 95% 4,26-12,17 et aOR 1,98, IC à 95% 1,26-3,10). Un score composite de risque alimentaire montre une relation dose-réponse : la proportion de personnes en excès pondéral passe de 19,8% pour un score de 0 à 50,0% pour un score de 4, avec un OR ajusté de 1,43 par point (IC à 95% 1,23-1,66 ; p < 0,001). Sur un échantillon de 921 personnes, même des associations faibles peuvent être statistiquement significatives, et le V de Cramér est précisément la mesure qui nous aide à distinguer les associations faibles des associations fortes. Le schéma est transversal : l'association n'établit pas le sens de la causalité, et certains comportements peuvent être la conséquence, et non la cause, de l'excès pondéral.

7.10 Rapporter les résultats des tests non paramétriques et du test du khi-deux

Un résultat bien rapporté permet au lecteur de refaire le raisonnement. Pour les tests sur les rangs, la description se fait par la médiane et l'écart interquartile (ou le minimum-maximum, pour de très petits échantillons), et non par la moyenne et l'écart-type. On donne ensuite le nom du test, la statistique, la valeur p exacte (avec trois décimales ; en dessous de 0,001, on écrit "p < 0,001") et, idéalement, une taille d'effet. Pour le khi-deux, on donne les effectifs absolus et les pourcentages (selon les lignes qui correspondent aux groupes comparés), le χ^2, les degrés de liberté, p et une mesure d'effet avec son intervalle de confiance.

Quelques modèles, sur les données de ce chapitre :

- "La satisfaction à 6 mois a été plus élevée dans le bras C que dans le bras A (médiane 4, EIQ 3-5, contre 3, EIQ 2-3 ; Mann-Whitney U = 2265,5, p < 0,001 ; corrélation bisériale des rangs 0,58)."
- "La baisse de la glycémie à jeun a différé entre les bras (Kruskal-Wallis H = 36,63, ddl 2, p < 0,001, ε^2 = 0,23) ; les comparaisons de Dunn avec correction de Bonferroni ont montré des différences entre A et B (p < 0,001) et entre A et C (p < 0,001), mais pas entre B et C (p = 0,12)."
- "La glycémie à jeun a diminué au cours de l'étude (médianes 104,5, 99,0 et 92,5 mg/dL à 0, 3 et 6 mois ; Friedman χ^2 = 74,67, ddl 2, p < 0,001 ; W de Kendall = 0,23)."
- "La proportion de patients ayant perdu au moins 5% de leur poids a différé entre les bras (11,3%, 30,2% et 29,6% ; χ^2 = 6,77, ddl 2, p = 0,034 ; V de Cramér = 0,21)."

Ce que nous ne faisons pas : nous n'écrivons pas "p = 0,000" (le logiciel arrondit, mais la probabilité n'est pas nulle), nous ne rapportons pas seulement "p < 0,05" lorsque nous disposons de la valeur exacte, nous ne présentons pas la moyenne de scores de Likert comme s'il s'agissait de mesures physiques et nous n'écrivons pas "le test a été non significatif, donc les groupes sont égaux".

Un exemple de compte rendu correct de scores ordinaux provient d'une étude qui vous concerne directement. Lungeanu et ses collaborateurs, avec la Dre Lupușoru, ont demandé à 121 étudiants en médecine, au moyen d'un questionnaire anonyme, comment ils évaluaient les cours liés aux compétences numériques et à la science des données [16]. Les appréciations, sur une échelle de 1 à 10, sont des variables ordinales, et les auteurs les ont rapportées exactement comme le recommande ce chapitre : médiane et EIQ. Un cours a obtenu une médiane de 9 (EIQ 7-10), le cours d'informatique médicale 7 (5-9) et la biostatistique 8 (5-9). Les relations entre les appréciations ont été évaluées par la corrélation de Spearman, qui travaille elle aussi sur les rangs : une corrélation très forte, R = 0,869 (p < 0,001), et une corrélation pratiquement nulle entre les connaissances mesurées et les connaissances autoévaluées, R = 0,107 (p = 0,246). Le second résultat est probablement le plus intéressant pour vous : le fait qu'un étudiant se sente prêt ne dit presque rien de ce qu'il sait réellement. Un résultat "non significatif" bien interprété est ici une information précieuse, et non un échec de l'étude.

7.11 Problèmes résolus

Pour les calculs à la main sur de petits échantillons, vous avez besoin des valeurs critiques des tests de Mann-Whitney et de Wilcoxon, qui ne figurent pas dans les tables classiques des distributions de l'annexe. Les tableaux 7.15 et 7.16 les donnent pour le test bilatéral à α = 0,05 ; elles ont été calculées à partir des distributions exactes des statistiques. La règle de décision est la même pour les deux : **on rejette H0 si la statistique calculée est inférieure ou égale à la valeur de la table**. Un tiret signifie que, pour ces tailles de groupes, aucune valeur ne peut atteindre la signification à ce seuil.

Tableau 7.15. Valeurs critiques de U de Mann-Whitney (test bilatéral, α = 0,05) ; n1 et n2 sont les effectifs des deux groupes.

n1 \ n2	3	4	5	6	7	8	9	10
3	-	-	0	1	1	2	2	3
4	-	0	1	2	3	4	4	5
5	0	1	2	3	5	6	7	8
6	1	2	3	5	6	8	10	11
7	1	3	5	6	8	10	12	14
8	2	4	6	8	10	13	15	17
9	2	4	7	10	12	15	17	20
10	3	5	8	11	14	17	20	23

Tableau 7.16. Valeurs critiques de T de Wilcoxon pour données appariées (test bilatéral, α = 0,05) ; n est le nombre de différences non nulles.

n	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
T critique	0	2	3	5	8	10	13	17	21	25	29	34	40	46	52

Problème résolu 7.1. Durée d'hospitalisation dans la pneumonie, test de Mann-Whitney à la main

Dans un service de médecine interne, 6 patients atteints de pneumonie communautaire sont passés précocement d'une antibiothérapie intraveineuse à une antibiothérapie orale, et 7 patients ont reçu le schéma intraveineux standard. La durée d'hospitalisation (jours) a été la suivante : relais précoce 4, 5, 5, 6, 8, 12 ; schéma standard 6, 7, 7, 9, 10, 11, 15. La durée d'hospitalisation diffère-t-elle entre les deux stratégies (α = 0,05) ?

Hypothèse de recherche. Le relais précoce par l'antibiotique oral raccourcit l'hospitalisation.

Hypothèses statistiques. H0 : les durées d'hospitalisation proviennent de la même distribution dans les deux groupes (Me1 = Me2). H1 : les distributions diffèrent (Me1 ≠ Me2), test bilatéral.

Choix du test. La variable est numérique, mais la durée d'hospitalisation est typiquement asymétrique (les valeurs de 15 jours et de 12 jours tirent la moyenne vers le haut), les groupes sont très petits (6 et 7) et indépendants. Nous choisissons le test U de Mann-Whitney.

Calcul. Série commune ordonnée (N = 13) : 4, 5, 5, 6, 6, 7, 7, 8, 9, 10, 11, 12, 15. Les rangs : 4 → 1 ; 5 et 5 → 2,5 ; 6 et 6 (une valeur de chaque groupe) → 4,5 ; 7 et 7 → 6,5 ; 8 → 8 ; 9 → 9 ; 10 → 10 ; 11 → 11 ; 12 → 12 ; 15 → 13.

R1 (relais précoce) = 1 + 2,5 + 2,5 + 4,5 + 8 + 12 = 30,5 R2 (standard) = 4,5 + 6,5 + 6,5 + 9 + 10 + 11 + 13 = 60,5

Vérification : $30{,}5 + 60{,}5 = 91 = \frac{13\times14}{2}$.

$U_1 = 30{,}5 - \frac{6\times7}{2} = 9{,}5$; $U_2 = 60{,}5 - \frac{7\times8}{2} = 32{,}5$; vérification 9,5 + 32,5 = 42 = 6 x 7. U = 9,5.

La valeur critique pour n1 = 6, n2 = 7, en bilatéral, α = 0,05 est 6. Comme 9,5 > 6, nous ne rejetons pas H0. La valeur p exacte est de 0,138. Taille d'effet : $r_{\text{rb}} = 1 - \frac{2\times9{,}5}{42} = 0{,}55$.

Conclusion statistique. Il n'existe pas de différence statistiquement significative entre les distributions de la durée d'hospitalisation (U = 9,5, p = 0,138).

Conclusion clinique. Les médianes sont de 5,5 jours contre 9 jours, et l'effet estimé est important (0,55), mais l'étude est trop petite pour le confirmer. Le résultat ne montre pas que les stratégies sont équivalentes ; il justifie une étude portant sur davantage de patients, dont l'effectif est calculé au préalable (chapitres 4 et 5).

Problème résolu 7.2. La douleur avant et après kinésithérapie, test de Wilcoxon et test des signes

Dix patients atteints de gonarthrose ont noté leur douleur (0-10) avant et après 4 semaines de kinésithérapie : avant 7, 6, 8, 5, 9, 6, 7, 8, 6, 7 ; après 4, 5, 5, 4, 6, 6, 3, 5, 4, 7. La douleur a-t-elle diminué ?

Hypothèses. H0 : la médiane des différences (avant - après) est égale à 0. H1 : médiane des différences ≠ 0.

Choix du test. Les mêmes patients, deux temps, donc des données appariées ; le score est ordinal. Nous choisissons le test de Wilcoxon pour données appariées.

Calcul. Les différences : 3, 1, 3, 1, 3, 0, 4, 3, 2, 0. Deux différences sont nulles et sont éliminées, n = 8. Les valeurs absolues ordonnées : 1, 1 (rangs 1 et 2, donc 1,5 chacune) ; 2 (rang 3) ; 3, 3, 3, 3 (rangs 4, 5, 6, 7, donc 5,5 chacune) ; 4 (rang 8). Toutes les différences non nulles sont positives, donc W+ = 1,5 + 1,5 + 3 + 4 x 5,5 + 8 = 36 et W- = 0. Vérification : $36 + 0 = \frac{8\times9}{2}$. T = 0.

La valeur critique pour n = 8, en bilatéral, α = 0,05 est 3. Comme 0 ≤ 3, nous rejetons H0 ; p exact = 0,008. Le test des signes donne ici le même résultat : 8 signes plus sur 8, p = 2 x 0,5□ = 0,008.

Conclusions. La douleur a diminué significativement après la kinésithérapie (médiane passant de 7 à 5 ; médiane des différences 2,5 points ; Wilcoxon T = 0, p = 0,008). Sur le plan clinique, une baisse de 2 à 3 points sur l'échelle 0-10 est habituellement perçue par le patient comme pertinente. L'absence de groupe témoin ne permet toutefois pas de séparer l'effet de la kinésithérapie de l'évolution naturelle ou de l'effet placebo.

Problème résolu 7.3. La cicatrisation des plaies avec trois types de pansement, test de Kruskal-Wallis

Le nombre de jours jusqu'à la cicatrisation complète d'une plaie superficielle a été enregistré chez 14 patients : pansement classique (n = 4) 12, 14, 15, 18 ; pansement hydrocolloïde (n = 5) 9, 10,

11, 13, 14 ; pansement à l'argent (n = 5) 7, 8, 8, 10, 12. Le délai de cicatrisation diffère-t-il entre les pansements ?

Hypothèses. H0 : la distribution du délai de cicatrisation est la même pour les trois pansements. H1 : au moins un pansement diffère.

Choix du test. Trois groupes indépendants, des échantillons très petits, une variable numérique de type "durée" : test de Kruskal-Wallis.

Calcul. Série commune ordonnée (N = 14) : 7, 8, 8, 9, 10, 10, 11, 12, 12, 13, 14, 14, 15, 18. Les rangs : 7 → 1 ; 8 et 8 → 2,5 ; 9 → 4 ; 10 et 10 → 5,5 ; 11 → 7 ; 12 et 12 → 8,5 ; 13 → 10 ; 14 et 14 → 11,5 ; 15 → 13 ; 18 → 14.

R1 (classique) = 8,5 + 11,5 + 13 + 14 = 47 R2 (hydrocolloïde) = 4 + 5,5 + 7 + 10 + 11,5 = 38 R3 (argent) = 1 + 2,5 + 2,5 + 5,5 + 8,5 = 20

Vérification : $47 + 38 + 20 = 105 = \frac{14\times15}{2}$.

$$H = \left[\frac{12}{14 \times 15}\right] \times \left(\frac{47^2}{4} + \frac{38^2}{5} + \frac{20^2}{5}\right) - 3 \times 15 = 0{,}05714 \times (552{,}25 + 288{,}8 + 80) - 45$$
$$= 0{,}05714 \times 921{,}05 - 45 = 7{,}63$$

Avec la correction pour les ex aequo, H = 7,70. Avec 2 degrés de liberté, 7,70 > 5,99, nous rejetons donc H0 (p = 0,021). Attention : un groupe ne compte que 4 patients, l'approximation par le khi-deux est donc à la limite de l'acceptable ; les logiciels qui proposent la version exacte doivent être préférés.

Post hoc (Dunn, Bonferroni). Les rangs moyens sont de 11,75 (classique), 7,6 (hydrocolloïde) et 4,0 (argent). Seule la différence classique contre argent reste significative après correction (z = 2,76, p ajusté = 0,017) ; classique contre hydrocolloïde p ajusté = 0,42, hydrocolloïde contre argent p ajusté = 0,52.

Conclusions. Le délai de cicatrisation a différé entre les pansements (médianes 14,5, 11 et 8 jours ; H = 7,70, p = 0,021), et la différence se situe entre le pansement classique et le pansement à l'argent. Sur 14 patients, le résultat est indicatif.

Problème résolu 7.4. La fatigue après la COVID-19 à trois temps, test de Friedman

Sept patients présentant des symptômes persistants après la COVID-19 ont noté leur fatigue (0-10) à 1, 3 et 6 mois après l'infection :

Tableau 7.17. Score de fatigue et rangs à l'intérieur de chaque patient (données fictives).

Patient	1 mois	3 mois	6 mois	Rangs
1	8	6	4	3; 2; 1
2	7	7	5	2,5; 2,5; 1
3	9	6	5	3; 2; 1
4	6	5	5	3; 1,5; 1,5
5	8	5	3	3; 2; 1
6	7	6	4	3; 2; 1
7	9	8	6	3; 2; 1
Rj				20,5; 14; 7,5

Hypothèses. H0 : la distribution du score de fatigue est la même aux trois temps. H1 : au moins un temps diffère.

Choix du test. Les mêmes patients, trois temps, un score ordinal : test de Friedman.

Calcul. Vérification : 20,5 + 14 + 7,5 = 42 = 7 x 6.

$$\chi^2 F = \left[\frac{12}{7 \times 3 \times 4}\right] \times (20{,}5^2 + 14^2 + 7{,}5^2) - 3 \times 7 \times 4$$
$$= 0{,}1429 \times (420{,}25 + 196 + 56{,}25) - 84 = 0{,}1429 \times 672{,}5 - 84 = 12{,}07$$

Avec la correction pour les ex aequo, 13,00 ; ddl 2 ; 13,00 > 5,99, nous rejetons donc H0 (p = 0,0015). Les médianes : 8, 6 et 5.

Post hoc. Test de Wilcoxon pour chaque paire : 1 contre 3 mois p = 0,031 ; 3 contre 6 mois p = 0,031 ; 1 contre 6 mois p = 0,016. Avec Bonferroni (x 3) : 0,094, 0,094 et 0,047. Seule la comparaison 1 contre 6 mois reste significative. Avec 7 patients, la plus petite valeur p que puisse donner le test de Wilcoxon exact bilatéral est de 0,016, les comparaisons post hoc ont donc une puissance très limitée.

Conclusions. La fatigue a diminué significativement au cours des 6 premiers mois (Friedman χ^2 = 13,00, p = 0,0015) ; c'est la différence entre 1 et 6 mois qui résiste à la correction. Sans groupe de comparaison, nous ne pouvons attribuer la baisse à aucun traitement.

Problème résolu 7.5. Tabagisme et BPCO, khi-deux 2 x 2 complet

Dans une étude transversale menée dans des cabinets de médecine générale, 300 personnes de plus de 40 ans ont bénéficié d'une spirométrie : sur les 120 fumeurs, 30 avaient une BPCO ; sur les 180 non-fumeurs, 15 avaient une BPCO. Le tabagisme est-il associé à la BPCO ?

Hypothèses. H0 : la proportion de BPCO est la même chez les fumeurs et chez les non-fumeurs (p1 = p2). H1 : p1 ≠ p2.

Choix du test. Deux variables nominales dichotomiques, des groupes indépendants : khi-deux sur un tableau 2 x 2 (nous vérifions les conditions).

Tableau 7.18. Tabagisme et BPCO : observé (attendu).

	BPCO oui	**BPCO non**	**Total**
Fumeurs	30 (18)	90 (102)	120
Non-fumeurs	15 (27)	165 (153)	180
Total	45	255	300

Effectifs attendus. $E_{11} = \frac{120 \times 45}{300} = 18$; $E_{12} = \frac{120 \times 255}{300} = 102$; $E_{21} = \frac{180 \times 45}{300} = 27$; $E_{22} = \frac{180 \times 255}{300} = 153$. Tous sont ≥ 5, le test est valide.

La statistique. |O - E| = 12 dans chaque cellule.

$$\chi^2 = \frac{144}{18} + \frac{144}{102} + \frac{144}{27} + \frac{144}{153} = 8{,}000 + 1{,}412 + 5{,}333 + 0{,}941 = 15{,}69$$

ddl = 1 ; valeur critique 3,84 ; p < 0,001. Avec la correction de Yates : (12 - 0,5)² = 132,25, et χ^2Yates = $132{,}25 \times \left(\frac{1}{18} + \frac{1}{102} + \frac{1}{27} + \frac{1}{153}\right) = 14{,}41$, p < 0,001. Le test de Fisher donne lui aussi p < 0,001. Le coefficient φ = √(15,69/300) = 0,23.

Mesures d'effet. Prévalence de la BPCO : 25,0% chez les fumeurs, 8,3% chez les non-fumeurs. Rapport de prévalences = $\frac{0{,}250}{0{,}083} = 3{,}0(;\ \text{IC}\,95\%\ 1{,}69 - 5{,}33)$; $\text{OR} = \frac{30 \times 165}{90 \times 15} = 3{,}67(;\ \text{IC}\,95\%\ 1{,}87 - 7{,}17)$.

Conclusions. Il existe une association significative entre le tabagisme et la BPCO (χ^2 = 15,69, p < 0,001) ; la prévalence de la BPCO est trois fois plus élevée chez les fumeurs. L'association est d'intensité faible à modérée selon φ, mais cliniquement importante, car la BPCO est fréquente. Le schéma transversal n'établit pas à lui seul la causalité, mais le résultat concorde avec les données probantes des études de cohorte.

Problème résolu 7.6. Statut tabagique et diabète dans mg2026, tableau 3 x 2

Dans le jeu de données mg2026, le diabète confirmé par HGPO est survenu chez 3 des 33 anciens fumeurs, 10 des 53 fumeurs et 16 des 74 non-fumeurs. Le statut tabagique est-il associé au diabète à l'inclusion ?

Hypothèses. H0 : la proportion de diabète est la même dans les trois catégories de tabagisme. H1 : au moins une proportion diffère.

Calcul des effectifs attendus (29 diabétiques sur 160, soit 18,1%) : anciens fumeurs $\frac{33\times29}{160} = 5{,}98$; fumeurs $\frac{53\times29}{160} = 9{,}61$; non-fumeurs $\frac{74\times29}{160} = 13{,}41$; dans la colonne "sans diabète" : 27,02, 43,39 et 60,59. Tous sont ≥ 5 (le plus petit étant 5,98), le test est donc valide.

Les contributions : anciens fumeurs (3 - 5,98)²/5,98 = 1,486 et (30 - 27,02)²/27,02 = 0,329 ; fumeurs (10 - 9,61)²/9,61 = 0,016 et (43 - 43,39)²/43,39 = 0,004 ; non-fumeurs (16 - 13,41)²/13,41 = 0,499 et (58 - 60,59)²/60,59 = 0,111. χ^2 = 2,44, ddl = (3 - 1)(2 - 1) = 2, p = 0,29. V de Cramér $= \sqrt{\frac{2{,}44}{160}} = 0{,}12$.

Conclusions. Nous n'avons pas trouvé d'association significative entre le statut tabagique et le diabète (9,1%, 18,9% et 21,6% ; χ^2 = 2,44, p = 0,29). Le pourcentage plus faible chez les anciens fumeurs repose sur seulement 3 cas ; avec 33 personnes dans cette catégorie, l'estimation est très imprécise. Le tabagisme reste un facteur de risque cardiovasculaire avéré ; ce petit échantillon n'a pas la puissance nécessaire pour mettre en évidence une association avec le diabète.

Problème résolu 7.7. Deux tests rapides chez les mêmes patients, test de McNemar

Chez 100 patients atteints de pharyngite, deux tests rapides de détection du streptocoque (A et B) ont été appliqués. Les deux positifs : 40 ; A positif et B négatif : 15 ; A négatif et B positif : 5 ; les deux négatifs : 40. La proportion de résultats positifs diffère-t-elle entre les deux tests ?

Choix du test. Chaque patient a deux résultats binaires, il s'agit donc de données appariées : test de McNemar. Le test du khi-deux habituel serait erroné, car les mêmes patients apparaîtraient dans les deux lignes.

Calcul. Les proportions de positifs : $\frac{A\,55}{100} = 55\%$, $\frac{B\,45}{100} = 45\%$. Les cellules discordantes : b = 15, c = 5.

$\chi^2 = \frac{(15-5)^2}{15+5} = \frac{100}{20} = 5{,}00$, ddl 1, p = 0,025 ; avec correction de continuité : (10 - 1)²/20 = 4,05, p = 0,044 ; version exacte (binomiale) p = 0,041

Conclusions. Le test A donne significativement plus de résultats positifs que le test B. Le test de McNemar ne dit pas quel test est le meilleur ; pour cela, il faudrait comparer les deux au standard de référence (la culture du prélèvement de gorge), c'est-à-dire calculer la sensibilité et la spécificité (chapitres 11 et 12).

Problème résolu 7.8. Le délirium postopératoire dans une petite étude, test exact de Fisher

Dans une étude pilote chez des personnes âgées opérées, un délirium postopératoire est survenu chez 1 des 12 patients ayant reçu un nouveau sédatif et chez 6 des 10 patients ayant reçu les soins habituels. La fréquence du délirium diffère-t-elle ?

Vérification des conditions. Le total des patients avec délirium est de 7 sur 22. Les effectifs attendus : $\frac{12\times7}{22} = 3{,}82$ et $\frac{10\times7}{22} = 3{,}18$ dans la colonne "délirium" ; 8,18 et 6,82 dans la colonne "sans délirium". Deux cellules sur quatre ont E < 5, nous utilisons donc le test exact de Fisher.

Calcul. Avec les totaux fixés (12 et 10 en lignes, 7 et 15 en colonnes), le nombre a de cas de délirium dans le groupe sédatif peut aller de 0 à 7. Les probabilités hypergéométriques sont : P(0) = 0,0007 ; P(1) = 0,0148 (tableau observé) ; P(2) = 0,0975 ; P(3) = 0,2709 ; P(4) = 0,3483 ; P(5) = 0,2090 ; P(6) = 0,0542 ; P(7) = 0,0046. Nous additionnons les tableaux au plus aussi probables que le tableau observé : 0,0007 + 0,0148 + 0,0046 = 0,020.

Conclusions. Fisher p = 0,020 : le délirium a été significativement plus rare dans le groupe sous nouveau sédatif (8,3% contre 60,0%). Le test du khi-deux appliqué à tort aurait donné p = 0,010, et avec la correction de Yates p = 0,033 ; la conclusion ne change pas ici, mais la valeur correcte est la valeur exacte. Avec 22 patients, l'effet doit être confirmé dans un essai randomisé d'effectif adéquat.

7.12 Exercices proposés

1. Attribuez des rangs aux valeurs 12, 15, 15, 9, 20, 15, 11, 9 (dans l'ordre donné) et vérifiez la somme des rangs.
2. Le score d'anxiété (0-10) avant une endoscopie était le suivant : groupe avec information vidéo (n = 4) 3, 4, 4, 6 ; groupe sans information (n = 5) 5, 7, 8, 8, 9. Calculez R1, R2, U1, U2 et U. La valeur critique pour n1 = 4, n2 = 5, en bilatéral, α = 0,05, est 1. Que décidez-vous ?
3. Pression systolique (mmHg) chez 8 patients, avant et après 4 semaines de traitement : avant 140, 152, 138, 160, 145, 150, 148, 155 ; après 132, 150, 140, 148, 139, 150, 141, 146. Appliquez le test de Wilcoxon à la main (la valeur critique pour n = 7 est 2). Que remarquez-vous chez le patient 6 ?
4. Dans une étude à trois groupes de 6 patients chacun (N = 18), les sommes des rangs sont de 40, 60 et 71. Vérifiez la somme des rangs, calculez H et décidez (α = 0,05).
5. Dix patients ont été évalués à trois temps ; les sommes des rangs à l'intérieur des patients sont de 14, 20 et 26. Calculez la statistique de Friedman et décidez (α = 0,05). Calculez le W de Kendall.
6. Au cours d'une saison grippale, 10 des 200 personnes vaccinées et 24 des 150 personnes non vaccinées ont eu une grippe confirmée. Construisez le tableau, calculez les effectifs attendus, le χ^2 (avec et sans Yates), le risque relatif et φ.
7. Un tableau 3 x 3 a les effectifs observés suivants : ligne 1 : 8, 12, 5 ; ligne 2 : 10, 20, 6 ; ligne 3 : 2, 4, 3. Calculez les effectifs attendus pour la ligne 3. Le test du khi-deux est-il applicable ? Quelles solutions proposez-vous ?
8. Chez 80 patients, la présence d'une hypertension a été évaluée avant et après un programme de perte de poids. Au total, 12 patients sont passés de "hypertendu" à "normotendu", et 4 de "normotendu" à "hypertendu". Quel test utilisez-vous ? Calculez la statistique, avec et sans correction.

9. Dans un échantillon de 200 donneurs, les groupes sanguins O, A, B, AB ont été observés respectivement 70, 85, 30 et 15 fois. Testez l'ajustement aux proportions hypothétiques de 35%, 40%, 17%, 8%.
10. Un effet indésirable est survenu chez 3 des 12 patients sous médicament X et chez 0 des 12 patients sous placebo. Calculez les effectifs attendus, choisissez le test et interprétez la valeur p donnée par le logiciel (Fisher bilatéral p = 0,217).
11. Sur le jeu de données mg2026, dans JASP, comparez la satisfaction à 6 mois entre les bras B et A. Rapportez les médianes, U, p et la corrélation bisériale des rangs.

Pièges fréquents

- Appliquer le test de Mann-Whitney ou le test du khi-deux habituel à des données appariées (les mêmes patients mesurés deux fois). Pour des données appariées : test de Wilcoxon et test de McNemar, respectivement.
- Rapporter la moyenne et l'écart-type de scores de Likert, puis appliquer un test sur les rangs : la description doit correspondre au test (médiane et EIQ).
- Vérifier la condition d'application du test du khi-deux sur les effectifs observés au lieu des effectifs attendus.
- Saisir des pourcentages au lieu des effectifs absolus dans le tableau de contingence.
- Conclure que "les médianes sont égales" après un test non significatif ; formulation correcte : "nous n'avons pas trouvé de différence significative".
- Faire plusieurs tests de Mann-Whitney entre trois groupes, chacun au seuil de 0,05, sans test global et sans correction.
- Comparer U ou T à la valeur critique dans le mauvais sens : pour les tests de Mann-Whitney et de Wilcoxon, on rejette H0 lorsque la statistique est **inférieure ou égale** à la valeur critique.
- Interpréter un χ^2 élevé comme une "association forte" : la force se lit sur φ, le V de Cramér ou le RR/OR, et non sur p.

À retenir

- Les tests non paramétriques travaillent sur les rangs ; ils conviennent aux données ordinales, aux distributions asymétriques sur de petits échantillons et aux valeurs extrêmes réelles. Leur efficacité sur des données normales est d'environ 95%.
- Mann-Whitney : 2 groupes indépendants ; Wilcoxon : 2 mesures appariées ; Kruskal-Wallis : ≥ 3 groupes indépendants (post hoc de Dunn) ; Friedman : ≥ 3 mesures répétées (post hoc de Wilcoxon avec correction).
- Les ex aequo reçoivent le rang moyen ; la somme des rangs est toujours égale à $N(N + 1)/2$.
- Effectif attendu : E = total de la ligne x total de la colonne/N ; ddl = (r - 1)(c - 1).
- Khi-deux valide : E ≥ 5 dans toutes les cellules d'un tableau 2 x 2 ; dans les tableaux plus grands, au plus 20% des cellules avec E < 5 et aucune avec E < 1. Sinon : test de Fisher ou regroupement de catégories.
- Test de McNemar pour les proportions appariées, test de tendance pour les catégories ordonnées, test d'ajustement pour une seule variable comparée à une distribution théorique.
- Compte rendu complet : description (médiane et EIQ, ou n et %), statistique de test, ddl, p exact et une taille d'effet (r_{rb}, ε^2, W, φ, V, RR ou OR avec IC à 95%).

Questions d'autoévaluation

1. (QRU) Une étude compare le score de qualité de vie (échelle ordinale 0-100, distribution fortement asymétrique) entre 15 patients opérés par cœlioscopie et 15 patients opérés par voie classique. Le test adapté est :

   a) le test t apparié
   b) le test de Wilcoxon pour données appariées
   c) le test U de Mann-Whitney
   d) le test de Friedman
   e) le test de McNemar

2. (QRU) Les valeurs 4, 7, 7, 2, 7, 9 reçoivent, dans l'ordre donné, les rangs :
   a) 2; 3; 4; 1; 5; 6
   b) 2; 4; 4; 1; 4; 6
   c) 2; 3,5; 3,5; 1; 3,5; 6
   d) 2; 4; 4; 1; 5; 6
   e) 1; 4; 4; 2; 4; 6

3. (QRU) Un tableau de contingence comporte 4 lignes et 3 colonnes. Le nombre de degrés de liberté pour le test du khi-deux est :
   a) 12
   b) 11
   c) 7
   d) 6
   e) 5

4. (QRU) Dans un tableau 2 x 2, le total de la première ligne est de 40, le total de la première colonne de 30 et le total général de 200. L'effectif attendu dans la cellule située à leur intersection est :
   a) 6
   b) 7,5
   c) 12
   d) 15
   e) 30

5. (QRU) Chez 60 patients, la présence d'une douleur thoracique a été enregistrée avant et après traitement. Le test adapté pour comparer les proportions est :
   a) le khi-deux de Pearson sur un tableau "temps x douleur" comptant 120 observations
   b) le test exact de Fisher
   c) le test de McNemar
   d) le test de Kruskal-Wallis
   e) le test de tendance de Cochran-Armitage

6. (QRM) À propos du test de Kruskal-Wallis, les affirmations vraies sont :
   a) c'est l'alternative non paramétrique de l'ANOVA à un facteur
   b) les rangs sont attribués à l'intérieur de chaque sujet
   c) la statistique H est comparée à la distribution du khi-deux à k - 1 degrés de liberté
   d) un résultat significatif montre exactement quels groupes diffèrent entre eux
   e) après un résultat significatif, on peut faire des comparaisons de Dunn avec correction pour comparaisons multiples

7. (QRM) Le test du khi-deux sur un tableau 2 x 2 n'est PAS adapté (sans autres mesures) si :
   a) un effectif attendu est de 3,2
   b) un effectif observé est égal à 0 et que tous les effectifs attendus sont supérieurs à 5
   c) les mêmes patients sont comptés à la fois dans la ligne "avant" et dans la ligne "après"

d) le tableau contient des pourcentages au lieu de nombres de patients

e) les groupes ont des effectifs inégaux (300 et 80 patients)

8. (QRM) Dans le jeu de données mg2026, le test de Mann-Whitney pour la satisfaction, bras C contre bras A, a donné U = 2265,5, $p < 0,001$, corrélation bisériale des rangs 0,58. Les affirmations correctes sont :

   a) la satisfaction a été plus élevée dans le bras C

   b) l'effet est important

   c) les médianes sont de 4 (EIQ 3-5) et de 3 (EIQ 2-3)

   d) le résultat prouve que le bras C est plus efficace pour le contrôle de la glycémie

   e) la probabilité qu'un patient du bras C soit plus satisfait qu'un patient du bras A est d'environ 0,79

9. (QRM) À propos du test de Friedman, les affirmations vraies sont :

   a) il s'applique à ≥ 3 mesures chez les mêmes sujets

   b) les rangs sont attribués dans l'échantillon commun de toutes les mesures

   c) le coefficient W de Kendall peut être utilisé comme taille d'effet

   d) les comparaisons post hoc peuvent se faire avec le test de Wilcoxon et une correction de Bonferroni

   e) c'est l'alternative non paramétrique de l'ANOVA à mesures répétées

10. (QRM) Pour le tableau obésité x diabète de mg2026 ($\chi^2 = 12,70$, $p < 0,001$, $\varphi = 0,28$), les affirmations correctes sont :

a) l'association est statistiquement significative

b) φ indique une association d'intensité faible à modérée

c) la faible valeur p montre que l'association est très forte

d) le schéma transversal ne permet pas de conclure que l'obésité a causé le diabète

e) l'effectif attendu pour la cellule "obèse avec diabète" est de 9,79

Chapitre 8. Sélection et application des tests statistiques : critères de choix, résolution de problèmes et études de cas

Objectifs

À la fin de ce chapitre, l'étudiant sera capable de :

- identifier le but d'une analyse (différence, association, concordance, prédiction) à partir de la question de recherche ;
- classer la variable résultat et la variable de regroupement et reconnaître le schéma d'étude (groupes indépendants ou données appariées, nombre de groupes ou de temps) ;
- appliquer pas à pas l'algorithme de choix du test statistique et justifier ce choix, y compris l'alternative non paramétrique ou catégorielle ;
- reconnaître l'unité d'analyse correcte et les situations dans lesquelles les observations ne sont pas indépendantes ;
- expliquer pourquoi le test doit être fixé avant l'analyse des données et ce qui se passe lorsqu'il est choisi d'après la valeur p ;
- reconnaître les erreurs les plus fréquentes de choix du test dans des descriptions de type "Méthodes" et proposer le test correct ;
- rédiger la résolution d'un problème selon le format standardisé de la discipline (a-g), de l'hypothèse de recherche à la conclusion clinique ;
- évaluer de façon critique l'adéquation entre le schéma d'étude, le type de variable et le test rapporté dans des articles publiés par des collègues de la discipline.

8.1 Le choix du test est une décision de conception, non de calcul

Les logiciels statistiques modernes calculent n'importe quel test en quelques secondes. JASP, MedCalc ou Epi Info ne protesteront pas si vous leur demandez un test t sur des données qui sont en réalité appariées ou un khi-deux sur un tableau dont les effectifs attendus sont de 1,5. Le résultat aura une apparence impeccable, avec trois décimales et un graphique en couleurs, et il sera faux. C'est pourquoi la partie difficile de l'analyse statistique n'est pas le calcul, mais le choix : quelle question posons-nous, de quel type de données disposons-nous, quel test correspond à cette combinaison.

Aux chapitres 5, 6 et 7, nous avons vu les tests un par un. Ce chapitre les rassemble. Nous allons construire un algorithme de décision, l'appliquer à 22 problèmes résolus sur le jeu de données mg2026, nous exercer à reconnaître les tests erronés et lire les sections méthodes des articles de nos collègues de la discipline. À l'examen pratique, la consigne typique est exactement celle-ci : "Quel test statistique appliquez-vous pour savoir si... ?", suivie de la résolution complète au format a-g. L'expérience des travaux pratiques montre que les étudiants qui échouent à l'examen ne se trompent généralement pas dans la lecture de la valeur p, mais dès la première étape : ils ont confondu des données appariées avec des groupes indépendants ou une variable ordinale avec une variable numérique.

Le choix du test se fait idéalement avant le recueil des données, dans le protocole de l'étude, en même temps que la définition du critère de jugement principal et le calcul de la taille de l'échantillon [18,56]. Le test découle du schéma d'étude, et le schéma découle de la question. Si la

question est bien formulée ("la baisse de la glycémie à 6 mois diffère-t-elle entre les trois bras ?"), le test se choisit presque tout seul.

8.2 L'algorithme de choix du test

La figure 8.1 résume l'algorithme utilisé dans cet ouvrage. Ce n'est pas le seul possible, mais il couvre la grande majorité des situations de la recherche clinique de base.

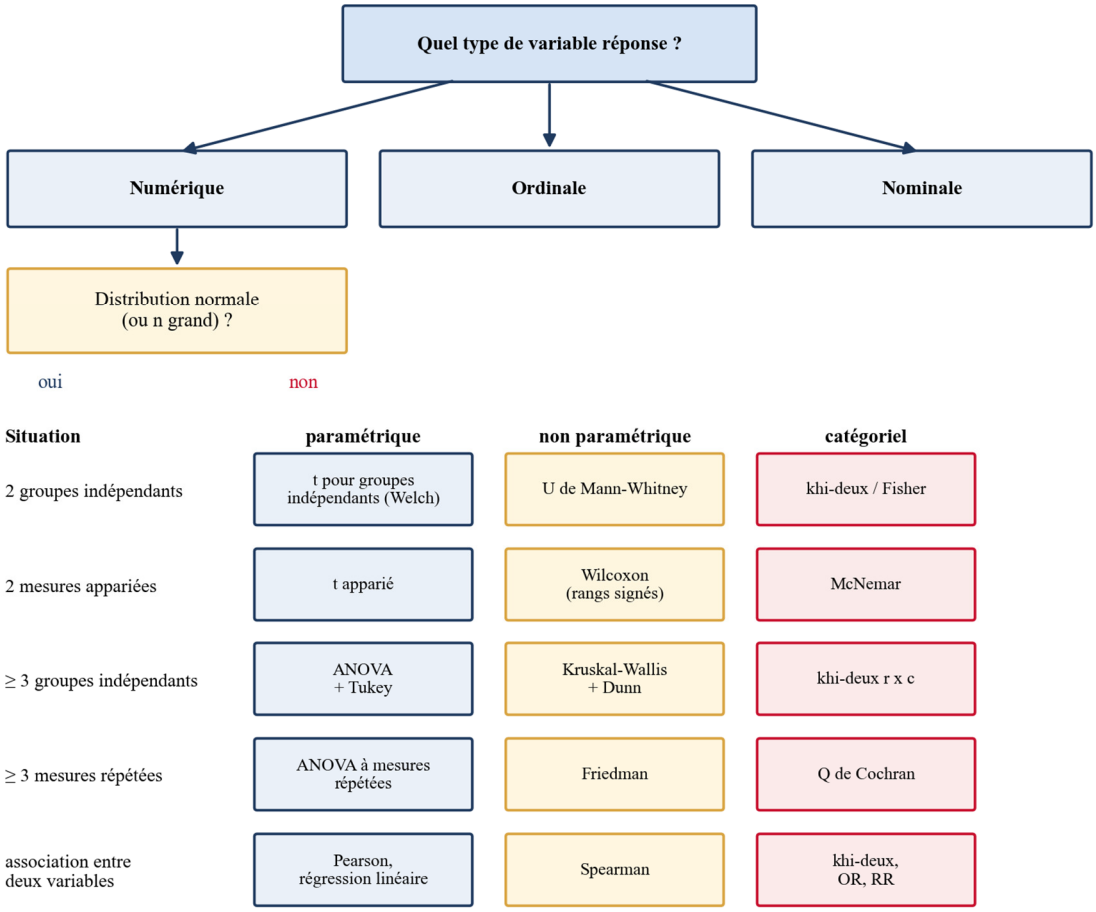


Figure 8.1. L'algorithme de choix du test statistique : type de la variable résultat (numérique, ordinale, nominale), puis situation (2 groupes indépendants, 2 mesures appariées, ≥ 3 groupes indépendants, ≥ 3 mesures répétées, association de deux variables) et, pour les variables numériques, distribution (normale ou grand n, par opposition à une asymétrie avec un petit n). Les colonnes donnent la version paramétrique, non paramétrique et catégorielle.

L'algorithme se parcourt en cinq étapes. L'ordre compte : une erreur commise à une étape initiale ne peut pas être réparée plus tard.

8.2.1 Étape 1 : quel est le but de l'analyse ?

La première question ne porte pas sur les données, mais sur l'intention. Il existe quatre grands buts, et chacun exige une famille différente de méthodes :

- la **différence** entre des groupes ou entre des temps : "La glycémie est-elle plus basse dans le bras C que dans le bras A ?" (tests t, ANOVA, Mann-Whitney, Kruskal-Wallis, Wilcoxon, Friedman, khi-deux, McNemar) ;
- l'**association** entre deux variables mesurées chez les mêmes sujets : "La pression systolique augmente-t-elle avec l'IMC ?" (corrélation de Pearson ou de Spearman, khi-deux pour les variables qualitatives, OR et RR) ;

- la **concordance** entre deux méthodes ou deux évaluateurs qui mesurent la même chose : "Le tensiomètre automatique et le tensiomètre manuel donnent-ils les mêmes valeurs ?" (méthode de Bland-Altman [102], coefficient de corrélation intraclasse, kappa de Cohen [103]) ;
- la **prédiction** d'un résultat à partir d'un ou de plusieurs facteurs : "Dans quelle mesure le FINDRISC prédit-il le diabète ?" (régression linéaire ou logistique, courbe ROC).

La confusion la plus fréquente se fait entre association et concordance. Deux tensiomètres peuvent avoir une corrélation de 0,95 et l'un d'eux donner pourtant systématiquement 10 mmHg de plus. La corrélation mesure si les valeurs augmentent ensemble ; la concordance mesure si elles sont identiques. Pour la concordance, on n'utilise ni le test t apparié seul ni la corrélation, mais l'analyse des différences (moyenne des différences et limites de concordance).

8.2.2 Étape 2 : quel est le type de la variable résultat ?

La variable résultat (dépendante) est celle que nous voulons comparer ou expliquer : la glycémie, la satisfaction, la présence d'un diabète. La variable de regroupement (indépendante) définit les groupes : le bras, le sexe, le milieu de résidence. Le type de la variable résultat détermine la colonne de la figure 8.1.

- **Numérique** (continue, ou discrète avec de nombreuses valeurs) : glycémie, PAS, poids, IMC, âge, score FINDRISC 0-26.
- **Ordinale** : satisfaction 1-5, niveau d'études 1-4, activité physique autoévaluée 1-5, stade d'une maladie.
- **Nominale** (y compris dichotomique) : diabète oui/non, perte ≥ 5% oui/non, statut tabagique, sexe.

Deux situations demandent de l'attention. Un score comportant de nombreuses valeurs (FINDRISC 0-26, un score de qualité de vie 0-100) est habituellement traité comme numérique, si la distribution le permet. Une variable numérique dichotomisée (glycémie ≥ 100 mg/dL oui/non) devient nominale, mais perd de l'information et de la puissance ; la dichotomisation ne se justifie que lorsque le seuil a un sens clinique et a été fixé à l'avance.

8.2.3 Étape 3 : combien de groupes ou combien de temps ?

Nous comptons les groupes ou les temps comparés : un seul (comparaison à une valeur de référence), deux, ou trois et plus. La distinction entre "deux" et "trois ou plus" n'est pas une formalité. Avec trois groupes, trois tests t séparés font passer le risque global d'erreur de type I de 5% à 14,3% ($1 - 0{,}95^3$) ; avec cinq groupes, les dix comparaisons possibles le portent à 40%. C'est pourquoi, pour trois groupes ou plus, nous utilisons un test global (ANOVA, Kruskal-Wallis, khi-deux r x c) et seulement ensuite des comparaisons post hoc avec correction.

8.2.4 Étape 4 : groupes indépendants ou données appariées ?

Les données sont **appariées** (dépendantes) lorsque chaque valeur d'une série a une correspondante naturelle dans l'autre : le même patient avant et après, l'œil droit et l'œil gauche, deux méthodes sur le même prélèvement, des paires appariées dans les études cas-témoins. Les données sont **indépendantes** lorsque les groupes sont formés de sujets différents, sans lien entre eux. La question de contrôle est simple : "puis-je relier par un trait chaque valeur de la première colonne à une valeur précise de la seconde ?" Si oui, les données sont appariées.

La différence de résultat peut être importante. Dans mg2026, la glycémie est passée de 105,44 mg/dL à l'inclusion à 95,25 mg/dL à 6 mois. Le test t apparié (correct) donne t = 11,14, avec une erreur standard de la différence de 0,91 mg/dL. Si nous traitons à tort les deux séries comme deux groupes indépendants (test de Welch), l'erreur standard monte à 2,03 mg/dL et t descend à 5,03. La raison en est la forte corrélation entre la glycémie d'un patient à 0 et à 6 mois (r = 0,82) : le test apparié élimine la variabilité entre patients et ne conserve que la variabilité du

changement. Ici, les deux versions sont significatives, mais sur un échantillon plus petit, le test erroné aurait pu manquer un effet réel.

8.2.5 Étape 5 : la distribution et la taille de l'échantillon

Pour les variables numériques seulement, la dernière question est de savoir si nous pouvons utiliser un test paramétrique. Les critères, abordés aux chapitres 3 et 6 :

- la forme de la distribution dans chaque groupe (histogramme, diagramme Q-Q, asymétrie) et, à titre indicatif, le test de Shapiro-Wilk [66] ; pour le test t apparié, on vérifie la distribution des **différences**, et non celle des deux séries ;
- la taille de l'échantillon : avec au moins 30 à 40 sujets par groupe et une asymétrie modérée, le test t et l'ANOVA restent robustes ; avec de petits groupes, la distribution compte beaucoup plus ;
- l'homogénéité des variances (test de Levene) ; si elle n'est pas respectée, on utilise la version de Welch du test t ou de l'ANOVA, et non automatiquement un test non paramétrique ;
- les valeurs extrêmes réelles et les variables bornées (scores avec effet plafond).

Un test de Shapiro-Wilk "significatif" sur un grand échantillon n'oblige pas à recourir à un test non paramétrique. Dans mg2026, la glycémie à l'inclusion a un SW p = 0,037 pour n = 160, mais l'asymétrie n'est que de 0,37 et l'histogramme montre une distribution presque symétrique. Pour 160 patients, le test t est parfaitement adapté. À l'inverse, un SW "non significatif" sur 8 patients ne garantit pas la normalité, il montre seulement la faible puissance du test de normalité.

8.2.6 L'algorithme en bref

Pour travailler rapidement en travaux pratiques, l'algorithme peut s'écrire sous forme de liste de vérification :

1. Écrivez la question de recherche en une phrase. Que voulez-vous savoir : différence, association, concordance ou prédiction ?
2. Identifiez la variable résultat et son type (numérique, ordinale, nominale).
3. Identifiez la variable de regroupement ou les temps : combien de catégories comporte-t-elle ?
4. Vérifiez si les données sont indépendantes ou appariées.
5. Pour une variable numérique : vérifiez la distribution dans chaque groupe (ou celle des différences) et l'homogénéité des variances ; choisissez entre la version paramétrique, la version de Welch ou la version non paramétrique.
6. Pour des variables qualitatives : calculez les effectifs attendus et choisissez entre le khi-deux, la correction de Yates, le test de Fisher ou le test de McNemar.
7. Pour trois groupes ou plus : test global, puis comparaisons post hoc avec correction.
8. Fixez à l'avance la taille d'effet que vous rapporterez (différence de moyennes avec IC à 95%, d de Cohen, r_rb, RR, OR, φ, V).

8.3 Le tableau de synthèse des tests

Le tableau 8.1 rassemble en un même endroit les situations, les tests et l'endroit où vous les trouverez dans les logiciels utilisés en travaux pratiques (JASP 0.98.1 [47], MedCalc 24.0 [50], Epi Info 7.2.7 [51]). Les noms des menus peuvent différer légèrement d'une version à l'autre ; les étapes sont décrites en détail dans les travaux pratiques 6, 7 et 8.

Tableau 8.1. Synthèse des tests statistiques usuels et des menus des logiciels.

Situation	Variable numérique, distribution normale (paramétrique)	Variable ordinale ou numérique asymétrique (non paramétrique)	Variable nominale (catégorielle)	JASP	MedCalc	Epi Info
1 échantillon comparé à une valeur de référence	t pour un échantillon	Wilcoxon pour un échantillon	test binomial ; khi-deux d'ajustement	T-Tests > One Sample T-Test; Frequencies > Binomial Test, Multinomial Test	One sample t-test; Signed rank sum test	StatCalc: Binomial
2 groupes indépendants	t non apparié (Student ; Welch si variances inégales)	U de Mann-Whitney	khi-deux 2 x 2 ; Fisher si E < 5	T-Tests > Independent Samples T-Test (Student, Welch, Mann-Whitney); Frequencies > Contingency Tables	Independent samples t-test; Mann-Whitney test; Chi-squared test; Fisher's exact test	Classic Analysis: MEANS avec variable de regroupement; TABLES; StatCalc: Tables 2 x 2
2 mesures appariées	t apparié	Wilcoxon (signes et rangs) ; test des signes	McNemar	T-Tests > Paired Samples T-Test (Student, Wilcoxon signed-rank)	Paired samples t-test; Wilcoxon test (paired samples); McNemar test	StatCalc: paires appariées (McNemar)
≥ 3 groupes indépendants	ANOVA à un facteur + Tukey (Welch si variances inégales)	Kruskal-Wallis + Dunn	khi-deux r x c	ANOVA > ANOVA (Post Hoc Tests; Nonparametrics: Kruskal-Wallis, Dunn)	One-way analysis of variance; Kruskal-Wallis test	MEANS (ANOVA et Kruskal-Wallis); TABLES
≥ 3 mesures répétées	ANOVA à mesures répétées	Friedman + Wilcoxon avec correction	Q de Cochran	ANOVA > Repeated Measures ANOVA (Nonparametrics: Friedman)	Repeated measures analysis of variance; Friedman test; Cochran's Q test	il n'existe pas de procédure dédiée
proportion selon des catégories ordonnées	-	Mann-Whitney ou Kruskal-Wallis sur la variable ordinale	khi-deux de tendance	Frequencies > Contingency Tables	Chi-squared test for trend	StatCalc: Chi square for trend
association de deux variables numériques ou ordinales	corrélation de Pearson	corrélation de Spearman, de Kendall	-	Regression > Correlation	Correlation; Rank correlation	Classic Analysis: REGRESS
association de deux variables nominales	-	-	khi-deux, Fisher ; φ, V de Cramér ; RR, OR avec IC à 95%	Frequencies > Contingency Tables (Phi and Cramér's V; Log odds	Chi-squared test; Relative risk; Odds ratio	TABLES; StatCalc: Tables 2 x 2

Situation	Variable numérique, distribution normale (paramétrique)	Variable ordinale ou numérique asymétrique (non paramétrique)	Variable nominale (catégorielle)	JASP	MedCalc	Epi Info
				ratio)		
concordance de deux méthodes ou de deux évaluateurs	Bland-Altman ; coefficient de corrélation intraclasse	kappa pondéré	kappa de Cohen	module Reliability	Bland-Altman plot; Intraclass correlation; Inter-rater agreement (kappa)	-
prédiction d'un résultat numérique	régression linéaire simple ou multiple	-	-	Regression > Linear Regression	Multiple regression	Classic Analysis: REGRESS
prédiction d'un résultat binaire	régression logistique ; courbe ROC	-	-	Regression > Logistic Regression	Logistic regression; ROC curve analysis	Classic Analysis: LOGISTIC
délai de survenue d'un événement	Kaplan-Meier, log-rank, régression de Cox	-	-	module Survival	Kaplan-Meier survival analysis; Cox proportional-hazards regression	Classic Analysis: Kaplan-Meier, Cox

Un étudiant qui maîtrise les cinq premières lignes du tableau peut résoudre presque tous les problèmes de l'examen pratique. Les lignes consacrées à la concordance, à la prédiction et à la survie sont détaillées aux chapitres 9 à 13.

8.4 L'unité d'analyse et l'indépendance des observations

Tous les tests du tableau 8.1 supposent que les observations sont indépendantes (ou, pour les tests appariés, que les paires sont indépendantes entre elles). Cette hypothèse paraît banale, mais elle est souvent violée, surtout lorsqu'un patient fournit plusieurs valeurs.

L'**unité d'analyse** est l'entité sur laquelle ont porté la randomisation ou l'échantillonnage et sur laquelle nous voulons tirer une conclusion : en règle générale, le patient. Les problèmes apparaissent lorsque nous analysons autre chose :

- **yeux, dents, articulations, lésions** : si un patient fournit deux yeux, les valeurs des deux yeux sont corrélées. Dans l'étude de nos collègues sur la biométrie oculaire, 170 yeux provenaient de 102 patients [81]. En odontologie, les faces d'une même dent et les dents d'un même patient sont liées entre elles (effet de grappe) [104] ; traiter 600 dents provenant de 30 patients comme 600 observations indépendantes gonfle artificiellement n et produit des valeurs p trop petites.
- **visites répétées** : un patient hospitalisé 5 fois en un an ne représente pas 5 patients. Si nous analysons les hospitalisations, il faut utiliser des méthodes pour données répétées.
- **grappes** : dans mg2026, les 160 patients proviennent de 12 cabinets de médecine générale. Les patients d'un même médecin peuvent se ressembler (même style de conseil, même zone géographique). Si la randomisation avait été faite par cabinet (tous les patients d'un cabinet dans le même bras), l'unité d'analyse serait devenue le cabinet, et l'analyse aurait dû tenir compte de ce regroupement. Dans notre scénario, la randomisation a été individuelle, les tests habituels sont donc acceptables, mais une analyse de sensibilité pourrait inclure le cabinet comme facteur.

- **événements au lieu de personnes** : 30 épisodes d'hypoglycémie peuvent provenir de 30 patients ou de 3 patients. Un khi-deux sur les épisodes violerait l'indépendance.

Règle pratique : avant tout test, demandez-vous "combien d'unités indépendantes ai-je réellement ?". La réponse est le n réel de l'analyse.

8.5 Le test se fixe avant de voir les données

Un test statistique ne maintient son erreur de type I à 5% que s'il a été choisi indépendamment du résultat. Si nous essayons plusieurs tests sur les mêmes données (t, Welch, Mann-Whitney, avec et sans valeurs extrêmes, avec et sans transformation) et que nous rapportons celui qui "ressort", la probabilité réelle d'un résultat faussement positif dépasse largement 5%, même si chaque test, pris isolément, est correct [70,74]. Ce phénomène est connu dans la littérature sous le nom de "jardin aux sentiers qui bifurquent" : chaque décision apparemment anodine ouvre une nouvelle occasion d'obtenir $p < 0,05$ par hasard.

Le jeu de données mg2026 en offre un exemple concret, présenté dans la figure 8.2. L'une des variables secondaires est la baisse de la pression systolique à 6 mois. Les moyennes sont de 2,79 mmHg (ÉT 8,96) dans le bras A, de 4,53 (ÉT 9,63) dans le bras B et de 6,96 (ÉT 8,18) dans le bras C. Les distributions sont compatibles avec la loi normale dans chaque bras (SW p entre 0,07 et 0,43), et les variances sont homogènes (Levene p = 0,67). Le test adapté, fixé dans le plan d'analyse, est l'ANOVA classique.

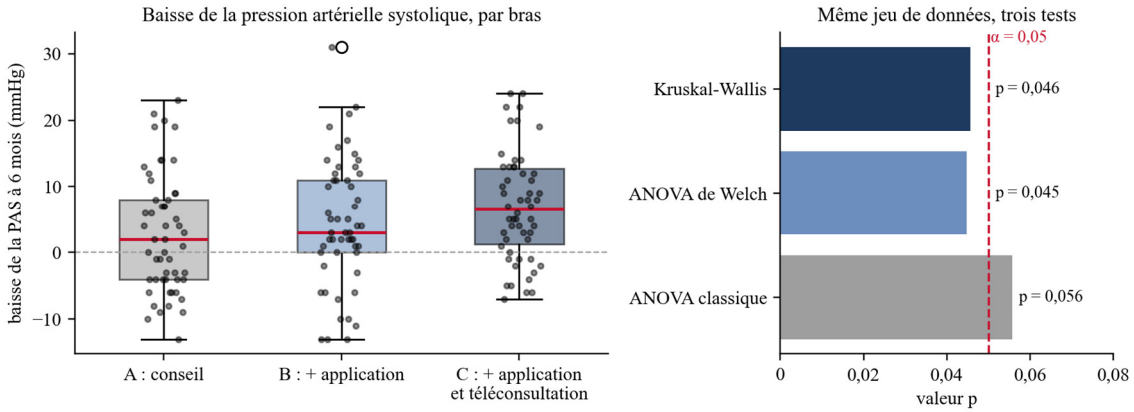


Figure 8.2. À gauche : baisse de la pression artérielle systolique à 6 mois, par bras (boîtes avec la médiane et les quartiles, points individuels). À droite : les valeurs p de trois tests appliqués aux mêmes données : ANOVA classique 0,056, ANOVA de Welch 0,045, Kruskal-Wallis 0,046. La ligne rouge marque le seuil de 0,05. Données synthétiques (jeu de données mg2026).

L'ANOVA classique donne $F(2 ; 157) = 2,94$, $p = 0,056$. La version de Welch donne $p = 0,045$, et le test de Kruskal-Wallis $p = 0,046$. Un chercheur qui choisit le test d'après le résultat rapportera "$p < 0,05$" et écrira que la téléconsultation fait baisser la pression artérielle. Un chercheur qui respecte le plan rapportera $p = 0,056$, $\eta^2 = 0,036$, et dira que les données suggèrent une baisse plus importante dans le bras C (différence C-A de 4,17 mmHg, IC à 95% de Tukey 0,08-8,26), mais que le test global n'a pas atteint le seuil fixé à l'avance. C'est la seconde formulation qui est correcte. La différence entre 0,045 et 0,056 est cliniquement négligeable ; le problème ne tient pas aux chiffres, mais à l'honnêteté de la procédure.

Les solutions sont simples et relèvent des bonnes pratiques : le test principal, le critère de jugement principal et le seuil sont inscrits dans le protocole avant le début de l'étude ; les analyses supplémentaires sont explicitement qualifiées d'"analyses de sensibilité" ou d'"analyses exploratoires" ; et si une condition d'application se révèle non respectée, la règle de repli ("si Levene $p < 0,05$, nous utilisons Welch") est elle aussi écrite à l'avance. Les recommandations

pour le compte rendu des essais cliniques exigent que les modifications par rapport au plan initial soient déclarées [56].

8.6 Erreurs fréquentes dans le choix et l'application du test

8.6.1 Un test pour groupes indépendants appliqué à des données appariées

Nous avons vu à la section 8.2.4 l'effet sur la glycémie de mg2026 : t = 5,03 au lieu de t = 11,14. La même erreur se produit avec les variables binaires : un khi-deux habituel sur un tableau "temps (0 ou 6 mois) x glycémie ≥ 100 (oui/non)" comptant 320 observations, au lieu du test de McNemar sur 160 patients. Le signal d'alarme à la lecture d'un article : les mêmes patients mesurés deux fois et le mot "indépendant" ou le test du khi-deux de Pearson dans la section méthodes.

8.6.2 Plusieurs tests t au lieu d'une ANOVA

Avec trois bras, la tentation est de comparer A à B, A à C et B à C par trois tests t. Chaque comparaison comporte un risque de 5% de résultat faussement positif, et le risque global atteint 14,3%. La procédure correcte : ANOVA (ou Kruskal-Wallis), puis comparaisons post hoc avec correction (Tukey, Bonferroni, Holm, Dunn). L'exception légitime est une comparaison principale fixée à l'avance (par exemple C contre A comme objectif principal), qui peut être testée directement.

8.6.3 Un khi-deux avec de faibles effectifs attendus

Au chapitre 7, sur un tableau de 18 patients avec des effectifs attendus de 3,1 et 3,9, le khi-deux a donné p = 0,040, et le test correct (Fisher) p = 0,066. Les logiciels affichent parfois un avertissement discret ("x cellules ont un effectif attendu inférieur à 5"), que beaucoup ignorent. La vérification des effectifs attendus fait partie de la résolution et s'écrit dans le format a-g, au moment de justifier la méthode.

8.6.4 Choisir le test d'après la valeur p

L'exemple de la section 8.5 est typique. Variantes fréquentes de la même erreur : passer à un test unilatéral après que le test bilatéral a donné p = 0,08 ; éliminer des "valeurs extrêmes" uniquement jusqu'à ce que le résultat devienne significatif ; remplacer le critère de jugement principal par un critère secondaire qui s'est révélé significatif ; tester 20 sous-groupes et rapporter celui où p < 0,05.

8.6.5 La moyenne de scores ordinaux traitée comme une mesure

La satisfaction dans mg2026 a pour moyennes 2,96 (A), 3,62 (B) et 3,98 (C). L'ANOVA sur ces scores donne F = 17,32, p < 0,001, et le test de Kruskal-Wallis H = 32,16, p < 0,001 : la conclusion ne change pas. Le problème n'est pas ici la valeur p, mais l'interprétation : "la satisfaction moyenne a augmenté de 1,02 point" suppose que la différence entre "neutre" et "satisfait" est égale à celle entre "satisfait" et "très satisfait", ce que l'échelle ne garantit pas. Le compte rendu correct utilise les médianes et la distribution par catégorie (figure 7.2 du chapitre 7). Sur de petits échantillons ou avec des distributions présentant un effet plafond, la moyenne de scores ordinaux peut aussi induire des conclusions statistiques erronées.

8.6.6 D'autres erreurs fréquemment rencontrées

Tester la normalité sur l'ensemble de l'échantillon au lieu de chaque groupe (ou des différences, pour les données appariées) est une erreur technique fréquente : deux groupes normaux de moyennes différentes forment ensemble une distribution bimodale, et le test de normalité

"échoue" sans raison. Une autre erreur, propre aux essais randomisés, consiste à tester statistiquement les différences à l'inclusion entre les bras (le tableau 1 avec des valeurs p). Si la randomisation a été correcte, toute différence à l'inclusion est, par définition, due au hasard ; les recommandations CONSORT préconisent de décrire ces différences, et non de les tester [56]. Dans mg2026, le niveau d'activité physique à l'inclusion a un p = 0,08 entre les bras (Kruskal-Wallis), et le statut tabagique un p = 0,07 (khi-deux) : ce sont deux fluctuations de la randomisation, que l'on discute en fonction de leur ampleur clinique, et non de la valeur p. La dernière erreur fréquente est d'ignorer les tests multiples : un article qui rapporte 7 rapports de cotes pour 7 troubles différents, chacun au seuil de 0,05, a une probabilité d'environ 30% d'obtenir au moins un résultat faussement positif même si aucune association n'était réelle.

8.7 Rapporter un résultat selon un format standardisé (a-g)

Dans les devoirs et à l'examen pratique de la discipline, chaque problème se résout selon le même format, en sept composantes. Ce format n'est pas une formalité bureaucratique : il impose le bon ordre de raisonnement (d'abord la question, puis les hypothèses, puis la méthode et seulement à la fin le calcul) et permet à quelqu'un d'autre de vérifier la résolution.

- **(a) L'hypothèse de recherche** : ce que pense le chercheur, en langage clinique. "Les patients du bras C ont une baisse de la glycémie plus importante que ceux du bras A."
- **(b) Les hypothèses statistiques**, en mots et en notation mathématique : H0 et H1, avec le paramètre de la population (μ, Me, p) et en précisant si le test est bilatéral.
- **(c) La méthode choisie et sa justification** : le type des variables, le schéma d'étude (indépendant ou apparié, nombre de groupes), la vérification des conditions (normalité par groupe, Levene, effectifs attendus) et l'alternative qui aurait été utilisée si une condition n'avait pas été remplie.
- **(d) Le mode de résolution et le logiciel** : le logiciel et le menu (par exemple JASP, T-Tests > Independent Samples T-Test, avec les options cochées), ou le calcul à la main, étape par étape.
- **(e) Le résultat, avec un graphique** : les statistiques descriptives par groupe, la statistique de test, les degrés de liberté, la valeur p, la taille d'effet avec son IC à 95% et le graphique adapté (boîte à moustaches ou graphique des moyennes avec IC pour les variables numériques, diagramme en barres pour les proportions).
- **(f) La conclusion statistique** : "p < 0,05, donc nous rejetons H0" ou "p ≥ 0,05, donc nous ne rejetons pas H0", avec la formulation "il existe / il n'existe pas de différence statistiquement significative entre...".
- **(g) La conclusion de recherche**, en mots : ce que signifie le résultat pour la question clinique, avec la taille d'effet et les limites du schéma d'étude.

Le tableau 8.2 donne un exemple complet, portant sur la question principale de l'étude mg2026.

Tableau 8.2. Modèle de résolution au format a-g : baisse de la glycémie à 6 mois selon le bras (jeu de données mg2026).

Composante	Contenu
(a) Hypothèse de recherche	La baisse de la glycémie à jeun à 6 mois diffère selon l'intervention reçue (conseil, conseil + application, conseil + application + téléconsultation).
(b) Hypothèses statistiques	H0 : la baisse moyenne de la glycémie est la même dans les trois bras (μA = μB = μC). H1 : au moins une moyenne diffère des autres.
(c) Méthode et justification	La variable résultat (baisse de la glycémie, mg/dL) est numérique ; la variable de regroupement (le bras) est nominale à 3 catégories ; les groupes sont indépendants. Les distributions sont approximativement normales dans chaque bras, les variances sont homogènes (Levene p =

Composante	Contenu
	0,163), d'où le choix d'une ANOVA à un facteur, suivie du test de Tukey. Alternative en cas de non-respect des conditions : Kruskal-Wallis avec Dunn.
(d) Résolution et logiciel	JASP 0.98.1, ANOVA > ANOVA; Dependent Variable: baisse de la glycémie; Fixed Factors: brat; Assumption Checks: Homogeneity tests; Post Hoc Tests: Tukey; Estimates of effect size: η².
(e) Résultat	A : $3{,}0 \pm 8{,}8 (n\ 53)$; B : $11{,}1 \pm 11{,}2 (n\ 53)$; C : $16{,}3 \pm 10{,}6 (n\ 54)$; F(2 ; 157) = 22,93, p < 0,001, η² = 0,226. Tukey : B-A 8,1 (IC à 95% 3,4-12,8 ; p = 0,0002), C-A 13,3 (8,6-18,0 ; p < 0,001), C-B 5,2 (0,6-9,9 ; p = 0,024). Graphique : boîte à moustaches de la baisse de la glycémie par bras, comme au chapitre 6.
(f) Conclusion statistique	p < 0,001 < 0,05, donc nous rejetons H0 : il existe des différences statistiquement significatives entre les bras ; les trois comparaisons deux à deux sont toutes significatives.
(g) Conclusion de recherche	L'ajout de l'application mobile et, en plus, de la téléconsultation mensuelle a conduit à une baisse plus importante de la glycémie à jeun à 6 mois (en moyenne de 8 et de 13 mg/dL de plus, respectivement, qu'avec le conseil seul). L'effet est important (η² = 0,226). Les données sont simulées ; dans une étude réelle, il faudrait aussi vérifier la pertinence clinique d'une baisse de cette ampleur.

8.8 Problèmes résolus

Les problèmes ci-dessous utilisent le jeu de données mg2026 (le fichier mg2026_curat.csv sur Moodle) et suivent le style des sujets d'examen et des fiches "Teste statistice - recapitulare" (récapitulatif des tests statistiques) et "Probleme BPOC" (problèmes sur la BPCO) de la discipline. Pour chaque problème sont donnés la question, le type des variables, le schéma d'étude, le test choisi avec son alternative, les hypothèses, le résultat et les conclusions. Tous les chiffres ont été calculés sur le jeu de données ; vous pouvez les reproduire dans JASP ou MedCalc. Le seuil de signification est α = 0,05, et les tests sont bilatéraux.

Problème résolu 8.1. La glycémie à jeun chez les patients des milieux urbain et rural

La question. La glycémie à jeun à l'inclusion diffère-t-elle entre les patients du milieu urbain et ceux du milieu rural ?

Variables et schéma d'étude. Résultat : la glycémie à l'inclusion (numérique, mg/dL). Regroupement : le milieu (nominale, 2 catégories). Deux groupes indépendants.

Test choisi et alternative. Shapiro-Wilk par groupe : urbain p = 0,076, rural p = 0,054 ; groupes de grande taille (95 et 65). Levene p = 0,117. Test t non apparié (nous rapportons la version de Welch, qui ne suppose pas l'égalité des variances) ; alternative : U de Mann-Whitney. Dans JASP : T-Tests > Independent Samples T-Test, en cochant Student, Welch, Mann-Whitney et Assumption checks.

Hypothèses. H0 : μurbain = μrural ; H1 : μurbain ≠ μrural.

Résultat. Urbain $105{,}82 \pm 15{,}30$ mg/dL$(n\ 95)$, rural $104{,}88 \pm 16{,}52$ mg/dL$(n\ 65)$. Welch t = -0,37, ddl 130,5, p = 0,715 ; différence rural - urbain -0,94 mg/dL (IC à 95% -6,05 à 4,16) ; d de Cohen -0,06. Mann-Whitney p = 0,81.

Conclusion statistique. p = 0,715 > 0,05, nous ne rejetons pas H0 : il n'existe pas de différence statistiquement significative.

Conclusion clinique. La glycémie à jeun a été pratiquement la même dans les deux milieux. L'intervalle de confiance montre qu'une différence réelle supérieure à environ 6 mg/dL, dans un sens ou dans l'autre, est peu probable, ce qui est plus informatif que le simple "non significatif".

Problème résolu 8.2. Le HDL-cholestérol chez les femmes et chez les hommes

La question. Le HDL-cholestérol diffère-t-il entre les femmes et les hommes ?

Variables et schéma d'étude. Résultat : HDL (numérique, mg/dL). Regroupement : le sexe (2 catégories). Groupes indépendants.

Test choisi et alternative. SW : femmes p = 0,41, hommes p = 0,48, distributions normales. Levene p = 0,034, variances inégales, d'où le test t de Welch. Alternative : Mann-Whitney.

Hypothèses. H0 : μF = μM ; H1 : μF ≠ μM.

Résultat. Femmes $53{,}92 \pm 11{,}70$ mg/dL$(n\,86)$, hommes $44{,}41 \pm 9{,}66$ mg/dL$(n\,74)$. Welch t = 5,63, ddl 157,7, p < 0,001 ; différence F - M 9,51 mg/dL (IC à 95% 6,18-12,85) ; d = 0,88. Mann-Whitney p < 0,001.

Conclusion statistique. p < 0,001, nous rejetons H0 : la différence est statistiquement significative.

Conclusion clinique. Les femmes ont un HDL supérieur en moyenne d'environ 9,5 mg/dL à celui des hommes, un effet important, conforme aux différences physiologiques connues. Conséquence pratique : les seuils de HDL "bas" diffèrent selon le sexe.

Problème résolu 8.3. L'IMC chez les femmes et chez les hommes, une variable asymétrique

La question. L'IMC à l'inclusion diffère-t-il entre les femmes et les hommes ?

Variables et schéma d'étude. Résultat : IMC (numérique, kg/m²), avec une asymétrie positive (1,42 sur l'ensemble de l'échantillon). Regroupement : le sexe. Groupes indépendants.

Test choisi et alternative. SW : femmes p = 0,0001, hommes p < 0,001 ; les distributions sont asymétriques. Nous choisissons le test U de Mann-Whitney, et nous rapportons le test t de Welch comme analyse de sensibilité (les groupes sont grands, il est donc lui aussi robuste).

Hypothèses. H0 : les distributions de l'IMC sont identiques chez les femmes et chez les hommes (MeF = MeM) ; H1 : les distributions diffèrent.

Résultat. Femmes : médiane 28,25 kg/m² (EIQ 26,02-31,25), hommes : 28,25 (EIQ 26,32-31,50). Mann-Whitney U = 3020,5, p = 0,58 ; r_rb = 0,05. Welch p = 0,68.

Conclusion statistique. p = 0,58, nous ne rejetons pas H0.

Conclusion clinique. L'IMC ne diffère pas entre les sexes dans cet échantillon ; les deux tests conduisent à la même conclusion. Lorsque les groupes sont grands et que les tests concordent, le choix entre eux ne change pas l'interprétation ; ce qui compte, c'est que la description (médiane et EIQ) soit cohérente avec le test choisi.

Problème résolu 8.4. Le score FINDRISC chez les patients avec et sans diabète

La question. Les patients ayant un diabète confirmé par HGPO ont-ils un score FINDRISC plus élevé que les patients sans diabète ?

Variables et schéma d'étude. Résultat : FINDRISC (score 0-26, traité comme numérique). Regroupement : le diabète (oui/non). Groupes indépendants, très inégaux (29 et 131).

Test choisi et alternative. SW par groupe : p = 0,17 et 0,33 ; Levene p = 0,67. Test t non apparié ; alternative : Mann-Whitney.

Hypothèses. H0 : μdiabète = μsans diabète ; H1 : μdiabète ≠ μsans diabète.

Résultat. Sans diabète 9,40 ± 3,70 (médiane 10), avec diabète 14,76 ± 3,96 (médiane 14). Student t = -6,97, p < 0,001 (Welch t = -6,67, ddl 39,5, p < 0,001) ; différence de 5,36 points (IC à 95% 3,74-6,99) ; d = 1,43. Mann-Whitney p < 0,001, r_rb = 0,71.

Conclusion statistique. Nous rejetons H0.

Conclusion clinique. Le score FINDRISC est en moyenne supérieur de plus de 5 points chez les patients diabétiques, un effet très important. Si la question avait été "dans quelle mesure le FINDRISC identifie-t-il les patients diabétiques ?", le but aurait été la prédiction, et la méthode adaptée aurait été la courbe ROC (AUC 0,852 dans mg2026), et non le test t. Un même jeu de données, deux questions, deux méthodes différentes.

Problème résolu 8.5. La pression systolique chez les patients avec et sans diabète, quand les tests ne concordent pas

La question. La PAS à l'inclusion diffère-t-elle entre les patients avec et sans diabète ?

Variables et schéma d'étude. Résultat : PAS (numérique, mmHg). Regroupement : le diabète (oui/non). Groupes indépendants (29 et 131).

Test choisi et alternative. SW : p = 0,19 et 0,20 ; Levene p = 0,53. Le test fixé à l'avance est le test t non apparié ; l'alternative, le test de Mann-Whitney.

Hypothèses. H0 : μdiabète = μsans diabète ; H1 : μdiabète ≠ μsans diabète.

Résultat. Sans diabète 129,96 ± 13,86 mmHg (médiane 131), avec diabète 136,41 ± 12,92 mmHg (médiane 132). Student t = -2,30, p = 0,023 (Welch t = -2,40, p = 0,021) ; différence de 6,45 mmHg (IC à 95% 1,04-11,87) ; d = 0,47. Mann-Whitney p = 0,065.

Conclusion statistique. Selon le test choisi à l'avance (t non apparié), p = 0,023 < 0,05, nous rejetons H0.

Conclusion clinique. Les patients diabétiques avaient une PAS supérieure en moyenne d'environ 6,5 mmHg, un effet faible à moyen, estimé de façon imprécise (IC de 1 à 12 mmHg). Pourquoi le test de Mann-Whitney donne-t-il p = 0,065 ? Les médianes sont très proches (131 et 132), et la différence de moyennes provient surtout de la queue supérieure de la distribution chez les diabétiques. Les tests répondent à des questions légèrement différentes. Nous n'avons pas le droit de choisir le test d'après le résultat ; si le protocole prévoyait le test t et que ses conditions sont remplies, nous rapportons le test t et mentionnons la discordance comme analyse de sensibilité.

Problème résolu 8.6. Le poids à l'inclusion et à 6 mois, sur l'ensemble de l'échantillon

La question. Le poids des patients a-t-il changé entre l'inclusion et 6 mois ?

Variables et schéma d'étude. Résultat : le poids (numérique, kg), mesuré deux fois chez les mêmes 160 patients. Données appariées.

Test choisi et alternative. Pour des données appariées, on vérifie la distribution des différences : SW p = 0,86, normale. Test t apparié ; alternative : Wilcoxon. Dans JASP : T-Tests > Paired Samples T-Test, paire greutate0 - greutate6.

Hypothèses. H0 : la moyenne des différences μd = 0 ; H1 : μd ≠ 0.

Résultat. Poids moyen de 83,05 kg à l'inclusion et de 80,74 kg à 6 mois. Différence moyenne 2,31 kg (ÉT 2,66), IC à 95% 1,90-2,73 ; t = 11,00, ddl 159, p < 0,001 ; dz = 0,87. Wilcoxon p < 0,001.

Conclusion statistique. Nous rejetons H0.

Conclusion clinique. Les patients ont perdu en moyenne 2,3 kg en 6 mois. Le test apparié sur l'ensemble de l'échantillon montre qu'un changement s'est produit, mais pas ce qui l'a provoqué : sans comparaison des bras, nous ne pouvons pas séparer l'effet des interventions de l'effet de la participation à l'étude. La question de l'efficacité se résout en comparant la perte de poids entre les bras (ANOVA, F = 18,25, p < 0,001), c'est-à-dire le problème L5 des travaux pratiques 8.

Problème résolu 8.7. La pression systolique à l'inclusion et à 6 mois

La question. La PAS a-t-elle diminué entre l'inclusion et 6 mois ?

Variables et schéma d'étude. PAS, numérique, mêmes patients à deux temps : données appariées.

Test choisi et alternative. Les différences ont un SW p = 0,13. Test t apparié ; alternative : Wilcoxon.

Hypothèses. H0 : $\mu d = 0$; H1 : $\mu d \neq 0$.

Résultat. PAS moyenne de 131,13 mmHg à l'inclusion et de 126,36 mmHg à 6 mois ; différence moyenne 4,78 mmHg (ÉT 9,05), IC à 95% 3,36-6,19 ; t = 6,68, ddl 159, p < 0,001 ; dz = 0,53. Wilcoxon p < 0,001.

Conclusion statistique. Nous rejetons H0.

Conclusion clinique. La PAS a diminué en moyenne d'environ 4,8 mmHg. À l'échelle de la population, une baisse de cette ampleur est considérée comme pertinente pour le risque cardiovasculaire, mais ici aussi l'absence d'un groupe de comparaison non traité limite l'interprétation causale (la régression vers la moyenne et l'effet de la mesure répétée peuvent y contribuer).

Problème résolu 8.8. La baisse de la glycémie dans les trois bras

La question. La baisse de la glycémie à 6 mois diffère-t-elle entre les bras A, B et C ?

Variables et schéma d'étude. Résultat numérique, regroupement nominal à 3 catégories, groupes indépendants.

Test choisi et alternative. ANOVA à un facteur avec Tukey (Levene p = 0,163) ; alternative : Kruskal-Wallis avec Dunn.

Hypothèses. H0 : $\mu A = \mu B = \mu C$; H1 : au moins une moyenne diffère.

Résultat. La résolution complète figure dans le tableau 8.2 : F(2 ; 157) = 22,93, p < 0,001, η^2 = 0,226 ; Tukey significatif pour les trois paires. Kruskal-Wallis H = 36,63, p < 0,001.

Conclusion statistique. Nous rejetons H0.

Conclusion clinique. Les interventions avec application ont produit des baisses plus importantes de la glycémie, et l'effet augmente avec l'intensité de l'intervention.

Problème résolu 8.9. La pression systolique selon le statut tabagique

La question. La PAS à l'inclusion diffère-t-elle entre les non-fumeurs, les anciens fumeurs et les fumeurs ?

Variables et schéma d'étude. Résultat : PAS (numérique). Regroupement : le tabagisme (nominale, 3 catégories). Groupes indépendants.

Test choisi et alternative. SW par groupe : 0,37, 0,86, 0,23 ; Levene p = 0,114. ANOVA avec Tukey ; alternative : Kruskal-Wallis.

Hypothèses. H0 : μnon-fumeurs = μanciens fumeurs = μfumeurs ; H1 : au moins une moyenne diffère.

Résultat. Anciens fumeurs $127{,}55 \pm 14{,}39 (n\,33)$, fumeurs $137{,}04 \pm 14{,}63 (n\,53)$, non-fumeurs $128{,}50 \pm 11{,}76 (n\,74)$. F(2 ; 157) = 7,85, p < 0,001 (p = 0,0006), η² = 0,091. Tukey : fumeurs - anciens fumeurs 9,49 mmHg (IC à 95% 2,50-16,48 ; p = 0,005) ; fumeurs - non-fumeurs 8,54 mmHg (IC à 95% 2,87-14,21 ; p = 0,001) ; non-fumeurs - anciens fumeurs 0,95 mmHg (p = 0,94). Kruskal-Wallis H = 12,77, p = 0,002.

Conclusion statistique. Nous rejetons H0 ; les fumeurs diffèrent des deux autres groupes.

Conclusion clinique. Les fumeurs actuels ont une PAS supérieure d'environ 9 mmHg à celle des non-fumeurs et des anciens fumeurs, et les anciens fumeurs ne diffèrent pas des non-fumeurs. Le résultat concorde avec la régression multiple du chapitre 9, dans laquelle le tabagisme reste associé à la PAS (+6,74 mmHg par rapport aux anciens fumeurs) après ajustement sur l'IMC, l'âge et le sexe. S'agissant d'une analyse transversale, elle n'établit pas à elle seule la causalité.

Problème résolu 8.10. La baisse de la pression systolique selon le bras, un résultat à la limite

La question. La baisse de la PAS à 6 mois diffère-t-elle entre les trois bras ?

Variables et schéma d'étude. Résultat numérique, trois groupes indépendants.

Test choisi et alternative. SW par bras 0,07-0,43 ; Levene p = 0,67. Le plan d'analyse prévoit une ANOVA classique ; alternative en cas de variances inégales : Welch ; en cas d'asymétrie : Kruskal-Wallis.

Hypothèses. H0 : μA = μB = μC ; H1 : au moins une moyenne diffère.

Résultat. A 2,79 ± 8,96 ; B 4,53 ± 9,63 ; C $6{,}96 \pm 8{,}18$ mmHg. F(2 ; 157) = 2,94, p = 0,056, η² = 0,036 (figure 8.2).

Conclusion statistique. p = 0,056 > 0,05, nous ne rejetons pas H0.

Conclusion clinique. Les données suggèrent une baisse plus importante de la PAS dans le bras C, mais le test global fixé à l'avance n'atteint pas le seuil de signification. Le fait que les tests de Welch (p = 0,045) et de Kruskal-Wallis (p = 0,046) passent sous 0,05 ne change pas la conclusion, car ce n'étaient pas les tests planifiés et les conditions de l'ANOVA classique sont remplies. Le résultat est un signal qui mérite d'être étudié comme objectif principal dans une étude future. Après un test global non significatif, les comparaisons post hoc ne s'interprètent pas comme une preuve.

Problème résolu 8.11. La satisfaction selon le niveau d'études

La question. La satisfaction à 6 mois diffère-t-elle selon le niveau d'études ?

Variables et schéma d'étude. Résultat : la satisfaction (ordinale 1-5). Regroupement : le niveau d'études (ordinal, 4 niveaux). Groupes indépendants.

Test choisi et alternative. Résultat ordinal, 4 groupes : Kruskal-Wallis. Comme le niveau d'études est lui aussi ordonné, une question complémentaire, "la satisfaction augmente-t-elle avec le niveau d'études ?", se teste plus efficacement par la corrélation de Spearman.

Hypothèses. H0 : la distribution de la satisfaction est la même dans les 4 niveaux d'études ; H1 : au moins un niveau diffère.

Résultat. Médianes : collège 3 (n 20), lycée ou enseignement professionnel 4 (n 57), universitaire 3 (n 64), postuniversitaire 4 (n 19). Kruskal-Wallis H = 4,12, ddl 3, p = 0,25. Spearman rho = 0,125, p = 0,115.

Conclusion statistique. Nous ne rejetons pas H0 (ni pour la différence ni pour la tendance).

Conclusion clinique. La satisfaction ne diffère pas significativement selon le niveau d'études. L'absence de différence est un résultat utile pour la mise en œuvre : le programme semble aussi bien accepté par des patients de niveaux d'études différents, même si les groupes extrêmes (20 et 19 patients) sont petits.

Problème résolu 8.12. L'activité physique à l'inclusion selon le bras : le test a-t-il un sens ?

La question. L'activité physique autoévaluée (1-5) à l'inclusion diffère-t-elle entre les bras ?

Variables et schéma d'étude. Résultat ordinal, trois groupes indépendants, mesuré avant l'intervention, dans un essai randomisé.

Test choisi et alternative. Techniquement, Kruskal-Wallis (alternative : khi-deux sur le tableau 3 x 5, mais une cellule a des effectifs très faibles, car seuls 3 patients ont le niveau 5).

Hypothèses. H0 : la distribution de l'activité est la même dans les trois bras ; H1 : au moins un bras diffère.

Résultat. Médiane de 3 dans tous les bras (moyennes des codes 2,98, 2,58 et 2,61) ; Kruskal-Wallis H = 5,01, ddl 2, p = 0,082.

Conclusion statistique. Nous ne rejetons pas H0.

Conclusion clinique. Plus important que le résultat est le fait que la question elle-même est discutable. Dans un essai correctement randomisé, les différences à l'inclusion sont dues au hasard par définition, et la valeur p ne dit rien de la qualité de la randomisation. Ce qui compte, c'est de savoir si une différence, si petite soit-elle, est suffisamment importante sur le plan clinique pour influencer le résultat ; si oui, on l'ajuste dans l'analyse (par exemple par régression), indépendamment de la valeur p [56].

Problème résolu 8.13. La glycémie à 0, 3 et 6 mois dans le bras C

La question. La glycémie à jeun s'est-elle modifiée au cours du temps chez les patients du bras C ?

Variables et schéma d'étude. Résultat numérique, mêmes 54 patients à 3 temps : mesures répétées.

Test choisi et alternative. Dans le bras C, la glycémie à l'inclusion a une distribution asymétrique (asymétrie 1,05 ; SW p = 0,003), les différences 0-3 mois ne sont pas normales (SW p = 0,027), et la condition de sphéricité de l'ANOVA à mesures répétées n'est pas respectée (Mauchly p = 0,002). Nous choisissons le test de Friedman avec post hoc de Wilcoxon et correction de Bonferroni ; l'alternative est l'ANOVA à mesures répétées avec correction de Greenhouse-Geisser.

Hypothèses. H0 : la distribution de la glycémie est la même à 0, 3 et 6 mois ; H1 : au moins un temps diffère.

Résultat. Médianes : 100,5 / 89,0 / 85,0 mg/dL. Friedman χ^2 = 64,63, ddl 2, p < 0,001. Post hoc (Bonferroni) : 0 contre 3 mois p < 0,001 ; 3 contre 6 mois p = 0,011 ; 0 contre 6 mois p < 0,001.

Conclusion statistique. Nous rejetons H0 ; les trois comparaisons restent toutes significatives après correction.

Conclusion clinique. Dans le bras C, la glycémie a diminué au cours des 3 premiers mois et a continué de baisser jusqu'à 6 mois, avec une baisse totale de la médiane de 15,5 mg/dL. La comparaison avec les bras A et B montre que la baisse ne s'explique pas seulement par le passage du temps (voir chapitre 7, section 7.6.3).

Problème résolu 8.14. Le statut tabagique selon le sexe

La question. Le statut tabagique est-il associé au sexe ?

Variables et schéma d'étude. Deux variables nominales (sexe : 2 catégories ; tabagisme : 3 catégories), sujets indépendants : tableau 2 x 3.

Test choisi et alternative. Khi-deux de Pearson ; le plus petit effectif attendu est de 15,26, la condition est remplie. Alternative en cas de faibles E : Fisher ou regroupement de catégories.

Hypothèses. H0 : le tabagisme et le sexe sont indépendants ; H1 : il existe une association.

Résultat. Femmes : anciennes fumeuses 16 (18,6%), fumeuses 30 (34,9%), non-fumeuses 40 (46,5%). Hommes : 17 (23,0%), 23 (31,1%), 34 (45,9%). $\chi^2 = 0,54$, ddl 2, $p = 0,76$; V de Cramér = 0,06.

Conclusion statistique. Nous ne rejetons pas H0.

Conclusion clinique. La distribution du statut tabagique est semblable chez les femmes et chez les hommes dans cet échantillon. Dans la population générale roumaine, la prévalence du tabagisme diffère habituellement entre les sexes ; un échantillon de patients à risque cardiométabolique recrutés en cabinet ne reflète pas nécessairement la population générale.

Problème résolu 8.15. Le diabète selon le sexe

La question. La prévalence du diabète confirmé par HGPO diffère-t-elle entre les femmes et les hommes ?

Variables et schéma d'étude. Deux variables dichotomiques, sujets indépendants : tableau 2 x 2.

Test choisi et alternative. Effectifs attendus : le plus petit est de 13,4, d'où le khi-deux de Pearson ; test de Fisher à titre de vérification.

Hypothèses. H0 : $pF = pM$; H1 : $pF \neq pM$.

Résultat. Femmes 15/86 (17,4%), hommes 14/74 (18,9%). $\chi^2 = 0,06$, ddl 1, $p = 0,81$ (Yates $p = 0,97$; Fisher $p = 0,84$).

Conclusion statistique. Nous ne rejetons pas H0.

Conclusion clinique. La prévalence du diabète non diagnostiqué auparavant a été pratiquement la même chez les femmes et chez les hommes. Dans le modèle de régression logistique du diabète (obésité, âge et sexe comme prédicteurs), le sexe reste sans effet après ajustement (OR 0,97, IC à 95% 0,40-2,38).

Problème résolu 8.16. Un âge d'au moins 50 ans et le diabète : quelle version rapporter ?

La question. Un âge ≥ 50 ans est-il associé au diabète confirmé par HGPO ?

Variables et schéma d'étude. L'âge dichotomisé à 50 ans (seuil fixé à l'avance, ayant un sens clinique pour le dépistage) et le diabète (oui/non) ; tableau 2 x 2, sujets indépendants.

Test choisi et alternative. Le plus petit effectif attendu est de 6,2 (34 x 29/160) : la condition $E \geq 5$ est remplie, mais de justesse. Nous rapportons le khi-deux et, à titre de vérification, le test de Fisher.

Hypothèses. H0 : $p \geq 50 = p < 50$; H1 : $p \geq 50 \neq p < 50$.

Résultat. Moins de 50 ans : 2/34 (5,9%) ; 50 ans ou plus : 27/126 (21,4%). $\chi^2 = 4,36$, ddl 1, $p = 0,037$; Yates $\chi^2 = 3,38$, $p = 0,066$; Fisher $p = 0,044$; $\varphi = 0,17$.

Conclusion statistique. Avec le test de Pearson et avec le test de Fisher, $p < 0,05$: nous rejetons H0. La correction de Yates, plus conservatrice, donne $p = 0,066$.

Conclusion clinique. Le diabète a été plus fréquent après 50 ans, mais le résultat repose sur seulement 2 cas dans le groupe jeune et il est fragile. L'âge traité comme variable continue dans la régression logistique (OR 1,11 par an, IC à 95% 1,05-1,18) donne une image plus stable que la dichotomisation. La leçon : dichotomiser une variable numérique fait perdre de l'information.

Problème résolu 8.17. Perte de poids ≥ 5% dans le bras C par rapport au bras A

La question. La téléconsultation, ajoutée à l'application, augmente-t-elle la proportion de patients qui perdent au moins 5% de leur poids, par rapport au conseil seul ?

Variables et schéma d'étude. Résultat dichotomique (perte ≥ 5% oui/non), deux bras randomisés indépendants (comparaison principale fixée dans le protocole).

Test choisi et alternative. Khi-deux 2 x 2 (le plus petit effectif attendu dépasse 10) et mesures de risque : RR, différence de risques, NST.

Hypothèses. H0 : pC = pA ; H1 : pC ≠ pA.

Résultat. C : 16/54 (29,6%) ; A : 6/53 (11,3%). $\chi^2 = 5,49$, ddl 1, $p = 0,019$ (Yates 4,43, $p = 0,035$). RR = 2,62 (IC à 95% 1,11-6,17) ; différence absolue de 18,3 points de pourcentage (IC à 95% 3,4-33,2) ; NST 5,5, arrondi à 6 (IC à 95% 3,0-29,1).

Conclusion statistique. Nous rejetons H0.

Conclusion clinique. Pour 6 patients pris en charge avec l'application et la téléconsultation au lieu du conseil seul, un patient supplémentaire atteint une perte de poids d'au moins 5%. L'intervalle de confiance du NST est large (de 3 à 29), l'ampleur du bénéfice est donc estimée de façon imprécise.

Problème résolu 8.18. L'obésité à l'inclusion et à 6 mois : test de McNemar

La question. La proportion de patients obèses (IMC ≥ 30) a-t-elle changé entre l'inclusion et 6 mois ?

Variables et schéma d'étude. Variable dichotomique mesurée deux fois chez les mêmes patients : données appariées. L'IMC à 6 mois a été calculé à partir du poids à 6 mois et de la taille.

Test choisi et alternative. Le test de McNemar (avec correction de continuité ou dans sa version exacte, car le nombre de paires discordantes est faible).

Hypothèses. H0 : la proportion d'obèses est la même aux deux temps ; H1 : la proportion diffère.

Résultat. Obèses aux deux temps 47 ; obèses seulement à l'inclusion 7 ; obèses seulement à 6 mois 4 ; non obèses aux deux temps 102. Proportion : 54/160 (33,8%) à l'inclusion, 51/160 (31,9%) à 6 mois. χ^2McNemar $= (7 - 4)^2/11 = 0,82$, $p = 0,37$ (avec correction 0,36, $p = 0,55$; exact $p = 0,55$).

Conclusion statistique. Nous ne rejetons pas H0.

Conclusion clinique. Bien que les patients aient perdu en moyenne 2,3 kg (problème 8.6), la classification comme obèse n'a changé que chez très peu de patients : une perte de poids modeste ne fait pas facilement passer un patient de l'autre côté du seuil de 30 kg/m². La comparaison avec le problème 8.6 montre une fois de plus combien d'information se perd par dichotomisation.

Problème résolu 8.19. IMC et pression systolique : association

La question. L'IMC est-il associé à la PAS à l'inclusion ?

Variables et schéma d'étude. Deux variables numériques mesurées chez les mêmes patients ; le but est l'association.

Test choisi et alternative. Corrélation de Pearson, avec examen du nuage de points ; alternative, en raison de l'asymétrie de l'IMC : Spearman.

Hypothèses. H0 : $\rho = 0$ (il n'existe pas de corrélation linéaire dans la population) ; H1 : $\rho \neq 0$.

Résultat. r = 0,405, p < 0,001 ; Spearman rho = 0,372, p < 0,001 ; $R^2 = 0{,}164$.

Conclusion statistique. Nous rejetons H0.

Conclusion clinique. La PAS augmente avec l'IMC, avec une corrélation modérée. L'IMC explique environ 16% de la variabilité de la PAS ; les 84% restants tiennent à d'autres facteurs (âge, tabagisme, génétique, sel, traitements). La corrélation ne montre pas à elle seule la causalité. Détails au chapitre 9.

Problème résolu 8.20. Prédire la pression systolique à partir de l'IMC

La question. Quelle PAS attendons-nous chez un patient dont l'IMC est de 30 kg/m² ?

Variables et schéma d'étude. Le but est la prédiction d'une variable numérique à partir d'une autre variable numérique.

Test choisi et alternative. Régression linéaire simple, avec vérification des résidus ; pour plusieurs prédicteurs, régression multiple.

Hypothèses. H0 : la pente $\beta = 0$; H1 : $\beta \neq 0$.

Résultat. PAS = 89,47 + 1,435 x IMC ; IC à 95% de la pente 0,925-1,944 ; p < 0,001 ; erreur standard de l'estimation 12,73 mmHg. Prédiction pour un IMC de 30 : 132,5 mmHg ; pour un IMC de 35 : 139,7 mmHg.

Conclusion statistique. Nous rejetons H0 : la pente est significativement différente de 0.

Conclusion clinique. Chaque unité d'IMC supplémentaire est associée, en moyenne, à une PAS plus élevée d'environ 1,4 mmHg. Pour un patient donné, la prédiction est toutefois imprécise : avec une erreur standard de l'estimation de 12,7 mmHg, la PAS réelle d'un patient dont l'IMC est de 30 peut facilement se situer entre 108 et 157 mmHg. Le modèle décrit la tendance moyenne, il ne remplace pas la mesure de la pression artérielle.

Problème résolu 8.21. La satisfaction dans le bras C par rapport au bras B

La question. La téléconsultation, ajoutée à l'application, apporte-t-elle une satisfaction plus grande ?

Variables et schéma d'étude. Résultat ordinal (1-5), deux groupes indépendants.

Test choisi et alternative. U de Mann-Whitney. Un test t sur un score à 5 valeurs n'a pas de sens.

Hypothèses. H0 : la distribution de la satisfaction est la même dans les bras C et B ; H1 : les distributions diffèrent.

Résultat. Médiane de 4 (EIQ 3-5) dans le bras C et de 4 (EIQ 3-4) dans le bras B ; U = 1698, p = 0,079 ; r_rb = 0,19.

Conclusion statistique. Nous ne rejetons pas H0.

Conclusion clinique. La satisfaction tend à être un peu plus élevée dans le bras C (31% de patients très satisfaits, contre 17% dans le bras B), mais la différence n'est pas statistiquement significative et l'effet estimé est faible. Pour décider de proposer la téléconsultation, le critère principal devrait être le résultat clinique et le coût, et non la satisfaction.

Problème résolu 8.22. Le niveau d'études selon le milieu de résidence

La question. Le niveau d'études diffère-t-il entre les patients du milieu urbain et ceux du milieu rural ?

Variables et schéma d'étude. Le niveau d'études est ordinal (4 niveaux), le milieu est nominal (2 catégories) ; sujets indépendants.

Test choisi et alternative. Deux approches sont acceptables : le khi-deux sur le tableau 2 x 4 (qui ignore l'ordre des niveaux) ou le test de Mann-Whitney sur le niveau d'études (qui utilise l'ordre et est plus puissant si la différence est un décalage d'ensemble).

Hypothèses. H0 : la distribution du niveau d'études est la même en milieu urbain et en milieu rural ; H1 : les distributions diffèrent.

Résultat. Rural : 7 (10,8%), 22 (33,8%), 30 (46,2%), 6 (9,2%) ; urbain : 13 (13,7%), 35 (36,8%), 34 (35,8%), 13 (13,7%). Le plus petit effectif attendu est de 7,7. $\chi^2 = 2,04$, ddl 3, p = 0,56 ; V de Cramér = 0,11. Mann-Whitney p = 0,69.

Conclusion statistique. Nous ne rejetons H0 avec aucune des deux approches.

Conclusion clinique. Le niveau d'études est semblable dans les deux milieux dans cet échantillon. C'est un résultat qui surprendrait dans une enquête en population ; ici, il reflète probablement la manière dont les patients ont été recrutés dans les cabinets (échantillon de commodité), et non la réalité démographique.

8.9 Reconnaître le test erroné : dix descriptions de type "Méthodes"

Les extraits ci-dessous imitent des formulations rencontrées dans des articles et dans des mémoires de fin d'études. Ils sont fictifs et ont été construits pour l'exercice. Pour chacun, lisez d'abord la description et essayez de trouver vous-mêmes le problème, puis comparez avec l'explication.

Description 1. "L'HbA1c a été mesurée à l'inclusion et après 6 mois de traitement chez 45 patients. Les valeurs ont été comparées par le test t de Student pour échantillons indépendants (p = 0,09)." Le problème : les mêmes patients, deux temps, donc des données appariées. Le test pour groupes indépendants ignore la corrélation entre les deux mesures et perd de la puissance. Correct : le test t apparié (si les différences sont approximativement normales) ou le test de Wilcoxon pour données appariées.

Description 2. "La durée d'hospitalisation (médiane 6 jours, EIQ 4-15) a été comparée entre les deux services (n = 18 et n = 22) par le test t de Student." Le problème : les auteurs décrivent eux-mêmes la variable par la médiane et un EIQ très asymétrique (de 4 à 15 autour d'une médiane de 6), et les groupes sont petits. La description et le test ne concordent pas. Correct : le test U de Mann-Whitney, en rapportant les médianes par groupe.

Description 3. "La pression systolique a été comparée entre les quatre classes d'IMC (poids normal, surpoids, obésité de grade I, obésité de grade II-III) par des tests t deux à deux ; une valeur p < 0,05 a été considérée comme significative." Le problème : quatre groupes impliquent

six comparaisons, et le risque global d'erreur de type I dépasse 26%. Correct : une ANOVA à un facteur (ou un test de Kruskal-Wallis), suivie de comparaisons post hoc avec correction ; si l'on s'intéresse à l'augmentation monotone de la PAS avec la classe d'IMC, un test de tendance ou une régression avec l'IMC en continu.

Description 4. "Les complications postopératoires sont survenues chez 2 des 15 patients opérés par cœlioscopie et chez 7 des 14 patients opérés par voie classique (khi-deux de Pearson, p = 0,033)." Le problème : les effectifs attendus dans la colonne "complications" sont de 4,66 et 4,34, tous deux inférieurs à 5. Correct : le test exact de Fisher, qui donne ici p = 0,050, c'est-à-dire exactement à la limite et, au sens strict, non significatif. La conclusion "la chirurgie cœlioscopique réduit significativement les complications" n'est pas étayée par les données.

Description 5. "Le cholestérol total a été comparé entre les non-fumeurs (n = 40), les anciens fumeurs (n = 25) et les fumeurs (n = 35) par le test de Friedman." Le problème : le test de Friedman concerne les mesures répétées chez les mêmes sujets, et non des groupes indépendants. Correct : une ANOVA à un facteur ou, si les distributions sont asymétriques, un test de Kruskal-Wallis.

Description 6. "La prévalence de l'hypertension non contrôlée à 0, 6 et 12 mois chez les mêmes 80 patients a été comparée par un test du khi-deux sur le tableau 3 x 2 (temps x contrôle tensionnel)." Le problème : le tableau contient 240 "observations" qui sont en réalité 80 patients comptés trois fois. Correct : le test Q de Cochran pour une variable binaire mesurée de façon répétée (ou le test de McNemar pour des paires de temps, avec correction).

Description 7. "La concordance entre le tensiomètre automatique et le tensiomètre manuel a été excellente (corrélation de Pearson r = 0,93, p < 0,001)." Le problème : la corrélation mesure l'association, et non la concordance. Un appareil qui donne systématiquement 10 mmHg de plus peut avoir r = 0,99. Correct : la méthode de Bland-Altman (moyenne des différences et limites de concordance à 95%) et, éventuellement, le coefficient de corrélation intraclasse [102].

Description 8. "Nous avons analysé 412 dents de 52 enfants, en comparant la fréquence des caries sur les dents scellées et non scellées par le test du khi-deux." Le problème : l'unité d'analyse. Les dents d'un même enfant ne sont pas indépendantes (même hygiène, même alimentation, même salive), le n réel est donc bien inférieur à 412, et la valeur p est trop optimiste [104]. Correct : des méthodes qui tiennent compte du regroupement (par exemple des modèles avec effet patient) ou une analyse au niveau du patient.

Description 9. "Le test de Shapiro-Wilk appliqué à l'ensemble de l'échantillon (n = 300, les deux groupes réunis) a donné p = 0,03, c'est pourquoi nous avons utilisé le test de Mann-Whitney." Le problème : la normalité se vérifie dans chaque groupe, et non sur les groupes réunis (si les moyennes diffèrent, leur mélange n'est pas normal même lorsque chaque groupe l'est). De plus, avec 150 sujets par groupe, le test t est robuste face à des écarts modérés. La décision doit reposer sur la forme de la distribution dans chaque groupe et être fixée à l'avance.

Description 10. "Dans une cohorte prospective suivie jusqu'à 5 ans, avec des durées de suivi différentes d'un patient à l'autre, la mortalité a été comparée entre les patients avec et sans diabète par un khi-deux, et l'association a été exprimée sous forme de rapport de cotes." Le problème : lorsque la durée de suivi varie, la simple proportion de décès ignore la durée pendant laquelle chaque patient a été suivi et ce qu'il est advenu des perdus de vue. Correct : l'analyse de survie (courbes de Kaplan-Meier, test du log-rank, régression de Cox avec rapport des risques instantanés, ou hazard ratio, et IC à 95%).

8.10 Lecture critique de la section Méthodes : articles de collègues de la discipline

8.10.1 Comment lire la section Méthodes

Lorsque nous lisons un article, nous appliquons l'algorithme de la figure 8.1 "à l'envers" : nous partons du schéma d'étude et des variables déclarés et vérifions si le test rapporté est celui que nous aurions choisi nous-mêmes. Six questions suffisent pour une première lecture :

1. Quel est le schéma d'étude (transversal, cohorte, cas-témoins, essai clinique, avant-après, comparaison de méthodes) ?
2. Quelle est la variable résultat principale et de quel type est-elle ?
3. Les données sont-elles indépendantes ou appariées ? Quelle est l'unité d'analyse ?
4. Le test nommé (ou suggéré par la manière de rapporter les résultats) correspond-il au type de la variable et au schéma d'étude ?
5. La description (moyenne ± ÉT ou médiane et EIQ) correspond-elle au test ?
6. Une taille d'effet est-elle rapportée avec son intervalle de confiance, et la mesure d'effet est-elle compatible avec le schéma d'étude (OR, RR, HR) ?

Les résumés disposent d'un espace limité, et de nombreux détails n'apparaissent que dans le texte intégral. L'exercice ci-dessous travaille exclusivement avec l'information du résumé, telle qu'elle est enregistrée dans le catalogue bibliographique de cet ouvrage ; les observations critiques sont des points de discussion, et non des verdicts sur les articles.

8.10.2 Le tableau de lecture

Tableau 8.3. Adéquation entre le schéma d'étude, la variable de résultat et la méthode statistique dans des articles de collègues de la discipline (d'après les résumés).

Article	Schéma d'étude et n	Variable de résultat	Test ou mesure rapportés	Adéquat ?	Commentaire
Padurean et al. 2026 [82]	transversale, 3 groupes : 73, 86, 86	score d'estime de soi (numérique)	ANOVA avec Tukey ; d de Cohen ; η^2 ; Pearson ; régression multiple	oui	$F = 299{,}00$, $p < 0{,}001$, $\eta^2 = 0{,}712$; $d = 2{,}515$ pour les porteurs d'appareils auditifs par rapport aux non-porteurs ; présentation exemplaire des résultats, avec tailles d'effet
Munteanu et al. 2020 [94]	étude pilote contrôlée, mesures répétées sur un an, n = 93	scores d'observance et de contrôle (ordinaux) ; exacerbations (binaire)	Mann-Whitney ; Friedman avec Bonferroni ; khi-deux 2 x 2	oui	$\chi^2(1) = 13{,}707$, $p < 0{,}001$ pour 10,30% contre 46,30% ; un RR avec IC à 95% et une taille d'effet pour le Mann-Whitney seraient utiles
Gherbon et al. 2024 [59]	rétrospective avant-après, 1 an, n = 200	glycémie, HbA1c, poids, PAS (numériques, asymétriques)	médianes (EIQ) avant et après ; test non nommé	probablement oui	schéma apparié : le test de Wilcoxon est le test indiqué ; les médianes sont

Article	Schéma d'étude et n	Variable de résultat	Test ou mesure rapportés	Adéquat ?	Commentaire
					appelées "moyennes" dans le résumé ; la médiane des différences individuelles manque
Buda et al. 2019 [86]	avant-après un an de périndopril, avec groupe témoin ; n non déclaré dans le résumé	thrombospondine-1 sérique (numérique)	comparaison T0 contre T1, $p < 0,001$	partiellement	8061,9 ± 3684,80 contre 15380 ± 5887 ng/mL : un ÉT élevé par rapport à la moyenne suggère une asymétrie (Wilcoxon ou transformation logarithmique) ; la comparaison avec les témoins (9221,03 ± 6510,21) demande un test pour groupes indépendants ; n doit être déclaré
Gherbon et al. 2021 [85]	étude contrôlée contre placebo, 20 + 20, 3 mois	IMC, lipides, glycémie, PAD (numériques)	comparaisons entre groupes à 3 mois et intra-groupe avant-après	partiellement	le résumé donne des valeurs p issues des deux types de comparaisons ; seule la comparaison entre groupes (idéalement sur les variations ou par ANCOVA) répond à la question de l'effet ; avec 20 patients par groupe, la vérification de la distribution compte
Danila et al. 2024 [105]	déclarée transversale, appariement 168 + 168	sévérité, évolution (binaire) ; indices inflammatoires	ROC (AUC 0,624 et 0,752) ; hazard ratio 2,62 (IC 1,29-5,04)	pas entièrement	le HR suppose un suivi dans le temps (schéma de cohorte) ; les pourcentages à trois décimales (70,312%) suggèrent une fausse précision
Ion et al. 2024 [106]	cohorte prospective, n =	COVID-19 sévère (binaire)	ROC à deux moments ;	partiellement	la régression logistique

Article	Schéma d'étude et n	Variable de résultat	Test ou mesure rapportés	Adéquat ?	Commentaire
	141		régression logistique ; HR		produit des OR, pas des HR (SOFA HR 3,07, IC 2,29-4,12) ; il faut préciser si une régression de Cox avec temps jusqu'à l'événement a été utilisée
Cioca et al. 2024 [8]	cohorte rétrospective, 3 groupes : 200, 62, 224	sévérité, mortalité (binaire) ; durée d'hospitalisation (numérique)	proportions, moyenne ± ÉT, OR	partiellement	OR 2,24 (réanimation) et 2,46 (ventilation) sans IC et sans p ; la durée d'hospitalisation, 11,6 ± 7,0 contre 9,1 ± 5,8 jours, est asymétrique (médiane, Mann-Whitney) ; 3 groupes demandent d'abord un test global
Stoichescu-Hogea et al. 2021 [90]	3 groupes : 62, 46, 40	strain longitudinal global (numérique)	comparaison de 3 moyennes avec comparaisons deux à deux ; Spearman ; ROC	oui, avec une réserve	ÉT différents (1,4% chez les témoins contre 3,0% chez les hypertendus) : une ANOVA de Welch ou un post hoc de Games-Howell aurait mérité d'être envisagé
Popa et al. 2025 [100]	cohorte rétrospective, 26 + 106	symptômes, traitements (binaire)	Fisher ; OR	oui	petit groupe et événements rares, le test de Fisher est donc correct ; il manque les IC des OR (par exemple OR = 3,97 pour les céphalées)
Piroș et al. 2026 [6]	transversale, n = 693	prédiabète (binaire)	khi-deux sur 3 clusters (p = 0,003) ; tendance ; régression logistique	oui	catégories ordonnées, tendance rapportée (p de tendance < 0,001) ; les clusters sont construits à partir des mêmes données
Moleriu et al.	transversale, n =	excès pondéral	khi-deux, V de	oui	rapporte la force

Article	Schéma d'étude et n	Variable de résultat	Test ou mesure rapportés	Adéquat ?	Commentaire
2026 [4]	921	(binaire)	Cramér ; régression logistique (ORa)		de l'association (V) à côté de p ; schéma transversal, donc association et non causalité
Miulescu et al. 2026 [91]	étude pilote à un seul bras, n = 21, 3 moments	scores POEM et SCORAD (numériques)	ANOVA à mesures répétées	oui, avec des limites	POEM de 14,67 ± 5,22 à 7,10 ± 3,82 ; avec 21 enfants, un test de Friedman serait une analyse de sensibilité utile ; sans groupe témoin, l'effet ne peut pas être attribué au traitement
Popa et al. 2025 [61]	diagnostique, 69 avec AVC et 16 sans	GFAP, UCH-L1 (numériques)	comparaison des moyennes ; seuil diagnostique	partiellement	GFAP 142,91 ± 102,19 contre $37,76 \pm 19,92$ pg/mL ; un ÉT proche de la moyenne indique une asymétrie : médiane (EIQ) et Mann-Whitney
Christodorescu et al. 2026 [23]	cohorte prospective, 12 mois, n = 278	événement principal au cours du temps	Kaplan-Meier avec log-rank ; Cox multivarié	oui	log-rank p = 0,027, mais après ajustement p = 0,240 : exemple de confusion
Marc et al. 2023 [39]	transversale répétée, 1401 contre 1698	prévalence du VHC, du VHB (binaire) ; âge (numérique)	comparaisons de proportions et de moyennes	oui	prévalence du VHC 16,9% contre 10,5%, p < 0,0001 ; âge 59,7 contre 61 ans, p = 0,005 : statistiquement significatif, mais cliniquement négligeable

Quelques conclusions générales se dégagent du tableau. La plupart des articles utilisent des tests adaptés au schéma d'étude. Les problèmes les plus fréquents ne tiennent pas au choix du test principal, mais à la présentation des résultats : le test n'est pas nommé, les intervalles de confiance des OR manquent, les variables asymétriques sont décrites par la moyenne ± ÉT et les comparaisons intra-groupe sont présentées comme preuve de l'effet. Ce sont exactement les points que vérifie un relecteur et que vous retrouverez en rédigeant votre propre mémoire de fin d'études.

Étude de cas : une présentation exemplaire pour trois groupes

Pădurean et ses collaborateurs, avec la Dre Frandeș et le Dr Muntean, ont comparé l'estime de soi chez 245 personnes : 73 ayant une audition normale, 86 porteurs d'appareils auditifs et 86 personnes atteintes d'hypoacousie qui ne portent pas d'appareil [82]. Le résultat,

un score numérique, est comparé entre trois groupes indépendants ; l'ANOVA à un facteur est donc le choix naturel, suivie du test de Tukey pour les comparaisons deux à deux. Les auteurs rapportent le test global ($F = 299,00$, $p < 0,001$), la taille d'effet globale ($\eta^2 = 0,712$) et la taille d'effet pour la comparaison d'intérêt clinique : les porteurs d'appareils ont obtenu des scores de $35,41 \pm 5,32$ contre $22,99 \pm 4,53$ chez les non-porteurs ($p < 0,001$, d de Cohen = 2,515). Les associations avec les scores de handicap auditif et d'acouphènes sont données par la corrélation de Pearson ($r = -0,573$ et $r = -0,443$), et le modèle de régression multiple explique environ 65% de la variabilité. C'est un exemple complet d'application du format présenté dans ce chapitre : test adapté, test global puis post hoc, tailles d'effet. La limite du schéma est celle de toute étude transversale : nous ne savons pas si les appareils augmentent l'estime de soi ou si les personnes ayant une meilleure estime de soi choisissent plus facilement d'en porter.

8.10.3 Quand le schéma d'étude et la mesure d'effet ne concordent pas

Deux des articles du tableau soulèvent la même question : la mesure d'effet est-elle compatible avec le schéma déclaré ? Le rapport de cotes (odds ratio), le risque relatif et le hazard ratio ne sont pas synonymes. L'OR s'obtient à partir de tableaux 2 x 2 ou par régression logistique et peut être calculé dans n'importe quel schéma d'étude ; le RR exige un schéma dans lequel l'incidence peut être estimée (cohorte, essai clinique) ; le HR exige des données de type "temps jusqu'à l'événement", c'est-à-dire un suivi des patients dans le temps, et s'obtient en général par la régression de Cox.

Étude de cas : hazard ratio dans une étude transversale et dans une régression logistique

Danila et ses collaborateurs, avec le Dr Cioca, ont évalué des indices inflammatoires dérivés de la numération formule sanguine chez 336 patients atteints de COVID-19 (168 diabétiques et 168 non diabétiques, appariés sur le sexe, l'IMC et la sévérité) [105]. L'étude est déclarée transversale. Le volet consacré à l'exactitude diagnostique est cohérent avec ce schéma : pour le dNLR, au seuil de 2,685, la sensibilité était de 70,312% et la spécificité de 65,978%, avec AUC = 0,624, et pour l'ALRI AUC = 0,752. Le résumé rapporte toutefois aussi un hazard ratio pour le SII à 7 jours, 2,62 (IC 1,29-5,04, $p < 0,001$). Un hazard ratio suppose des patients suivis dans le temps et des dates de survenue de l'événement, ce qui ne correspond pas à l'étiquette "transversale" ; soit le schéma est en réalité une cohorte rétrospective, soit la mesure rapportée est une autre (par exemple un OR). Remarquons aussi qu'un IC allant de 1,29 à 5,04 est assez large pour une valeur p aussi petite, un détail qu'un lecteur vérifierait dans le texte intégral. Ion et ses collaborateurs, également avec le Dr Cioca, ont suivi de manière prospective 141 patients diabétiques atteints de COVID-19 et ont comparé les scores cliniques au moyen de courbes ROC (par exemple SOFA AUC 0,771 à l'admission et 0,873 au cinquième jour) [106]. Les méthodes mentionnent la régression logistique, mais les résultats sont donnés sous forme de hazard ratios (SOFA HR 3,07, IC à 95% 2,29-4,12 ; APACHE II HR 2,96, IC à 95% 2,21-3,96). La régression logistique produit des rapports de cotes (OR) ; le hazard ratio provient de la régression de Cox. Ces deux observations sont des exercices de lecture attentive et ne remettent pas en cause la valeur clinique des scores, mais elles illustrent une règle utile : le nom de la mesure d'effet doit correspondre au modèle statistique et au schéma d'étude.

8.10.4 La comparaison intra-groupe ne remplace pas la comparaison entre groupes

Une erreur d'interprétation très répandue consiste à conclure qu'un traitement fonctionne parce qu'"on a obtenu une baisse significative par rapport à la valeur initiale dans le groupe traité". Le test avant-après réalisé à l'intérieur d'un groupe montre seulement qu'un changement s'est produit. Le même changement peut apparaître dans le groupe témoin, par régression vers la moyenne, effet

placebo, saisonnalité ou simple participation à l'étude. Dans mg2026, le poids a diminué significativement même dans le bras avec simple conseil (test t apparié, p = 0,043). Seule la comparaison entre les bras (ANOVA, F = 18,25, p < 0,001) met en évidence l'effet des interventions.

Étude de cas : Aloe ferox et périndopril, comparaisons intra-groupe et entre groupes

Gherbon et ses collaborateurs, avec la Dre Frandeş, ont testé, dans une étude contrôlée contre placebo rapportée selon les recommandations CONSORT, l'effet de 2 gélules/jour d'Aloe ferox (460 mg) chez 20 personnes obèses, par rapport à 20 personnes appariées ayant reçu un régime et un placebo [85]. Les méthodes décrivent deux types de comparaisons : entre groupes à 3 mois et avant-après à l'intérieur des groupes. Le résumé donne, pour les mêmes variables, deux séries de valeurs p : par exemple, pour l'IMC P = 0,03 et P = 0,002, pour le LDLc P = 0,01 et P = 0,039, pour la glycémie à jeun P = 0,018 et P < 0,001, auxquelles s'ajoutent le cholestérol total (P = 0,032) et la pression diastolique (P = 0,006). Le lecteur doit vérifier dans le texte intégral quelle valeur répond à quelle question, car elles n'ont pas le même poids. Une baisse par rapport à la valeur initiale dans le groupe traité ne prouve pas à elle seule l'effet du complément : une baisse de l'IMC peut aussi survenir sous régime avec placebo. La comparaison qui répond à la question clinique est celle entre groupes, idéalement sur les variations entre le début et 3 mois ou par ANCOVA avec la valeur initiale comme covariable. Avec 20 patients par groupe, le choix entre le test t et le test de Mann-Whitney doit être justifié par la forme des distributions. Buda et ses collaborateurs, avec la Dre Tudor, ont comparé la thrombospondine-1 sérique chez des patients hypertendus présentant une dysfonction endothéliale, avant et après un an de périndopril : 8061,9 ± 3684,80 contre 15380 ± 5887 ng/mL, p < 0,001 [86]. Les données sont appariées ; le test adapté est donc le test t apparié ou le test de Wilcoxon, appliqué aux différences individuelles. Un écart-type proche de la moitié de la moyenne suggère une distribution asymétrique ; la variante non paramétrique ou la transformation logarithmique sont donc à envisager. Le groupe témoin sans hypertension $(9221{,}03 \pm 6510{,}21$ ng/mL) soulève une deuxième question, qui appelle un autre test : la comparaison avec les témoins se fait par un test pour groupes indépendants. Le résumé ne déclare pas le nombre de patients, information sans laquelle aucune valeur p ne peut être interprétée.

8.11 Exercices proposés

Pour chacune des situations 1-6, précisez : le type de la variable de résultat, le schéma (indépendant ou apparié, nombre de groupes), le test choisi et son alternative, les hypothèses statistiques.

1. Un médecin généraliste veut savoir si le nombre de consultations par an au cabinet (0, 1, 2,..., avec beaucoup de patients à 0-2 consultations et quelques-uns à plus de 15) diffère entre les patients avec et sans diabète.
2. La saturation en oxygène a été mesurée chez 25 patients atteints de BPCO au repos et après le test de marche de 6 minutes.
3. On compare la proportion de patients vaccinés contre la grippe dans trois départements (450, 380 et 520 patients).
4. Un questionnaire de qualité du sommeil (score 0-21) a été administré au même groupe de 30 internes après une garde, après un jour de repos et après une semaine de congé.
5. Deux radiologues ont classé indépendamment 120 radiographies thoraciques en "pneumonie" ou "pas de pneumonie". Quelle mesure utilisez-vous pour évaluer leur accord ?

6. On souhaite estimer le risque de développer un diabète en fonction du score FINDRISC, de l'âge et du sexe, dans une cohorte suivie pendant 10 ans.
7. Dans une étude, les auteurs ont comparé l'HbA1c entre 3 types de traitement au moyen de trois tests t et ont obtenu p = 0,03, p = 0,20 et p = 0,04. Calculez le risque global d'erreur de type I pour trois comparaisons indépendantes à $\alpha = 0,05$ ainsi que le seuil corrigé de Bonferroni. Quelles comparaisons restent significatives après correction ?
8. Un tableau 2 x 2 compte, en lignes, 12 et 48 patients et, en colonnes, 10 et 50. Calculez la plus petite fréquence attendue et décidez si vous appliquez le khi-deux ou le test de Fisher.
9. Sur le jeu de données mg2026, traitez entièrement, au format a-g, la question : "Le cholestérol total diffère-t-il entre les hommes et les femmes ?" (dans JASP, T-Tests > Independent Samples T-Test).
10. Sur le jeu de données mg2026, traitez au format a-g la question : "Le diabète confirmé par HGPO est-il associé au milieu de résidence ?"
11. Lisez le résumé d'un article récent de la spécialité qui vous intéresse (par exemple via PubMed) et remplissez pour cet article une ligne semblable à celles du tableau 8.3.

Pièges fréquents

- Choisir le test d'après le type de la variable de regroupement en oubliant le type de la variable de résultat (par exemple un khi-deux pour "la glycémie selon le sexe", simplement parce que le sexe est nominal).
- Confondre données appariées et groupes indépendants, surtout lorsque les mêmes patients apparaissent dans deux colonnes de la feuille de calcul.
- Vérifier la normalité sur l'échantillon total au lieu de chaque groupe (ou des différences, pour les données appariées).
- Passer automatiquement à un test non paramétrique lorsque le test de Levene donne $p < 0,05$: la solution naturelle est la variante de Welch.
- Rapporter le test qui "a marché" au lieu de celui fixé à l'avance.
- Tester les différences à l'inclusion entre les bras d'un essai randomisé et les interpréter comme un "échec de la randomisation".
- Interpréter une comparaison avant-après intra-groupe comme une preuve d'efficacité.
- Utiliser la corrélation pour démontrer l'accord entre deux méthodes.
- Compter les yeux, les dents ou les consultations comme s'il s'agissait de patients indépendants.

À retenir

- L'ordre de la décision : l'objectif (différence, association, accord, prédiction), le type de la variable de résultat, le nombre de groupes ou de moments, indépendant ou apparié, puis la distribution et la taille de l'échantillon.
- Pour chaque situation, il existe une variante paramétrique, une variante non paramétrique et une variante catégorielle (figure 8.1 et tableau 8.1).
- L'unité d'analyse est le patient (ou l'unité randomisée) ; les observations multiples d'un même patient ne sont pas indépendantes.
- Le test, la variable principale et la règle de repli sont fixés dans le protocole, avant de voir les données.
- Le format a-g : hypothèse de recherche, hypothèses statistiques, méthode et justification, résolution et logiciel, résultat avec graphique, conclusion statistique, conclusion de recherche.
- Un résultat complet comprend la statistique du test, les degrés de liberté, la valeur p exacte et une taille d'effet avec son intervalle de confiance ; la mesure d'effet doit correspondre au schéma d'étude (OR, RR, HR).

Questions d'autoévaluation

1. (QRU) Une étude mesure la pression artérielle chez 40 patients avant et après 8 semaines de traitement ; les différences suivent une distribution normale. Le test adapté est :
    a) le test t pour échantillons indépendants
    b) le test t apparié
    c) le test U de Mann-Whitney
    d) le test du khi-deux
    e) l'ANOVA à un facteur
2. (QRU) On compare le score de douleur (0-10) entre trois protocoles antalgiques, avec 12 patients par protocole. Le test adapté est :
    a) le test de Friedman
    b) le test de Wilcoxon pour données appariées
    c) le test de Kruskal-Wallis
    d) trois tests de Mann-Whitney sans correction
    e) le test de McNemar
3. (QRU) Pour évaluer l'accord entre le glucomètre portable et l'analyseur de laboratoire sur les mêmes 80 échantillons, la méthode la plus adaptée est :
    a) la corrélation de Pearson
    b) le test t pour échantillons indépendants
    c) l'analyse de Bland-Altman
    d) le test du khi-deux
    e) la régression logistique
4. (QRU) Le risque global d'au moins une erreur de type I lorsque l'on effectue 3 comparaisons indépendantes, chacune à $\alpha = 0{,}05$, est d'environ :
    a) 5%
    b) 10%
    c) 14%
    d) 15%
    e) 20%
5. (QRU) Dans une cohorte suivie pendant 5 ans, avec des durées de suivi différentes, l'effet d'un facteur sur la mortalité s'exprime le plus correctement par :
    a) un rapport de cotes issu d'un tableau 2 x 2
    b) un hazard ratio issu d'une régression de Cox
    c) le coefficient de corrélation de Spearman
    d) le V de Cramér
    e) la statistique H de Kruskal-Wallis
6. (QRM) Les données sont appariées dans les situations suivantes :
    a) la glycémie d'un même patient à 0 et à 6 mois
    b) l'acuité visuelle de l'œil droit et de l'œil gauche des mêmes patients
    c) l'HbA1c chez les hommes et chez les femmes
    d) des cas et des témoins appariés individuellement sur l'âge et le sexe
    e) la PAS chez les fumeurs et chez les non-fumeurs

7. (QRM) À propos du choix entre le test t et la variante non paramétrique, sont correctes les affirmations suivantes :
    a) la normalité se vérifie dans chaque groupe, et non sur l'échantillon total
    b) pour le test t apparié, on vérifie la normalité des différences
    c) un test de Shapiro-Wilk avec $p < 0,05$ pour $n = 300$ impose toujours un test non paramétrique
    d) des variances inégales (Levene $p < 0,05$) se traitent en règle générale par la variante de Welch
    e) le test se choisit après avoir vu quelle variante donne $p < 0,05$
8. (QRM) Un article décrit un essai randomisé à trois bras et rapporte $p < 0,05$ pour la baisse de la glycémie à l'intérieur de chaque bras. Sont correctes les affirmations suivantes :
    a) le résultat montre que chaque intervention produit un changement par rapport à la valeur initiale
    b) le résultat ne montre pas qu'une intervention est supérieure aux autres
    c) la comparaison adaptée est celle entre les bras (par exemple une ANOVA sur les variations)
    d) les valeurs p intra-bras prouvent l'efficacité de toutes les interventions
    e) la régression vers la moyenne peut contribuer aux baisses observées
9. (QRM) L'unité d'analyse n'est pas respectée (les observations ne sont pas indépendantes) si :
    a) on analyse 400 dents provenant de 40 patients comme 400 observations indépendantes
    b) on analyse 170 yeux provenant de 102 patients sans tenir compte du patient
    c) chaque patient fournit une seule valeur
    d) on compte les hospitalisations au lieu des patients, alors que certains patients ont plusieurs hospitalisations
    e) les patients sont randomisés individuellement et chacun n'apparaît qu'une seule fois dans l'analyse
10. (QRM) Le format standardisé a-g comprend :

a) les hypothèses statistiques en mots et sous forme mathématique
b) la justification du choix du test
c) le logiciel et la méthode de résolution
d) uniquement la valeur p, sans statistiques descriptives
e) la conclusion de recherche formulée en mots

Chapitre 9. Analyse des relations entre variables : la corrélation et la régression linéaire pour les variables numériques et ordinales

Objectifs

À la fin de ce chapitre, l'étudiant sera capable de :

- reconnaître un problème d'analyse bivariée et nommer le prédicteur (X) et la variable réponse (Y) ;
- construire et lire un diagramme de dispersion, en distinguant les variables indépendantes, fonctionnellement dépendantes et corrélées ;
- calculer à la main la covariance, le coefficient de corrélation de Pearson et le coefficient de Spearman sur un petit jeu de données ;
- interpréter le signe, l'intensité (selon la règle de Colton) et la signification statistique d'un coefficient de corrélation, avec son intervalle de confiance ;
- expliquer ce que signifie le coefficient de détermination R^2 et répondre à la question "quelle proportion de la variation est due à d'autres facteurs" ;
- estimer la droite de régression par la méthode des moindres carrés, interpréter la pente et l'ordonnée à l'origine et faire une prédiction, en reconnaissant le danger de l'extrapolation ;
- vérifier les conditions de la régression linéaire par l'analyse des résidus ;
- lire un tableau de régression linéaire multiple ou de régression logistique dans un article médical ;
- argumenter pourquoi une corrélation, aussi forte soit-elle, ne prouve pas à elle seule une relation causale et ne mesure pas la concordance entre deux méthodes.

9.1 L'analyse bivariée : deux variables mesurées chez le même patient

Jusqu'ici, nous avons presque toujours travaillé avec une variable numérique et un ou plusieurs groupes : la pression artérielle chez les femmes et chez les hommes, la glycémie dans les trois bras de l'étude, la glycémie à 0 et à 6 mois. Dans ce chapitre, nous changeons de question. Pour chaque patient, nous disposons de deux mesures numériques, par exemple l'indice de masse corporelle (IMC) et la pression artérielle systolique (PAS), et nous voulons savoir si elles varient ensemble. Les patients dont l'IMC est plus élevé ont-ils, en moyenne, une pression artérielle plus élevée ? Si oui, quelle est l'étroitesse de ce lien et de combien la pression augmente-t-elle pour chaque unité d'IMC ?

L'analyse qui suit simultanément le comportement de deux variables s'appelle analyse bivariée. Par convention, la variable dont nous pensons qu'elle influence l'autre, ou que nous utilisons pour prédire l'autre, est notée X et s'appelle prédicteur (variable explicative, variable indépendante). La variable que nous voulons expliquer ou prédire est notée Y et s'appelle variable réponse (variable dépendante, variable de résultat). Dans notre exemple, l'IMC est X et la PAS est Y. Le choix n'est pas toujours évident : entre la glycémie à jeun et l'hémoglobine glyquée (HbA1c), il n'existe pas de "prédicteur naturel", les deux reflètent le même processus, et le choix de X dépend de la question clinique.

Deux méthodes répondent à deux questions différentes. La corrélation mesure le sens et l'intensité de l'association entre deux variables numériques, traitées de façon symétrique : les deux sont considérées comme aléatoires, et le coefficient de corrélation entre X et Y est le même que celui entre Y et X. La régression construit un modèle mathématique qui décrit comment Y varie,

en moyenne, lorsque X varie, et permet de prédire Y pour une valeur donnée de X. En régression, les rôles ne sont pas symétriques : la droite de Y en fonction de X n'est pas la même que la droite de X en fonction de Y. Les manuels classiques de la discipline présentent les deux méthodes ensemble, car elles utilisent les mêmes sommes de carrés et se vérifient l'une l'autre [18,42,53].

9.1.1 Le diagramme de dispersion

Toute analyse bivariée commence par un graphique, et non par un nombre. Le diagramme de dispersion (angl. scatter plot) porte sur l'axe horizontal la variable X et sur l'axe vertical la variable Y ; chaque patient est un point. Face à un tel graphique, l'étudiant doit se poser quatre questions : quelle forme a le nuage de points (allongé le long d'une droite, courbe, informe), dans quel sens il va (croissant ou décroissant), à quel point les points sont resserrés autour d'une ligne imaginaire et s'il existe des points isolés, loin des autres.

La figure 9.1 montre les trois profils décrits dans les cours de la discipline. Dans le panneau A, les points sont dispersés uniformément, sans aucune direction : connaître X ne dit rien sur Y, les variables sont indépendantes. Dans le panneau B, tous les points se trouvent exactement sur une courbe : Y est entièrement déterminé par X, il s'agit d'une dépendance fonctionnelle (du type de la relation entre le côté et l'aire d'un carré). En biologie et en médecine, une relation aussi parfaite n'apparaît pratiquement jamais. Le panneau C correspond à la situation habituelle : les points se regroupent autour d'une droite, mais avec une dispersion visible. Y tend à augmenter avec X, mais pour une même valeur de X les patients ont des valeurs différentes de Y. Nous disons que les variables sont corrélées : le lien est statistique (stochastique), et non fonctionnel.

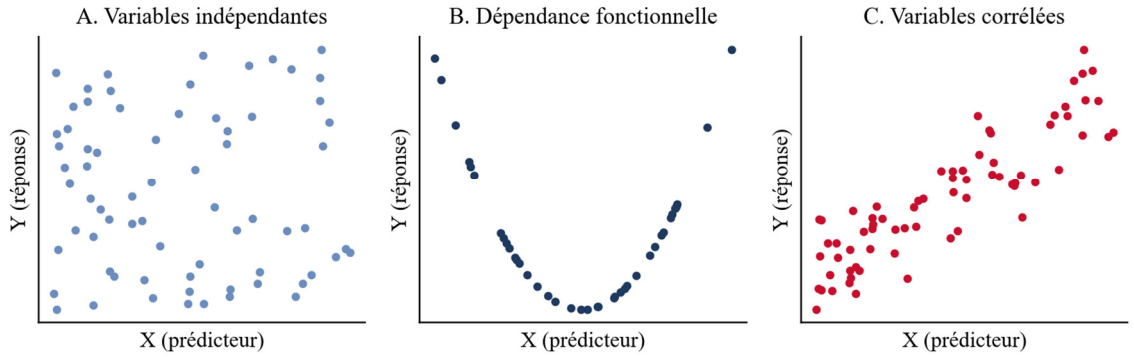


Figure 9.1. Profils de relation entre deux variables numériques. A : variables indépendantes, le nuage de points n'a aucune direction. B : dépendance fonctionnelle, chaque valeur de X détermine exactement une valeur de Y. C : variables corrélées, Y augmente en moyenne avec X, mais avec une dispersion autour de la tendance. Données simulées.

9.1.2 Ce que le graphique peut dire et ce qu'il ne peut pas dire

Le graphique montre des choses que les coefficients cachent. Une relation courbe (par exemple la mortalité en fonction de l'IMC, élevée aussi bien chez les sujets en insuffisance pondérale que chez les obèses, en forme de U) peut avoir un coefficient de Pearson proche de zéro, alors que le lien entre les variables est fort. Un seul patient aux valeurs extrêmes peut créer une corrélation apparente ou la masquer. Deux sous-groupes distincts (hommes et femmes, par exemple, pour la relation taille-poids) peuvent produire ensemble un nuage allongé, même si, à l'intérieur de chaque sous-groupe, le lien est faible. C'est pourquoi, en travaux pratiques, nous n'acceptons aucun coefficient de corrélation rapporté sans le diagramme correspondant.

Sur le jeu de données pédagogique mg2026 (160 adultes à risque cardiométabolique, chap. 1), nous suivrons tout au long du chapitre plusieurs paires de variables. Nous les annonçons dès maintenant, pour que vous puissiez comparer les résultats : IMC et PAS (corrélation faible, mais nettement significative), âge et PAS (corrélation faible), glycémie à jeun et HbA1c (corrélation

modérée, presque forte), HDL-cholestérol et triglycérides (pas de corrélation), activité physique autoévaluée et IMC (variable ordinale, pas de corrélation).

9.2 La covariance : un premier pas vers une mesure de l'association

Supposons que, pour chaque patient, nous calculions l'écart de X par rapport à la moyenne de X et l'écart de Y par rapport à la moyenne de Y. Si le patient a ses deux valeurs au-dessus de la moyenne, le produit des écarts est positif. S'il a ses deux valeurs au-dessous de la moyenne, le produit est également positif (moins par moins). S'il a une valeur au-dessus et une au-dessous de la moyenne, le produit est négatif. En additionnant les produits pour tous les patients, nous obtenons un nombre grand et positif lorsque les variables augmentent ensemble, grand et négatif lorsque l'une augmente pendant que l'autre diminue, et proche de zéro lorsqu'il n'existe pas de tendance commune.

En divisant cette somme par n - 1, nous obtenons la covariance d'échantillon :

$$\mathrm{cov}(X,\ Y) = \frac{\sum(x_i - m_x)\,(y_i - m_y)}{n-1}$$

où mx et my sont les moyennes des deux variables. Désormais, nous notons Sxy la somme des produits des écarts, et Sxx et Syy les sommes des carrés des écarts :

$$S_{\mathrm{xy}} = \sum(x_i - m_x)\,(y_i - m_y);\ S_{\mathrm{xx}} = \sum(x_i - m_x)^2;\ S_{\mathrm{yy}} = \sum(y_i - m_y)^2$$

Remarquez que Sxx / (n - 1) est précisément la variance de X vue au chapitre 2. La covariance généralise la variance : la variance est la covariance d'une variable avec elle-même.

La covariance présente un sérieux défaut pour l'interprétation : elle dépend des unités de mesure. Si nous exprimons le poids en grammes au lieu de kilogrammes, la covariance entre la taille et le poids devient 1000 fois plus grande, alors que la relation biologique reste la même. Face à une covariance de 43, nous ne pouvons pas dire si l'association est forte ou faible. La solution est la standardisation : nous divisons la covariance par le produit des écarts-types et obtenons un nombre sans unité, le coefficient de corrélation.

9.3 Le coefficient de corrélation de Pearson

9.3.1 La formule et le calcul à la main

Le coefficient de corrélation linéaire de Pearson est noté r pour l'échantillon et ρ (rho) pour la population :

$$r = \mathrm{cov}\frac{X, Y}{s_x \times s_y} = \frac{S_{\mathrm{xy}}}{\sqrt{S_{\mathrm{xx}} \times S_{\mathrm{yy}}}}$$

Les deux formes sont équivalentes, car les facteurs n - 1 se simplifient. Pour le calcul à la main, la seconde forme est plus commode.

Prenons un petit exemple, avec des chiffres choisis de sorte que le calcul puisse être suivi sans calculatrice. Huit patients d'un cabinet de médecine générale ont l'IMC et la PAS indiqués dans le tableau 9.1.

Tableau 9.1. Calcul à la main du coefficient de Pearson pour 8 patients $\left(X = \text{IMC},\ \frac{\text{kg}}{m^2};\ Y = \text{PAS},\ \text{mmHg}\right)$.

Patient	xi	yi	xi - mx	yi - my	(xi - mx)(yi - my)	$(x_i - m_x)^2$	$(y_i - m_y)^2$
1	23	118	-6	-13	78	36	169
2	25	126	-4	-5	20	16	25

Patient	xi	yi	xi - mx	yi - my	(xi - mx)(yi - my)	$(x_i - m_x)^2$	$(y_i - m_y)^2$
3	26	121	-3	-10	30	9	100
4	28	129	-1	-2	2	1	4
5	29	135	0	4	0	0	16
6	31	131	2	0	0	4	0
7	33	140	4	9	36	16	81
8	37	148	8	17	136	64	289
Total	232	1048	0	0	$S_{xy} = 302$	$S_{xx} = 146$	$S_{yy} = 684$

Les étapes sont les suivantes. Les moyennes : $m_x = \frac{232}{8} = \frac{29\ \text{kg}}{m^2}$ et $m_y = \frac{1048}{8} = 131$ mmHg. Les écarts se calculent ligne par ligne, et leur somme doit être nulle dans chaque colonne ; si elle ne l'est pas, vous avez fait une erreur de soustraction. On multiplie ensuite les écarts et on les élève au carré. La covariance vaut 302 / 7 = 43,14 kg/m² x mmHg (une unité de mesure difficile à se représenter, ce qui confirme la discussion ci-dessus). Les écarts-types sont $s_x = \sqrt{\frac{146}{7}} = \frac{4{,}57\ \text{kg}}{m^2}$ et $s_y = \sqrt{\frac{684}{7}} = 9{,}89$ mmHg. Finalement :

$$r = \frac{302}{\sqrt{146 \times 684}} = \frac{302}{\sqrt{99\,864}} = \frac{302}{316{,}01} = 0{,}956$$

Tous les produits sont positifs ou nuls : chaque patient dont l'IMC est au-dessus de la moyenne a aussi une PAS au-dessus de la moyenne ou égale à celle-ci. Le coefficient est proche de 1, ce qui indique une forte corrélation positive. Sur des données réelles, vous n'obtiendrez pas souvent de telles valeurs ; l'exemple a été construit pour la clarté. Nous continuerons avec les mêmes huit paires pour la régression (section 9.7) et dans le problème résolu 9.1.

9.3.2 Les propriétés du coefficient de Pearson

Le coefficient r prend des valeurs entre -1 et +1. Le signe indique le sens de l'association : r > 0 signifie une corrélation directe (positive), les deux variables augmentent ensemble ; r < 0 signifie une corrélation inverse (négative), quand l'une augmente l'autre tend à diminuer. La valeur r = +1 correspond à une corrélation directe parfaite (tous les points sur une droite croissante), r = -1 à une corrélation inverse parfaite, et r = 0 à l'absence de relation linéaire.

Quelques propriétés sont importantes pour l'examen. Le coefficient n'a pas d'unité de mesure et ne change pas si nous transformons linéairement l'une des variables (le poids en grammes ou en kilogrammes, la température en degrés Celsius ou Fahrenheit donnent le même r). Il est symétrique : r(X, Y) = r(Y, X). Il ne mesure que la relation linéaire ; une relation courbe forte peut avoir un r faible. Il est sensible aux valeurs extrêmes, car il repose sur des moyennes et sur les carrés des écarts. Enfin, point très important, il dépend de l'étendue des valeurs présentes dans l'échantillon : si nous étudions la relation IMC-PAS uniquement chez des patients dont l'IMC est compris entre 27 et 29, r sera beaucoup plus petit que dans l'échantillon complet, car nous avons supprimé précisément la variation qui rend la relation visible (ce phénomène s'appelle la restriction de l'étendue).

9.3.3 Quelle est la force d'une corrélation ?

La figure 9.2 montre à quoi ressemblent les nuages de points pour quelques valeurs de r. Elle mérite d'être mémorisée visuellement : un r de 0,5 paraît beaucoup plus "dispersé" que ne s'y attendent d'habitude les étudiants, tandis qu'un r de 0,95 apparaît comme une bande étroite.

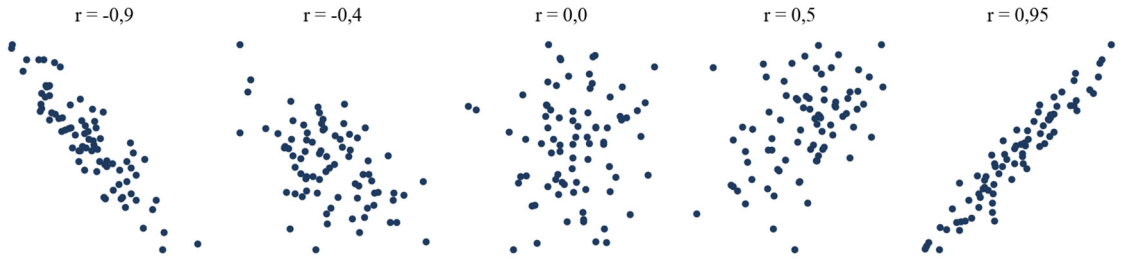


Figure 9.2. Nuages de points simulés pour r = -0,9 ; -0,4 ; 0 ; 0,5 et 0,95 (80 points chacun). L'intensité de la corrélation se lit au degré de resserrement des points autour d'une droite, et non à la pente de celle-ci.

Pour traduire le nombre en mots, les cours de la discipline utilisent la règle empirique proposée par Colton en 1974. La valeur absolue de r s'interprète de la façon suivante :

Tableau 9.2. Règle empirique de Colton pour l'interprétation du coefficient de corrélation.

Valeur de r	**Interprétation**
$-0,25 < r < 0,25$	pas de corrélation (corrélation négligeable)
$0,25 \leq r < 0,50$ ou $-0,50 < r \leq -0,25$	corrélation faible
$0,50 \leq r < 0,75$ ou $-0,75 < r \leq -0,50$	corrélation modérée
$r \geq 0,75$ ou $r \leq -0,75$	corrélation forte
$r = 1$ ou $r = -1$	corrélation parfaite

Sur mg2026, la corrélation IMC-PAS (r = 0,405) est faible, la corrélation âge-PAS (r = 0,256) est faible, tout juste à la limite inférieure, et la corrélation glycémie-HbA1c (r = 0,738) est modérée, très proche du seuil de 0,75. Le HDL et les triglycérides (r = 0,076) ne sont pas corrélés.

La règle de Colton n'est pas la seule. En psychologie et dans les sciences du comportement, on utilise fréquemment l'échelle de Cohen, qui considère r = 0,1 comme un petit effet, r = 0,3 comme un effet moyen et r = 0,5 comme un grand effet [76]. D'autres manuels utilisent des seuils de 0,2 en 0,2 (très faible, faible, modérée, forte, très forte). Avec l'échelle de Cohen, le r = 0,405 entre l'IMC et la PAS serait un effet "moyen à grand", et non une corrélation "faible". Il n'existe pas une échelle correcte et une échelle fausse : il existe des conventions. L'important est de dire quelle échelle vous utilisez et de ne pas changer d'échelle selon le résultat qui vous arrange. À l'examen de la discipline, c'est la règle de Colton qui s'applique.

9.3.4 Les conditions d'application du coefficient de Pearson

Le coefficient de Pearson peut être calculé pour n'importe quelle paire de colonnes numériques, mais son interprétation et le test de signification exigent trois conditions. Première condition : la relation entre les variables doit être linéaire, ce qui se vérifie sur le diagramme de dispersion. Deuxième condition : pour le test et pour l'intervalle de confiance, les paires doivent provenir d'une distribution normale bivariée ; en pratique, nous vérifions que chaque variable a une distribution approximativement normale (histogramme, graphique Q-Q, test de Shapiro-Wilk, chap. 3). Troisième condition : il ne doit pas exister de valeurs extrêmes qui dominent le calcul. S'y ajoute la condition générale d'indépendance des observations : chaque patient apparaît une seule fois, et non plusieurs fois avec des mesures répétées.

La figure 9.3 montre pourquoi la troisième condition compte autant. Dans le panneau A, 25 points simulés n'ont aucun lien entre eux (r = -0,16). Dans le panneau B, on a ajouté un seul point, loin de tous les autres, et r a bondi à 0,64. Un étudiant qui ne rapporte que le nombre dirait qu'il

existe une corrélation modérée ; le graphique montre que la "corrélation" est produite par un seul patient.

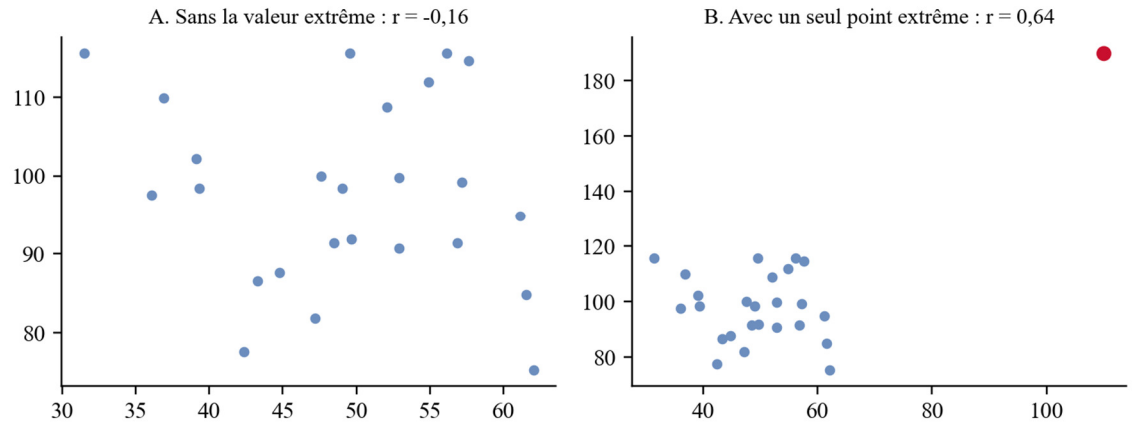


Figure 9.3. Influence d'une seule valeur extrême sur le coefficient de Pearson. A : 25 points sans relation, r = -0,16. B : les mêmes points plus un seul point extrême (en rouge), r = 0,64. Données simulées.

Que faire d'une valeur extrême ? Nous vérifions d'abord qu'il ne s'agit pas d'une erreur de saisie (une taille de 1,72 saisie en mètres au lieu de 172 cm, comme nous l'avons vu dans le fichier mg2026_brut.csv). Si la valeur est réelle, nous rapportons l'analyse avec et sans elle et nous préférons un coefficient fondé sur les rangs (Spearman), beaucoup moins sensible. Supprimer sans justification les patients qui dérangent n'est pas une option.

Sur mg2026, l'IMC ne respecte pas la normalité (asymétrie 1,42 ; Shapiro-Wilk $p < 0,001$), à cause de quelques patients atteints d'obésité sévère (le maximum est de 48,3 kg/m²). La PAS suit une distribution normale ($p = 0,303$). Par conséquent, pour la paire IMC-PAS, nous rapporterons à la fois Pearson et Spearman, et nous verrons que les deux conduisent à des conclusions semblables.

9.3.5 Le test de signification du coefficient de corrélation

Un r calculé sur un échantillon n'est jamais exactement nul, même si, dans la population, les variables n'ont aucun lien. La question inférentielle est la suivante : la valeur observée pourrait-elle apparaître du seul fait des fluctuations d'échantillonnage si, dans la population, $\rho = 0$? Les hypothèses sont :

H0 : $\rho = 0$ (dans la population, les deux variables ne sont pas corrélées linéairement) H1 : $\rho \neq 0$ (dans la population, les variables sont corrélées)

Sous H0, la statistique du test suit une distribution de Student à n - 2 degrés de liberté :

$$t = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}}$$

Nous perdons deux degrés de liberté parce que nous avons estimé deux moyennes. Pour l'exemple à huit patients : $t = \frac{0,956 \times \sqrt{6}}{\sqrt{1-0,913}} = \frac{0,956 \times 2,449}{0,294} = 7,95$. La valeur critique pour 6 degrés de liberté et $\alpha = 0,05$ en bilatéral est 2,447 ; comme $7,95 > 2,447$, on rejette H0 ($p = 0,0002$).

Sur mg2026, pour IMC-PAS : $t = \frac{0,405 \times \sqrt{158}}{\sqrt{1-0,164}} = \frac{0,405 \times 12,57}{0,914} = 5,56$, avec 158 degrés de liberté, $p < 0,001$. Pour âge-PAS : $t = 3,34$, $p = 0,001$. Pour HDL-triglycérides : $t = 0,96$, $p = 0,34$, on ne rejette pas H0.

La formule cache une leçon que vous rencontrerez souvent dans les articles : la signification dépend massivement de n. Le tableau 9.3 indique le plus petit r qui devient significatif à α = 0,05 pour différentes tailles d'échantillon.

Tableau 9.3. Valeur minimale de |r| significative à α = 0,05 (test bilatéral), en fonction de n.

n	**10**	**20**	**30**	**50**	**100**	**160**	**867**
r minimal significatif	0,632	0,444	0,361	0,279	0,197	0,155	0,067

Avec 10 patients, une corrélation modérée de 0,60 n'est pas significative. Avec 867 patients, une corrélation de 0,07, que Colton qualifierait de "pas de corrélation", est significative. La signification statistique répond seulement à la question "ρ est-il différent de zéro ?", et non à la question "la relation est-elle importante ?". Nous verrons un exemple publié à la section 9.5.

9.3.6 L'intervalle de confiance de r : la transformation de Fisher

La distribution d'échantillonnage de r n'est pas symétrique lorsque ρ est loin de zéro : r ne peut pas dépasser 1, si bien que les valeurs "s'entassent" vers l'extrémité. Le statisticien R. A. Fisher a proposé une transformation qui rend la distribution approximativement normale :

$z = 0{,}5 \times \ln\left[\frac{1+r}{1-r}\right]$, avec l'erreur standard $\mathrm{ES}(z) = \frac{1}{\sqrt{n-3}}$

On calcule l'intervalle sur l'échelle z, $z \pm 1{,}96 \times \mathrm{ES}(z)$, puis on retransforme les bornes en r avec la formule inverse $r = \frac{e^{2z}-1}{e^{2z}+1}$. Pour IMC-PAS sur mg2026 : $z = 0{,}5 \times \ln\left(\frac{1{,}405}{0{,}595}\right) = 0{,}429$; $\mathrm{ES} = \frac{1}{\sqrt{157}} = 0{,}0798$; l'intervalle sur l'échelle z est 0,429 ± 0,156, soit de 0,273 à 0,586 ; après retransformation, l'IC à 95% de ρ est 0,266-0,527.

L'intervalle en dit plus que la valeur p. Nous savons maintenant que, très probablement, la corrélation dans la population est au moins faible (au-dessus de 0,25) et au plus modérée (au-dessous de 0,53). Pour glycémie-HbA1c, l'IC à 95% est 0,659-0,802 : la corrélation est au moins modérée et peut être forte. Pour HDL-triglycérides, l'IC à 95% va de -0,080 à 0,229 et contient zéro, en accord avec p = 0,34. Pour l'échantillon de 8 patients, bien que r = 0,956, l'intervalle est large : 0,769-0,992, car n - 3 = 5 est très petit.

9.3.7 Le coefficient de détermination R^2

En élevant r au carré, nous obtenons le coefficient de détermination, R² (ou r²). Il indique quelle proportion de la variation de Y est "expliquée", au sens statistique, par la relation linéaire avec X. Il prend des valeurs entre 0 et 1 (ou entre 0% et 100%) et ne conserve plus le signe : une corrélation de -0,6 et une de +0,6 ont le même $R^2 = 0{,}36$. C'est pourquoi nous disons que R² renseigne sur la force de la corrélation, et non sur son sens.

Pour l'IMC et la PAS sur mg2026, $R^2 = 0{,}405^2 = 0{,}164$. Environ 16,4% de la variabilité de la pression systolique entre les patients peut être attribuée aux différences d'IMC ; les 83,6% restants relèvent d'autres facteurs (l'âge, le sexe, le tabagisme, le sel alimentaire, la génétique, les médicaments, la variation d'une mesure à l'autre). C'est exactement la question que posent les travaux pratiques de la discipline à propos de la taille et du poids : "dans quelle proportion la variation du poids est-elle causée par des facteurs autres que la variation de la taille ?". La réponse est 1 - R². Sur mg2026, la taille et le poids ont r = 0,652, donc $R^2 = 0{,}424$, et $1 - R^2 = 0{,}576$: environ 57,6% de la variation du poids relève de facteurs autres que la taille.

Pour la glycémie et l'HbA1c, $R^2 = 0{,}545$: un peu plus de la moitié de la variation de l'HbA1c est associée linéairement à la variation de la glycémie à jeun. Pour l'âge et la PAS, $R^2 = 0{,}066$, soit seulement 6,6%. Un r "faible" de 0,26 signifie, traduit en R², que l'âge explique très peu des différences de pression artérielle entre nos patients. Les étudiants sont souvent surpris de

voir à quel point un r modéré devient petit une fois élevé au carré : r = 0,5 n'explique qu'un quart de la variation.

Une dernière précision concerne le vocabulaire. "Expliqué" ne signifie pas "causé". L'expression des énoncés classiques ("la variation du poids causée par d'autres facteurs") doit être lue au sens statistique : la proportion de la variation qui n'est pas liée linéairement à la taille. Le fait que la taille "explique" 42% de la variation du poids ne signifie pas que 42% des kilogrammes d'une personne proviennent de sa taille.

Étude de cas : corrélations en élastographie hépatique (l'équipe d'échographie de Timișoara)

Les collègues de la discipline qui travaillent avec l'équipe de gastro-entérologie ont publié plusieurs études dans lesquelles une nouvelle méthode échographique est comparée à une méthode de référence, la corrélation étant le premier résultat rapporté. Dans une étude d'exactitude diagnostique portant sur 64 personnes (37 atteintes de cirrhose, 27 volontaires sains), le paramètre obtenu par échographie de contraste (CEUS-PAT) a présenté une corrélation r = 0,81 avec l'élastographie impulsionnelle à vibration contrôlée (VCTE) [2]. Dans une étude prospective portant sur 165 patients, l'élastographie time-harmonic a obtenu, par rapport à la même référence, r = 0,82, p < 0,0001 [3]. Pour la stéatose, le paramètre UGAP a présenté avec le CAP r = 0,73, p < 0,0001, sur 179 sujets consécutifs [58], et dans une étude de 2026 portant sur 169 patients, la fraction de graisse échographique (USFF) a présenté avec le CAP r = 0,81, p < 0,001, tandis que deux autres paramètres ont obtenu r = 0,66 et r = 0,83 [65].

Selon la règle de Colton, les valeurs de 0,81, 0,82 et 0,83 indiquent une corrélation forte, et 0,73 et 0,66 une corrélation modérée. Le résumé de la dernière étude qualifie pourtant de "modérée" une corrélation de 0,83, ce qui illustre précisément le problème des échelles différentes discuté plus haut : les auteurs peuvent utiliser un autre seuil, et le lecteur doit juger le nombre, et non l'adjectif. La seconde leçon est plus importante. Une corrélation de 0,82 entre deux méthodes de mesure de la rigidité hépatique ne signifie pas qu'elles donnent les mêmes valeurs et qu'elles peuvent être utilisées l'une à la place de l'autre. La corrélation montre que les patients sont ordonnés de façon semblable ; la concordance s'évalue autrement (section 9.6), et l'étude de 2020 a rapporté, pour la reproductibilité, des coefficients de corrélation intraclasse de 0,92-0,97 [3]. Ce qui serait encore utile dans un tel résumé : l'intervalle de confiance de r et le diagramme de dispersion, qui montrerait si la relation reste linéaire sur tout le domaine de fibrose.

9.4 La corrélation pour des données ordinales ou asymétriques : les coefficients de Spearman et de Kendall

9.4.1 L'idée : travailler avec des rangs, et non avec des valeurs

Le coefficient de Pearson suppose des variables numériques, une relation linéaire et une distribution approximativement normale. Beaucoup de variables médicales ne respectent pas ces conditions. Les scores sur des échelles de Likert, les stades de la maladie, la classe NYHA, le niveau d'études ou l'activité physique autoévaluée (1-5) de mg2026 sont ordinaux : nous connaissons l'ordre des catégories, mais la distance entre "2" et "3" n'est pas nécessairement la même que celle entre "4" et "5". D'autres variables sont numériques, mais très asymétriques (les triglycérides, l'IMC dans mg2026), ou présentent des valeurs extrêmes.

Le coefficient de corrélation des rangs de Spearman (noté rs ou ρ de Spearman, angl. Spearman's rank correlation) résout le problème par un artifice simple : il remplace chaque valeur par son rang (sa place dans l'ordre croissant) et calcule le coefficient de Pearson sur les rangs. Le résultat mesure à quel point le lien est monotone : si Y augmente chaque fois que X augmente,

même de façon non linéaire, le coefficient de Spearman vaut +1. Comme toutes les méthodes fondées sur les rangs du chapitre 7, il est robuste aux valeurs extrêmes : l'IMC le plus élevé reçoit le rang 160, qu'il soit de 40 ou de 48,3.

9.4.2 Le calcul à la main

Lorsqu'il n'existe pas de valeurs égales (ex aequo), le coefficient de Spearman peut se calculer avec une formule abrégée :

$$r_s = 1 - \frac{6 \times \sum {d_i}^2}{n \times (n^2 - 1)}$$

où di est la différence entre le rang de X et le rang de Y chez le patient i. Prenons huit patients chez qui on a mesuré le score FINDRISC et la glycémie à jeun.

Tableau 9.4. Calcul du coefficient de Spearman pour le score FINDRISC et la glycémie à jeun chez 8 patients.

Patient	FINDRISC	Glycémie (mg/dL)	Rang X	Rang Y	di	${d_i}^2$
1	5	92	1	2	-1	1
2	8	88	2	1	1	1
3	9	101	3	4	-1	1
4	12	99	4	3	1	1
5	14	115	5	6	-1	1
6	15	108	6	5	1	1
7	18	126	7	8	-1	1
8	21	119	8	7	1	1
Total					0	8

$$r_s = 1 - \frac{6 \times 8}{8 \times (64 - 1)} = 1 - \frac{48}{504} = 1 - 0{,}095 = 0{,}905$$

La corrélation est positive et forte : les patients dont le score FINDRISC est plus élevé tendent à avoir une glycémie plus élevée. L'ordre n'est pas parfaitement respecté (les paires 1-2, 3-4, 5-6, 7-8 échangent leurs places), mais les écarts ne portent que sur un seul rang. Pour un petit n, la signification de rs se lit dans des tables spéciales ou dans le logiciel (ici p = 0,002) ; pour n au-delà d'environ 10-20, on peut utiliser la même formule t que pour Pearson.

9.4.3 Les valeurs égales (ex aequo)

Les variables ordinales comportent presque toujours des valeurs égales. Les patients ayant la même valeur reçoivent la moyenne des rangs qu'ils auraient occupés : si deux patients ont une activité de 2 et occuperaient les places 2 et 3, tous deux reçoivent le rang 2,5. Avec beaucoup de valeurs égales, la formule abrégée devient inexacte et il faut calculer le coefficient de Pearson sur les rangs (ce que font automatiquement tous les logiciels). Le problème résolu 9.2 montre un exemple dans lequel la différence entre les deux calculs est importante : -0,85 avec la formule abrégée contre -0,91 avec le calcul correct.

9.4.4 Le coefficient tau de Kendall

Une autre solution, moins utilisée dans les articles cliniques mais préférée par certains statisticiens pour les petits échantillons et en présence de nombreuses valeurs égales, est le coefficient tau (τ) de Kendall. Son principe est le dénombrement des paires : pour deux patients quelconques, nous nous demandons si leur ordre selon X est le même que leur ordre selon Y (paire concordante) ou s'il est inversé (paire discordante). Tau est la différence entre la proportion de paires concordantes et celle de paires discordantes. Dans l'exemple du tableau 9.4, il existe 28 paires possibles de

patients ; seules les quatre paires voisines qui échangent leurs places sont discordantes, donc $\tau = (24 - 4) / 28 = 0{,}71$. Les valeurs de τ sont en général plus petites que celles de Spearman pour les mêmes données (sur mg2026, pour IMC-PAS, $\tau = 0{,}258$ contre $rs = 0{,}372$), si bien que les seuils de Colton ne s'appliquent pas directement à τ. La variante τ-b, rapportée par JASP et par d'autres logiciels, corrige pour les valeurs égales.

9.4.5 Ce que montrent les coefficients de rang sur mg2026

L'activité physique autoévaluée (1-5) est ordinale ; sa relation avec l'IMC s'évalue donc avec le coefficient de Spearman : $rs = -0{,}012$, $p = 0{,}88$. Le signe est celui attendu (plus d'activité, un IMC un peu plus faible), mais la valeur est pratiquement nulle : dans notre échantillon, ce que les patients déclarent de leur activité physique ne dit rien de leur IMC. Les explications possibles tiennent à la qualité de la mesure (l'autoévaluation est imprécise) et au fait que l'échantillon ne comprend que des personnes en surpoids. Entre le niveau d'études et la satisfaction à 6 mois, $rs = 0{,}125$, $p = 0{,}115$: ni corrélation selon Colton, ni signification.

Pour l'IMC et la PAS, où l'IMC est asymétrique, le coefficient de Spearman donne $rs = 0{,}372$ ($p < 0{,}001$), très proche du coefficient de Pearson, $r = 0{,}405$. Lorsque les deux coefficients sont proches, les valeurs extrêmes ne faussent pas beaucoup le résultat. Lorsqu'ils diffèrent nettement, il faut regarder le graphique avec attention.

Étude de cas : coefficients de rang dans des études de validation et d'éducation médicale

Quatre articles de collègues montrent comment rapporter correctement des corrélations de rang. Dans une étude d'exactitude diagnostique portant sur 157 patients atteints d'hépatite chronique B ou C, avec la biopsie hépatique comme standard de référence, la rigidité mesurée par deux méthodes a été corrélée au stade de fibrose (une variable ordinale, F0-F4) au moyen du coefficient de Spearman : $\rho = 0{,}826$, $p < 0{,}001$, avec un IC à 95% de rho de 0,769-0,870 [107]. C'est l'un des rares résumés qui donnent l'intervalle de confiance, et dans le problème résolu 9.4 nous vérifierons à la main que cet intervalle correspond à la transformation de Fisher. Dans une étude de validation d'un questionnaire d'acceptabilité de la technologie mobile dans le diabète, portant sur 103 patients, les auteurs ont rapporté à la fois des coefficients de Spearman ($\rho = -0{,}429$; $\rho = -0{,}322$; $\rho = 0{,}466$; tous avec $P < 0{,}001$) et un coefficient de Kendall ($\tau = 0{,}51$; $P < 0{,}001$) [54] ; remarquez que tau peut être élevé, même s'il est en général plus petit que rho.

Une étude par questionnaire anonyme menée auprès de 121 étudiants en médecine a utilisé la médiane et l'écart interquartile pour des scores sur une échelle de 1 à 10, ainsi que la corrélation de Spearman [16]. Deux corrélations du résumé méritent d'être mises côte à côte : $R = 0{,}869$, $p < 0{,}001$, et $R = 0{,}107$, $p = 0{,}246$, cette dernière entre les connaissances mesurées et les connaissances autoévaluées. La seconde corrélation est pratiquement nulle et non significative : ce que les étudiants croient savoir ne dit pas grand-chose de ce qu'ils savent réellement. Pour un cours de biostatistique, ce résultat est une invitation à la modestie et, du point de vue statistique, il montre qu'un résultat non significatif est lui aussi informatif, si l'échantillon n'est pas minuscule. Enfin, une étude sur les données de surveillance de la COVID-19 du département de Timiș a trouvé des corrélations de Spearman de rho = 0,852 et 0,821 ($p < 0{,}001$) entre l'incidence chez le personnel soignant et la tendance dans la communauté [40]. Ici, l'unité d'analyse n'est pas la personne, mais la période de temps : deux séries temporelles qui augmentent et diminuent ensemble. La corrélation est forte, mais elle ne dit pas où un soignant donné a été infecté ; nous discutons ce piège dans la section suivante.

9.5 Corrélation ne signifie pas causalité

9.5.1 Quatre explications pour une corrélation

Lorsque deux variables sont corrélées, il existe au moins quatre explications possibles, et la causalité directe n'est que l'une d'elles. La première est le hasard : si l'on teste suffisamment de paires de variables, certaines ressortiront corrélées de façon significative du seul fait des fluctuations d'échantillonnage (avec $\alpha = 0{,}05$, une paire sur 20 sans aucun lien réel). La deuxième est la causalité inverse : Y peut influencer X, et non l'inverse. Les patients atteints de maladie coronarienne font moins d'effort physique parce qu'ils ont un angor, et pas seulement l'inverse. La troisième est la confusion : une troisième variable (facteur de confusion, angl. confounder) influence les deux variables et crée entre elles une corrélation sans lien direct. Les ventes de glaces et le nombre de noyades sont corrélés parce que les deux augmentent en été. Sur mg2026, l'âge est corrélé à la fois à la PAS et au risque de diabète ; une corrélation brute entre la PAS et la glycémie devrait être vérifiée après ajustement sur l'âge. La quatrième explication est la causalité proprement dite, qui doit être argumentée, et non présumée.

Pour peser les arguments de causalité dans les études observationnelles, l'épidémiologiste Austin Bradford Hill a proposé en 1965 un ensemble de critères : la force de l'association, la constance entre les études, la spécificité, la temporalité (la cause précède l'effet), le gradient biologique (la relation dose-réponse), la plausibilité, la cohérence avec les connaissances existantes, les preuves expérimentales et l'analogie [108]. Seule la temporalité est indispensable. Ces critères sont discutés en détail au chapitre 10 ; retenons ici qu'un coefficient de corrélation, aussi élevé soit-il, couvre au mieux la force de l'association et, parfois, le gradient.

9.5.2 Le biais écologique

Les corrélations calculées sur des données agrégées (pays, départements, mois) peuvent être complètement différentes de celles observées au niveau individuel. Si les pays où la consommation de graisses est plus élevée ont une mortalité coronarienne plus élevée, il n'en découle pas automatiquement que les personnes qui mangent davantage de graisses meurent plus souvent de maladie coronarienne ; d'autres facteurs diffèrent peut-être aussi dans ces pays (le revenu, le tabagisme, l'accès aux soins). L'inférence erronée du groupe à l'individu s'appelle le biais écologique (erreur écologique). La corrélation entre l'incidence de la COVID-19 chez le personnel soignant et celle dans la communauté, discutée plus haut, est une corrélation écologique de ce type : utile pour la surveillance épidémiologique, mais sans éclairer la source de l'infection de chaque employé [40].

9.5.3 La régression vers la moyenne

Un phénomène statistique qui a trompé de nombreuses générations de médecins est la régression vers la moyenne. Si nous sélectionnons des patients parce qu'ils ont présenté une valeur extrême lors d'une mesure (par exemple une PAS supérieure à 160 mmHg lors d'une seule consultation), lors de la mesure suivante leurs valeurs seront, en moyenne, plus proches de la moyenne de la population, même sans aucun traitement. La raison : toute mesure comporte aussi une composante aléatoire (stress, café, brassard posé à la hâte), et les sujets sélectionnés sur une valeur extrême ont eu, en moyenne, une composante aléatoire défavorable, qui ne se répète pas. Plus la corrélation entre les deux mesures est faible, plus la régression vers la moyenne est importante. Ce phénomène explique pourquoi de nombreux traitements semblent fonctionner dans des études sans groupe témoin et pourquoi la randomisation, avec un groupe témoin mesuré de la même manière (chap. 13), est la seule protection sûre. Le terme "régression" lui-même vient de là : Francis Galton a observé au XIXe siècle que les fils de parents très grands sont, en moyenne, moins grands que leurs parents.

9.5.4 Signification statistique et importance clinique

Le dernier piège de cette section tient à la taille de l'échantillon. Nous avons vu dans le tableau 9.3 qu'avec des centaines de patients, des corrélations très faibles deviennent significatives. Un p petit montre que ρ est probablement différent de zéro, mais non que la relation compte sur le plan clinique. Rapportez toujours r (ou R²) avec p et, idéalement, avec l'intervalle de confiance.

Étude de cas : une corrélation significative, mais négligeable

Une étude rétrospective portant sur 867 patients atteints d'un cancer du côlon, opérés dans un hôpital universitaire de Timişoara sur quatre ans (deux avant et deux pendant la pandémie de COVID-19), a suivi les changements dans la prise en charge chirurgicale [55]. Les résultats principaux sont des comparaisons de proportions entre périodes : la diminution des interventions programmées ($p < 0,001$), l'augmentation des interventions en urgence ($p < 0,001$) et l'augmentation des cas au stade IV pendant la deuxième année de pandémie ($p = 0,045$). Le résumé rapporte aussi une corrélation inverse entre l'âge du patient et la durée de l'intervention : $r = -0,088$, $p = 0,01$.

Vérifions. Avec $n = 867$, $t = \frac{0,088\times\sqrt{865}}{\sqrt{1-0,0077}} = \frac{0,088\times 29,41}{0,996} = 2,60$, et pour 865 degrés de liberté $p = 0,0095$, soit environ 0,01, en accord avec le résumé. La corrélation est donc statistiquement significative. Mais, selon Colton, $|r| = 0,088 < 0,25$ signifie "pas de corrélation", et $R^2 = 0,0077$: l'âge explique moins de 1% de la variation de la durée de l'opération. Ce que dit le résultat : dans cette série, les patients plus âgés n'ont pas eu d'opérations plus longues, mais, très légèrement, plus courtes. Ce qu'il ne dit pas : que l'âge serait un prédicteur utile de la durée. Ce résultat est un très bon exemple didactique de la différence entre "significatif" et "important" ; un lecteur attentif voudrait voir aussi le diagramme de dispersion et, éventuellement, une analyse ajustée sur le type d'intervention (en urgence ou programmée), qui peut être un facteur de confusion.

9.6 Corrélation et concordance : la méthode de Bland-Altman

Une erreur fréquente dans la littérature médicale, signalée dès 1986 par Bland et Altman, consiste à utiliser le coefficient de corrélation pour montrer que deux méthodes de mesure "concordent" [102]. Supposons qu'un nouveau tensiomètre donne systématiquement 10 mmHg de plus que l'appareil de référence. Les points se trouveront parfaitement sur une droite parallèle à la première bissectrice, r sera proche de 1, mais les méthodes ne seront pas interchangeables : un patient à 135 mmHg sur l'appareil de référence serait classé hypertendu avec le nouvel appareil. La corrélation mesure à quel point les patients sont bien ordonnés, et non à quel point les valeurs sont proches.

Mg2026 en fournit un exemple naturel. La PAS à l'inclusion et la PAS à 6 mois ont $r = 0,835$ (corrélation forte, IC à 95% 0,781-0,877), mais la moyenne a baissé de 131,1 à 126,4 mmHg. Les patients conservent leur place dans la hiérarchie, mais les valeurs ne sont pas les mêmes ; une corrélation élevée n'exclut pas une différence systématique.

La méthode de Bland-Altman, qui évalue la concordance, comporte trois étapes. Pour chaque patient, on calcule la différence entre les deux méthodes et leur moyenne. On calcule la moyenne des différences, appelée biais ou erreur systématique (angl. bias), et l'écart-type des différences. On calcule les limites de concordance (angl. limits of agreement) : la moyenne des différences $\pm 1,96 \times$ ÉT des différences. Environ 95% des différences individuelles se situent entre ces limites. Le graphique de Bland-Altman porte en abscisse la moyenne des deux méthodes et en ordonnée la différence, avec trois lignes horizontales (le biais et les deux limites).

Un court exemple numérique : huit patients mesurés avec un glucomètre portable et au laboratoire présentent les différences (glucomètre moins laboratoire) 4, -2, 6, 3, 8, 1, 5 et -1

mg/dL. La moyenne des différences est de 3,0 mg/dL et l'ÉT des différences de 3,46 mg/dL. Les limites de concordance sont $3{,}0 \pm 1{,}96 \times 3{,}46$, soit de -3,8 à 9,8 mg/dL. Si une différence allant jusqu'à 10 mg/dL est cliniquement acceptable, l'appareil peut être utilisé pour la surveillance ; sinon, non. La décision est clinique, et non statistique, et le coefficient de corrélation n'y intervient pas. Pour les variables catégorielles, la concordance se mesure par le coefficient kappa [103], présenté avec les tests diagnostiques.

9.7 La régression linéaire simple

9.7.1 Le modèle

La corrélation dit à quel point les points sont resserrés autour d'une droite. La régression trouve cette droite. Le modèle de régression linéaire simple écrit chaque valeur de Y comme une partie expliquée par X plus une erreur :

$$y = a + b\,x + \varepsilon$$

Dans cette écriture, a est l'ordonnée à l'origine (angl. intercept), c'est-à-dire la valeur prédite de Y lorsque X = 0 ; b est la pente (le coefficient de régression), c'est-à-dire la variation moyenne de Y pour une augmentation d'une unité de X ; ε est l'erreur résiduelle, la partie de Y que le modèle n'explique pas. Le signe de b est toujours le même que le signe de r. La droite proprement dite, $\hat{y} = a + b\,x$, donne la valeur moyenne prédite de Y pour une valeur donnée de X ; un patient particulier se trouvera au-dessus ou au-dessous, à une distance égale à son résidu, $e_i = y_i - \hat{y}_i$.

9.7.2 La méthode des moindres carrés

On peut tracer d'innombrables droites à travers le nuage de points. Laquelle est "la meilleure" ? La réponse classique, donnée par Gauss et Legendre au début du XIXe siècle, est la droite pour laquelle la somme des carrés des résidus est minimale :

$\text{SSE} = \sum e_i{}^2 = \sum (y_i - a - b\,x_i)^2 = \text{minimum}$

L'élévation au carré empêche les résidus positifs et négatifs de s'annuler et pénalise plus fortement les grands écarts. En résolvant ce problème de minimum, on obtient les formules :

$$b = \frac{S_{xy}}{S_{xx}} = r \times \left(\frac{s_y}{s_x}\right) \quad a = m_y - b \times m_x$$

La première forme de la pente sert au calcul à la main, la seconde lorsque nous ne disposons que de statistiques résumées (moyennes, écarts-types et r), situation fréquente à la lecture des articles. La seconde formule montre que la droite passe toujours par le point des moyennes (mx, my).

Pour les huit patients du tableau 9.1 : $b = \frac{302}{146} = 2{,}068$ mmHg pour chaque kg/m², et a = 131 - 2,068 x 29 = 71,01 mmHg. La droite est :

$$\text{PAS prédite} = 71{,}01 + 2{,}068 \times \text{IMC}$$

Pour un IMC de 30 kg/m², la PAS prédite est de 71,01 + 2,068 x 30 = 133,1 mmHg.

9.7.3 L'interprétation de la pente et de l'ordonnée à l'origine

La pente est le résultat clinique de la régression. "Pour chaque kg/m² supplémentaire, la PAS est, en moyenne, plus élevée de 2,07 mmHg" est une phrase que tout clinicien comprend. Attention aux unités : si X avait été exprimé dans d'autres unités (par exemple le poids en grammes), la pente aurait changé en conséquence, alors que la relation serait restée la même. Attention aussi au sens : il s'agit d'une différence entre patients, observée de façon transversale, et non de la

promesse qu'un patient qui perd 1 kg/m² verra sa pression baisser de 2 mmHg. Cette seconde affirmation serait une affirmation causale, que seule une étude d'intervention peut étayer.

L'ordonnée à l'origine a souvent une interprétation absurde. Dans notre exemple, 71 mmHg serait la pression "d'un patient avec un IMC nul", qui n'existe pas. L'ordonnée à l'origine est un paramètre mathématique nécessaire pour positionner la droite, mais elle n'a de sens clinique que si X = 0 se trouve dans le domaine des données (par exemple la dose nulle d'un médicament). Si vous voulez une ordonnée à l'origine interprétable, vous pouvez centrer la variable X (soustraire la moyenne de chaque valeur) ; l'ordonnée à l'origine devient alors la valeur prédite pour un patient dont X est égal à la moyenne.

9.7.4 L'exemple mg2026 : IMC et PAS

Sur les 160 patients de mg2026, la régression de la PAS en fonction de l'IMC donne les résultats du tableau 9.5, et la figure 9.4 montre le nuage de points, la droite et la bande de confiance de la droite.

Tableau 9.5. Régression linéaire simple de la PAS (mmHg) en fonction de l'IMC (kg/m²), mg2026, n = 160.

Coefficient	Estimation	Erreur standard	t	p	IC à 95%
Ordonnée à l'origine (a)	89,47	7,56	11,84	< 0,001	74,54-104,40
IMC (b)	1,435	0,258	5,56	< 0,001	0,925-1,944

$R^2 = 0{,}164$; erreur standard de l'estimation (SEE) = 12,73 mmHg ; F(1 ; 158) = 30,93, p < 0,001.

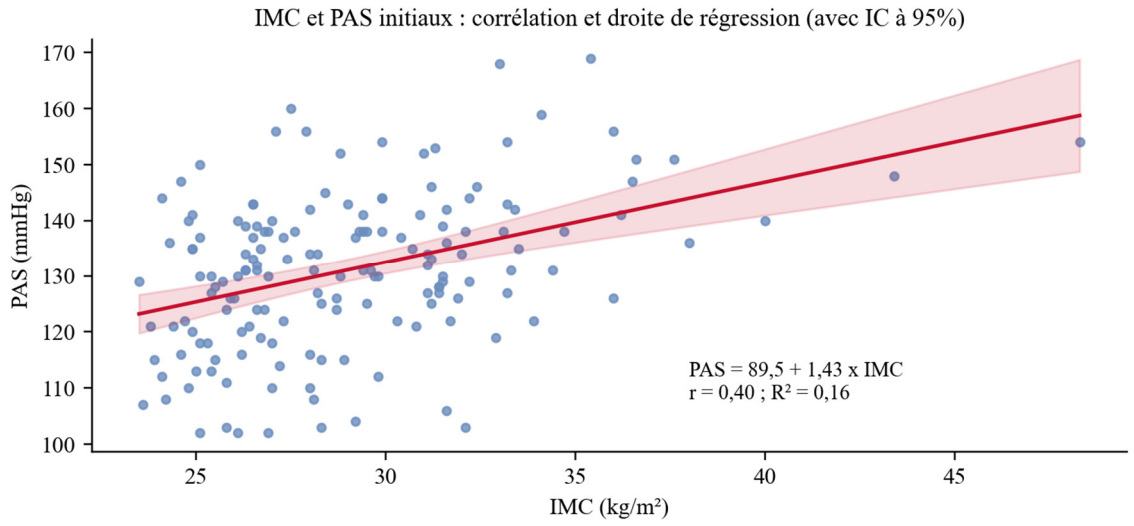


Figure 9.4. Indice de masse corporelle et pression artérielle systolique à l'inclusion dans le jeu de données mg2026 (n = 160) : chaque point est un patient, la ligne rouge est la droite de régression PAS = 89,47 + 1,435 x IMC et la bande colorée est l'intervalle de confiance à 95% de la valeur moyenne prédite. Données synthétiques (jeu de données mg2026).

L'équation est PAS = 89,47 + 1,435 x IMC. Pour chaque kg/m² supplémentaire, la PAS est en moyenne plus élevée de 1,4 mmHg ; l'intervalle de confiance de la pente (0,925-1,944) ne contient pas zéro, la pente est donc significativement différente de zéro. Ce n'est pas un hasard si le t de la pente (5,56) est exactement le t du test de r : en régression simple, le test de H0 : b = 0 et le test de H0 : ρ = 0 sont le même test, et $F = t^2 (5{,}56^2 = 30{,}93)$. Les prédictions sont : pour un IMC de 30, PAS = 89,47 + 1,435 x 30 = 132,5 mmHg ; pour un IMC de 35, PAS = 139,7 mmHg. La différence entre elles, 7,2 mmHg, vaut 5 x 1,435.

À titre de vérification, la pente peut aussi être obtenue à partir des statistiques résumées : $b = \frac{r \times s_y}{s_x} = \frac{0{,}405 \times 13{,}88}{3{,}91} = 1{,}44$, et a = 131,13 - 1,435 x 29,04 = 89,46 (les différences à la deuxième décimale proviennent des arrondis).

Pour l'âge, l'équation est PAS = 108,3 + 0,406 x âge ($R^2 = 0{,}066$) : pour chaque tranche de 10 ans supplémentaires, la PAS est en moyenne plus élevée d'environ 4 mmHg. Pour la glycémie et l'HbA1c : HbA1c = 3,38 + 0,0209 x glycémie, soit 0,21 point de pourcentage d'HbA1c pour chaque 10 mg/dL de glycémie.

9.7.5 L'erreur standard de l'estimation et R^2

Deux nombres décrivent la qualité d'ajustement de la droite. Le coefficient de détermination R^2 est une mesure relative : la proportion de la variation de Y expliquée par le modèle. Il peut s'écrire

$$R^2 = 1 - \frac{\mathrm{SSE}}{S_{\mathrm{yy}}}$$

c'est-à-dire 1 moins le rapport entre la variation restée inexpliquée après le tracé de la droite et la variation totale de Y autour de sa moyenne. Pour les huit patients, SSE = 59,3 et Syy = 684, donc $R^2 = 1 - \frac{59{,}3}{684} = 0{,}913$, exactement $r^2 = 0{,}956^2$.

L'erreur standard de l'estimation (SEE, angl. standard error of the estimate, appelée aussi écart-type résiduel) est une mesure absolue, exprimée dans les unités de Y :

$$\mathrm{SEE} = \sqrt{\left[\frac{\mathrm{SSE}}{n-2}\right]}$$

Elle indique à quelle distance de la droite se trouve typiquement un patient. Pour les huit patients, $\mathrm{SEE} = \sqrt{\frac{59{,}3}{6}} = 3{,}14$ mmHg. Pour mg2026, SEE = 12,73 mmHg, très proche de l'écart-type de la PAS (13,88 mmHg). La comparaison est instructive : si nous ne connaissions pas l'IMC, la meilleure estimation de la PAS serait la moyenne, avec une dispersion typique de 13,9 mmHg ; en connaissant l'IMC, la dispersion ne descend qu'à 12,7 mmHg. L'IMC explique une part réelle, mais faible, de la variation. À mesure que R^2 augmente, la SEE diminue, mais les deux disent des choses différentes : R^2 répond à la question "dans quelle mesure le modèle explique-t-il ?", et la SEE à la question "quelle est la précision d'une prédiction individuelle ?". Pour un clinicien qui veut estimer la pression artérielle d'un patient, la seconde question est souvent la plus importante.

9.7.6 L'intervalle de confiance de la pente

La pente estimée sur l'échantillon a sa propre erreur standard :

$$\mathrm{ES}(b) = \frac{\mathrm{SEE}}{\sqrt{S_{\mathrm{xx}}}}$$

et l'intervalle de confiance à 95% est $b \pm t(0{,}975;\ n-2) \times \mathrm{ES}(b)$. Pour les huit patients, $\mathrm{ES}(b) = \frac{3{,}14}{\sqrt{146}} = 0{,}260$, et le t pour 6 degrés de liberté vaut 2,447 ; $\mathrm{IC}\,95\% = 2{,}068 \pm 2{,}447 \times 0{,}260 = 2{,}068 \pm 0{,}637$, soit 1,43-2,71 mmHg par kg/m². Pour mg2026, ES(b) = 0,258 et, avec t ≈ 1,975 pour 158 degrés de liberté, $\mathrm{IC}\,95\% = 1{,}435 \pm 0{,}510 = 0{,}925 - 1{,}944$. L'intervalle est beaucoup plus utile que p : il nous dit que, dans la population dont provient l'échantillon, la différence de PAS associée à un kg/m² se situe plausiblement entre environ 0,9 et 1,9 mmHg.

9.7.7 Les résidus et la vérification des conditions

La régression linéaire repose sur quatre conditions, faciles à retenir grâce à leurs initiales en anglais (LINE) : la relation est linéaire (L), les observations sont indépendantes (I), les résidus suivent une distribution normale (N) et ont la même variance sur tout le domaine de X (E, de l'angl. equal variance, homoscédasticité). Remarquez que la normalité est exigée des résidus, et non de la variable X. X peut même être asymétrique, comme l'IMC dans mg2026.

Les conditions se vérifient sur les résidus. La figure 9.5 montre les deux graphiques standard pour le modèle PAS-IMC. Dans le panneau A, les résidus sont représentés en fonction des valeurs prédites : si le modèle est adéquat, les points forment une bande horizontale, sans courbure ni forme d'entonnoir, centrée sur zéro. Une courbe indiquerait une relation non linéaire ; un entonnoir (une dispersion qui augmente avec les valeurs prédites) indiquerait une hétéroscédasticité. Dans le panneau B, l'histogramme des résidus doit avoir approximativement la forme d'une cloche symétrique ; sur mg2026, le test de Shapiro-Wilk appliqué aux résidus donne p = 0,878, il n'y a donc pas de raison de rejeter la normalité.

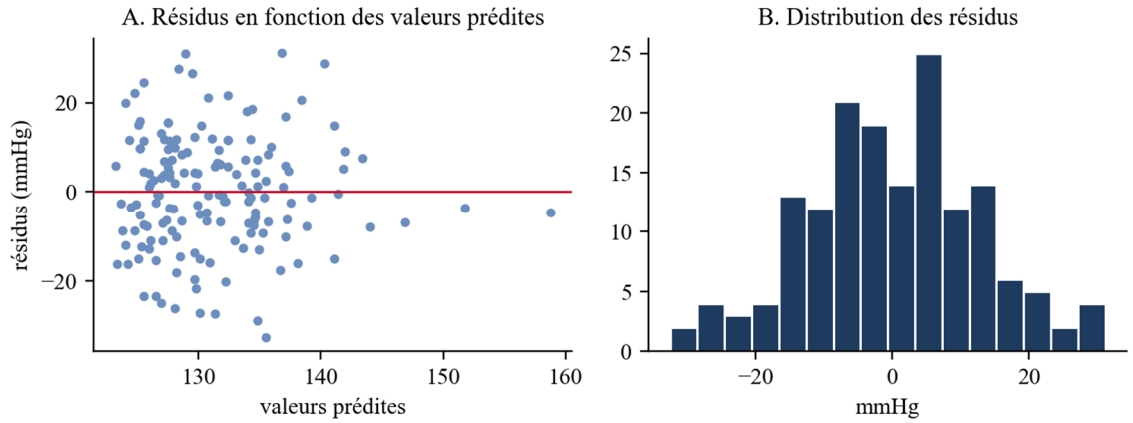


Figure 9.5. Analyse des résidus pour la régression PAS = 89,47 + 1,435 x IMC (mg2026). A : les résidus en fonction des valeurs prédites forment une bande horizontale, sans courbure ni forme d'entonnoir. B : l'histogramme des résidus est approximativement symétrique (Shapiro-Wilk p = 0,878). Données synthétiques (jeu de données mg2026).

L'indépendance ne se voit pas sur ces graphiques, mais dans le schéma de l'étude : chaque patient doit apparaître une seule fois. Si nous mettions dans la même régression la PAS à 0 et à 6 mois des mêmes patients, comme s'il s'agissait de 320 observations indépendantes, nous violerions cette condition et obtiendrions des valeurs p trop petites. Pour des données ordonnées dans le temps, les logiciels rapportent la statistique de Durbin-Watson (des valeurs proches de 2 indiquent l'absence d'autocorrélation des résidus ; sur mg2026, dans le modèle multiple de la section 9.8, la valeur est de 1,97).

9.7.8 La prédiction et le danger de l'extrapolation

Une fois la droite estimée, la prédiction semble simple : nous remplaçons X et obtenons Y. Deux subtilités séparent pourtant un bon étudiant d'un étudiant moyen.

La première concerne le type d'incertitude. La valeur de 132,5 mmHg prédite pour un IMC de 30 est une estimation de la pression moyenne de tous les patients ayant un IMC de 30. Pour cette moyenne, l'intervalle de confiance à 95% est étroit : 130,5-134,6 mmHg. Pour un seul nouveau patient avec un IMC de 30, l'incertitude est beaucoup plus grande, car s'y ajoute la variation individuelle (SEE = 12,7 mmHg) : l'intervalle de prédiction à 95% est de 107,3-157,7 mmHg. Autrement dit, le modèle connaît assez bien la moyenne, mais il ne peut pas dire si un patient donné ayant un IMC de 30 sera normotendu ou hypertendu. Pour un IMC de 35, l'intervalle de confiance de la moyenne est de 136,1-143,3 mmHg, et l'intervalle de prédiction de 114,3-165,1

mmHg. La bande de confiance de la figure 9.4 s'élargit vers les extrémités, car nous y disposons de moins de données.

La seconde subtilité est l'extrapolation. La droite a été estimée sur des patients dont l'IMC était compris entre 23,5 et 48,3 kg/m², et nous n'avons aucune garantie que la relation reste linéaire en dehors de cet intervalle. Les travaux pratiques de la discipline sur la taille et le poids demandent, à dessein, quel poids devrait avoir une personne mesurant 1,78 m et une autre mesurant 2,35 m. Sur mg2026, la droite est poids = -100,40 + 1,087 x taille (cm). Pour 178 cm, le poids prédit est de -100,40 + 1,087 x 178 = 93,1 kg, une prédiction légitime, car 178 cm se trouve dans le domaine des données (145-192 cm). Pour 235 cm, la formule donne 155,0 kg. Ce nombre se calcule tout aussi facilement, mais il ne s'appuie sur aucune donnée : personne dans l'échantillon ne dépasse 192 cm, et les personnes de plus de 2,30 m présentent souvent des affections (gigantisme, acromégalie) qui modifient complètement la relation. La figure 9.6 rend le problème visible : la ligne pointillée rouge est une extrapolation.

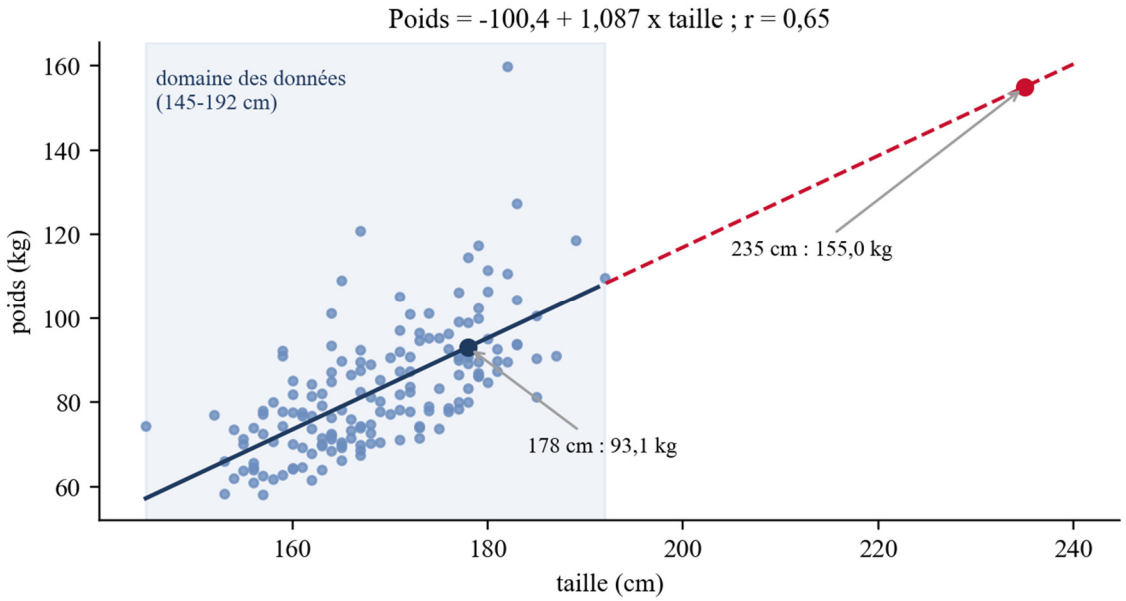


Figure 9.6. La taille et le poids dans le jeu de données mg2026 (n = 160), avec la droite de régression poids = -100,4 + 1,087 x taille (r = 0,65). La zone colorée correspond au domaine des données (145-192 cm). La prédiction pour 178 cm (93,1 kg) est une interpolation ; celle pour 235 cm (155,0 kg) est une extrapolation qui ne s'appuie pas sur les données. Données synthétiques (jeu de données mg2026).

Un exemple encore plus clair est celui de l'IMC et de la PAS : si nous introduisons un IMC de 60 kg/m², la formule donne 175,6 mmHg. C'est peut-être vrai, peut-être pas ; nos données ne nous disent rien sur les patients atteints d'obésité extrême. À l'examen, la question "quelle valeur le modèle prédit-il pour..." doit presque toujours s'accompagner d'un commentaire sur le domaine dans lequel le modèle est valable.

9.8 La régression linéaire multiple : introduction

9.8.1 Pourquoi plusieurs prédicteurs sont nécessaires

La PAS ne dépend pas seulement de l'IMC. L'âge, le sexe, le tabagisme, la consommation de sel et bien d'autres facteurs y contribuent. De plus, les prédicteurs sont liés entre eux : les patients plus âgés peuvent aussi avoir un IMC plus élevé, et une partie de l'association "brute" entre l'IMC et la PAS pourrait en réalité provenir de l'âge. La régression linéaire multiple estime l'effet de chaque prédicteur en maintenant constants les autres prédicteurs du modèle :

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + \varepsilon$$

Chaque coefficient bj se lit ainsi : "la variation moyenne de Y pour une augmentation d'une unité de xj, chez des patients ayant les mêmes valeurs des autres variables du modèle". L'expression "ajusté sur" dans les articles signifie exactement cela. Le calcul se fait uniquement avec le logiciel, mais tout médecin qui lit la littérature doit savoir en lire les résultats.

9.8.2 Les variables catégorielles en régression

Le sexe et le tabagisme ne sont pas numériques. Ils entrent dans le modèle par des variables indicatrices (angl. dummy variables), codées 0/1. Pour le sexe, une seule variable suffit (1 = homme, 0 = femme) ; son coefficient est la différence moyenne entre les hommes et les femmes, à valeurs égales des autres prédicteurs. Le tabagisme a trois catégories et nécessite donc deux variables indicatrices ; la catégorie qui ne reçoit pas d'indicatrice devient la catégorie de référence, à laquelle les autres sont comparées. Dans le modèle ci-dessous, la référence est "ancien fumeur". Le choix de la référence ne change pas le modèle, mais seulement les comparaisons affichées. On choisit habituellement la catégorie la plus nombreuse ou la plus naturelle sur le plan clinique (ici, cela aurait pu être les non-fumeurs).

9.8.3 L'exemple mg2026 : la PAS en fonction de l'IMC, de l'âge, du sexe et du tabagisme

Tableau 9.6. Régression linéaire multiple de la PAS à l'inclusion (mmHg), mg2026, n = 160.

Prédicteur	Coefficient (B)	Erreur standard	IC à 95%	p
Ordonnée à l'origine	64,51	9,39	45,96-83,06	< 0,001
IMC $\left(\frac{\text{kg}}{m^2}\right)$	1,389	0,240	0,915-1,862	< 0,001
Âge (années)	0,449	0,108	0,235-0,662	0,0001
Sexe masculin (par rapport au sexe féminin)	-1,481	1,878	-5,19 à 2,23	0,43
Fumeur (par rapport à ancien fumeur)	6,738	2,636	1,53-11,95	0,012
Non-fumeur (par rapport à ancien fumeur)	-1,119	2,487	-6,03 à 3,79	0,65

$R^2 = 0{,}308$; R^2 ajusté $= 0{,}286$; F(5 ; 154) = 13,72, p < 0,001.

Lisons le tableau ligne par ligne. À âge, sexe et statut tabagique identiques, chaque kg/m² supplémentaire est associé à une PAS plus élevée de 1,39 mmHg, très proche de la pente brute de 1,435. L'âge n'"absorbe" pas l'effet de l'IMC, ce qui signifie que, dans mg2026, l'âge n'est pas un facteur de confusion important pour la relation IMC-PAS. Chaque année d'âge ajoute, à IMC égal, 0,45 mmHg (4,5 mmHg par décennie). Les hommes ont une PAS inférieure de 1,5 mmHg à celle des femmes, mais l'intervalle (-5,2 à 2,2) contient zéro et p = 0,43 : nous n'avons pas de preuve d'une différence entre les sexes, en accord avec le test t du chapitre 6. Les fumeurs actuels ont une PAS supérieure de 6,7 mmHg à celle des anciens fumeurs (IC 1,5-11,9 ; p = 0,012), tandis que les non-fumeurs ne diffèrent pas significativement des anciens fumeurs.

Avec un modèle estimé, nous pouvons aussi faire des prédictions pour un patient concret. Un homme fumeur de 60 ans, avec un IMC de 32, a une PAS prédite = 64,51 + 1,389 x 32 + 0,449 x 60 - 1,481 + 6,738 = 141,2 mmHg. Une femme ancienne fumeuse du même âge et avec le même IMC a 64,51 + 1,389 x 32 + 0,449 x 60 = 135,9 mmHg.

9.8.4 Le R^2 ajusté

R^2 augmente chaque fois que nous ajoutons un prédicteur, même un prédicteur sans aucun lien avec Y (par exemple le dernier chiffre du code numérique personnel, le CNP roumain), car toute

variable "capte" une petite partie de la variation aléatoire. Le R² ajusté pénalise le nombre de prédicteurs :

$$R^2 \text{ ajusté } = 1 - \frac{(1 - R^2) \times (n - 1)}{n - k - 1}$$

où k est le nombre de prédicteurs (ici 5, car le tabagisme compte deux fois). Pour mg2026 : $1 - \frac{(1-0{,}308)\times 159}{154} = 1 - 0{,}692 \times 1{,}0325 = 0{,}286$. La faible différence entre R² et R² ajusté montre que le modèle ne comporte pas de prédicteurs en excès par rapport à la taille de l'échantillon. Une règle pratique fréquemment citée demande au moins 10-15 observations par prédicteur ; avec 160 patients et 5 prédicteurs, nous sommes loin de la limite.

Le modèle multiple explique 30,8% de la variation de la PAS, presque le double de ce qu'explique l'IMC seul (16,4%). Il reste toutefois près de 70% de variation inexpliquée. En recherche clinique, des valeurs de R² de 0,2-0,4 pour des variables biologiques sont habituelles ; cela ne signifie pas que le modèle est "mauvais", mais que la pression artérielle a de nombreuses causes que nous n'avons pas mesurées.

9.8.5 Coefficients non standardisés et standardisés

Les articles rapportent parfois le coefficient non standardisé B (dans les unités de Y par unité de X, comme dans le tableau 9.6), parfois le coefficient standardisé β (bêta), obtenu après transformation de toutes les variables en scores z. Bêta se lit "de combien d'écarts-types Y varie pour un écart-type de X" et permet de comparer des prédicteurs mesurés dans des unités différentes. Malheureusement, la notation n'est pas uniforme : certains auteurs notent β le coefficient non standardisé lui-même. Lorsque vous lisez un article, vérifiez dans la section des méthodes ou d'après les unités (un β exprimé en cm ou en kg/m² est non standardisé).

9.8.6 La colinéarité

Si deux prédicteurs sont très fortement corrélés entre eux, le modèle ne peut pas séparer leurs effets, et les coefficients deviennent instables, avec de grandes erreurs standard. Sur mg2026, l'IMC et le tour de taille ont r = 0,885 : tous deux mesurent l'adiposité. Les introduire simultanément dans un modèle de la PAS produirait des coefficients difficiles à interpréter. Les logiciels rapportent le facteur d'inflation de la variance (VIF, angl. variance inflation factor) ; des valeurs supérieures à 5-10 signalent une colinéarité.

9.8.7 Comment lire un tableau de régression dans un article

Pour un tableau de régression publié dans un article, voici un ordre de lecture utile. Identifiez la variable dépendante et le type de régression (linéaire pour un Y numérique, logistique pour un Y binaire, de Cox pour un temps jusqu'à l'événement). Vérifiez combien d'observations sont effectivement entrées dans le modèle (les patients ayant des données manquantes sont habituellement exclus) et combien de prédicteurs il comporte. Lisez le coefficient de chaque prédicteur d'intérêt avec son unité de mesure et son intervalle de confiance ; un coefficient sans unité ne peut pas être interprété. Cherchez les catégories de référence des variables catégorielles. Regardez R² (ou R² ajusté) pour juger de ce qu'explique le modèle. Enfin, demandez-vous quelles variables manquent : un facteur de confusion omis peut faire paraître significatif un coefficient. Les recommandations de rédaction pour les études observationnelles exigent explicitement la description des variables d'ajustement et de la manière dont elles ont été choisies [25].

> **Étude de cas : la régression multiple dans les articles des collègues**
>
> Une étude transversale portant sur 220 adultes atteints de maladies dermatologiques chroniques a analysé l'observance thérapeutique (score MARS-5, 21,50 ± 4,19) en fonction de la satisfaction à l'égard des services pharmaceutiques et du soutien social [72]. Le résumé

rapporte un coefficient non standardisé B = 0,076 (IC à 95% 0,017-0,135 ; p = 0,011) et une augmentation de la variance expliquée de 13,3% à 28,0% après l'ajout d'autres variables au modèle ; une association est arrivée à la limite de la signification (p = 0,053) après ajustement. La leçon : R^2 augmente avec l'ajout de prédicteurs pertinents, et un coefficient peut s'atténuer lorsque nous introduisons une variable médiatrice ou un facteur de confusion de la relation. L'interprétation de B exige de connaître l'échelle du prédicteur : un coefficient de 0,076 paraît petit, mais si le prédicteur est un score de grande amplitude (le score de satisfaction PSPSQ-2 vaut 62,92 ± 9,57), la différence entre des patients situés aux deux extrémités de l'échelle peut devenir pertinente. Sans unité, le coefficient ne dit rien ; avec son unité, il devient une grandeur clinique.

Une étude transversale multivariée portant sur 245 personnes a comparé l'estime de soi chez les porteurs d'appareils auditifs intelligents, chez les non-porteurs et chez des personnes ayant une audition normale [82]. Outre l'ANOVA (F = 299,00, p < 0,001, η^2 = 0,712), les auteurs rapportent des corrélations de Pearson : l'estime de soi est corrélée négativement au handicap auditif (r = -0,573, p < 0,001) et aux acouphènes (r = -0,443, p < 0,001), et les deux échelles sont corrélées positivement entre elles (r = 0,729, p < 0,001). Le modèle de régression multiple, avec un coefficient d'environ 11,9 pour le port d'un appareil, explique environ 65% de la variation de l'estime de soi. Un R^2 aussi élevé est inhabituel en recherche clinique et s'explique par la très grande différence entre les groupes. La corrélation de 0,729 entre deux prédicteurs potentiels signale aussi un risque de colinéarité, si tous deux entrent dans le même modèle.

En cardiologie, une étude cas-témoins portant sur 60 patients atteints d'insuffisance cardiaque ischémique et 60 témoins du même âge a rapporté des valeurs d'ostéoprotégérine plus élevées chez les malades (4,7 ± 0,25 contre $1{,}3 \pm 0{,}67$ ng/mL) et des valeurs de vitamine D plus basses (20,49 ± 7,31 contre $37{,}09 \pm 4{,}59$ ng/mL), puis des corrélations de Pearson très fortes des marqueurs étudiés avec le volume télédiastolique (r = 0,862), avec la fraction d'éjection (r = -0,832) et avec la vitesse de l'onde de pouls (r = 0,833), toutes avec p < 0,001, une corrélation modérée avec la vitamine D (r = -0,636) et une corrélation faible avec le débit de filtration glomérulaire (r = -0,488) [87]. Les auteurs ont poursuivi avec une régression multiple dans laquelle la vitamine D était la variable dépendante. Remarquez que, comme l'étude inclut à la fois des malades et des sujets sains, une partie des corrélations fortes peut provenir de la différence entre les deux groupes : à l'intérieur de chaque groupe, la relation pourrait être beaucoup plus faible (la situation des deux sous-groupes de la section 9.1.2).

Une étude de cohorte rétrospective portant sur 186 patients après un infarctus aigu du myocarde a utilisé la régression linéaire multiple pour étudier le remodelage ventriculaire en fonction du nombre de classes du traitement recommandé par les lignes directrices (GDMT) [109]. Les résultats sont donnés sous la forme β = 0,120 cm (p = 0,004) et β = 0,204 cm (p < 0,001) pour chaque classe supplémentaire. L'unité (cm) montre qu'il s'agit de coefficients non standardisés, interprétés directement comme une différence de diamètre pour chaque classe de traitement supplémentaire. Seuls 18,8% des patients ont reçu les quatre piliers thérapeutiques, ce qui limite la précision à l'extrémité supérieure de l'échelle.

Deux études de l'équipe de nutrition illustrent la régression linéaire en épidémiologie nutritionnelle. Dans une cohorte prospective portant sur 921 étudiants suivis pendant leur première année d'université, la variation moyenne de l'IMC a été de -0,45 ± 1,07 kg/m² (IC à 95% -0,52 à -0,38), et un comportement de chrononutrition a été associé à une différence ajustée de $\beta = +\frac{0{,}22\ \text{kg}}{m^2}\ (p = 0{,}004)$ [5]. Dans une étude transversale portant sur 693 étudiants, les profils alimentaires ont été associés à l'HbA1c : β = 0,147 (p < 0,001) pour un profil et β = -0,109 (p < 0,001) pour un autre, de sens opposé [6]. Les deux résumés rapportent des

coefficients avec leur valeur p, sans intervalle de confiance ; pour un lecteur, l'intervalle aurait montré avec quelle précision l'effet est estimé.

Enfin, la régression peut servir à construire un score clinique. Dans une étude portant sur 124 patients atteints de cirrhose compensée, les auteurs ont combiné la rigidité splénique (SS), la rigidité hépatique (LS) et la taille de la rate (SSZ) en un score obtenu par régression multiple : 0,053 x SS + 0,054 x LS + 0,059 x SSZ - 1,84, avec un seuil > 0,34 et une aire sous la courbe ROC de 0,93, supérieure à celle de chaque paramètre pris séparément (0,84 ; 0,86 ; 0,71) [110]. La formule a exactement la structure b0 + b1 x1 + b2 x2 + b3 x3 de la section 9.8.1. Sa validité chez d'autres patients que ceux sur lesquels elle a été construite doit être vérifiée séparément, dans une cohorte de validation.

9.9 La régression logistique : une passerelle vers le chapitre 10

9.9.1 Quand Y n'a que deux valeurs

Que faire lorsque la variable de résultat est binaire : diabète oui/non, décès oui/non, perte de poids d'au moins 5% oui/non ? La régression linéaire ne convient pas : elle pourrait prédire des "probabilités" négatives ou supérieures à 1, et les résidus ne peuvent pas être normaux. La régression logistique ne modélise pas directement la probabilité p, mais le logarithme de la cote (angl. log odds ou logit) :

$$\ln\left[\frac{p}{1-p}\right] = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

La cote (angl. odds) est le rapport entre la probabilité de présenter l'événement et la probabilité de ne pas le présenter ; une probabilité de 20% correspond à une cote de 0,2 / 0,8 = 0,25, c'est-à-dire "1 contre 4". Le chapitre 10 explique cette notion en détail. Sur le plan pratique, il faut retenir dès maintenant que l'exponentielle d'un coefficient, e^b, est un rapport de cotes (angl. odds ratio, OR) : le nombre de fois par lequel la cote de l'événement est multipliée lorsque le prédicteur augmente d'une unité, à valeurs constantes des autres prédicteurs.

9.9.2 L'exemple mg2026 : le diabète confirmé par HGPO

Tableau 9.7. Régression logistique du diabète confirmé par HGPO (29 cas sur 160), mg2026.

Prédicteur	OR ajusté	IC à 95%	p
Obésité (IMC ≥ 30), par rapport à IMC < 30	6,01	2,35-15,39	0,0002
Âge, pour chaque année	1,11	1,05-1,18	0,0003
Sexe masculin, par rapport au sexe féminin	0,97	0,40-2,38	0,95

Lecture : à âge et sexe identiques, la cote d'avoir un diabète est environ 6 fois plus élevée chez les patients obèses que chez les non-obèses ; l'intervalle large (2,35-15,39) reflète le petit nombre de cas (29). Chaque année d'âge multiplie la cote par 1,11, soit une augmentation de 11% par an, à statut pondéral et sexe identiques. Pour 10 ans, les effets se multiplient, ils ne s'additionnent pas : 1,11^10 ≈ 2,8, donc un patient plus âgé de 10 ans a une cote presque trois fois plus élevée. Le sexe ne joue pas de rôle (OR 0,97, intervalle contenant 1, p = 0,95). Remarquez la règle qui reviendra au chapitre 10 : pour les rapports (OR, RR), la valeur "sans effet" est 1, et non 0, et un intervalle qui contient 1 correspond à un résultat non significatif.

Au chapitre 10, nous calculerons à la main, à partir du tableau 2 x 2, l'OR brut entre l'obésité et le diabète : 4,32 (IC à 95% 1,86-10,03). L'OR ajusté sur l'âge et le sexe (6,01) est plus élevé que l'OR brut. Cette différence ne signifie pas une erreur : l'ajustement sur un facteur fortement lié au résultat (l'âge) peut modifier l'OR, même si ce facteur est faiblement lié à l'exposition. Pour

l'instant, il suffit de retenir que l'OR brut et l'OR ajusté répondent à des questions différentes et qu'un bon article rapporte les deux.

9.9.3 La même question, deux modèles

L'étude prospective portant sur 921 étudiants citée plus haut montre comment le choix entre régression linéaire et régression logistique dépend de la forme sous laquelle nous traitons la variable de résultat [5]. La variation de l'IMC, en tant que variable continue, a été analysée par régression linéaire $\left(\beta = +\frac{0{,}22\ \text{kg}}{m^2}\right)$. La même variation, dichotomisée en "prise de poids cliniquement significative" (présente chez 7,2% des étudiants), a été analysée par régression logistique. La première approche utilise toute l'information ; la seconde répond à une question clinique plus directe (qui arrive à une prise de poids importante ?), mais perd de l'information en découpant la variable continue. Une étude transversale portant sur 2976 étudiants du département de Timiș a utilisé la régression logistique pour identifier les facteurs associés au tabagisme intensif [22] ; les prédicteurs ainsi identifiés sont des facteurs associés, et non des causes démontrées, une discussion que nous poursuivons au chapitre 10.

9.10 Problèmes résolus

Problème résolu 9.1. Pearson et régression, à la main, sur huit patients

Énoncé. Pour les huit patients du tableau 9.1 (IMC et PAS) : (a) calculez le coefficient de Pearson et interprétez-le selon Colton ; (b) testez sa signification à α = 0,05 et calculez l'IC à 95% de ρ ; (c) calculez R² ; (d) estimez la droite de régression de la PAS en fonction de l'IMC ; (e) calculez la SEE et l'IC à 95% de la pente ; (f) prédisez la PAS pour un IMC de 30 et commentez une prédiction pour un IMC de 45.

Solution. (a) D'après le tableau 9.1 : Sxy = 302, Sxx = 146, Syy = 684. $r = \frac{302}{\sqrt{146 \times 684}} = 0{,}956$. Selon Colton, |r| ≥ 0,75 : corrélation directe forte.

(b) Hypothèses : H0 : ρ = 0 ; H1 : ρ ≠ 0. $t = \frac{0{,}956 \times \sqrt{6}}{\sqrt{1 - 0{,}9133}} = \frac{2{,}342}{0{,}2945} = 7{,}95$; ddl = 6 ; t critique = 2,447. Comme 7,95 > 2,447, on rejette H0 (p = 0,0002). IC à 95% par la transformation de Fisher : $z = 0{,}5 \times \ln\left(\frac{1{,}956}{0{,}044}\right) = 1{,}893$; ES $= \frac{1}{\sqrt{5}} = 0{,}447$; $z \pm 0{,}877$ donne 1,016 et 2,770 ; après retransformation : 0,769-0,992.

(c) $R^2 = 0{,}956^2 = 0{,}913$. Environ 91% de la variation de la PAS dans ce petit groupe est associée linéairement à l'IMC.

(d) $b = \frac{302}{146} = 2{,}068$; a = 131 - 2,068 x 29 = 71,01. Équation : PAS = 71,01 + 2,068 x IMC.

(e) Les valeurs prédites sont 118,6 ; 122,7 ; 124,8 ; 128,9 ; 131,0 ; 135,1 ; 139,3 et 147,5, et les résidus -0,6 ; 3,3 ; -3,8 ; 0,1 ; 4,0 ; -4,1 ; 0,7 et 0,5. La somme des résidus est nulle (une vérification utile). SSE = 59,3 ; SEE $= \sqrt{\frac{59{,}3}{6}} = 3{,}14$ mmHg. ES$(b) = \frac{3{,}14}{\sqrt{146}} = 0{,}260$; IC 95% $= 2{,}068 \pm 2{,}447 \times 0{,}260 = 1{,}43 - 2{,}71$.

(f) PAS$(30) = 71{,}01 + 2{,}068 \times 30 = 133{,}1$ mmHg. Pour un IMC de 45, la formule donne 164,1 mmHg, mais la valeur 45 se situe en dehors du domaine observé (23-37 kg/m²) : c'est une extrapolation, et avec seulement huit patients nous n'avons aucune raison de croire que la relation y reste linéaire.

Interprétation clinique : dans ce groupe, les patients dont l'IMC est plus élevé ont une pression artérielle plus élevée, d'environ 2 mmHg par kg/m². Le résultat est significatif, mais l'échantillon est très petit, et les intervalles de ρ (0,77-0,99) et de la pente (1,4-2,7) sont larges.

Problème résolu 9.2. Spearman avec des valeurs égales

Énoncé. Huit patients ont autoévalué leur activité physique sur une échelle de 1 (sédentaire) à 5 (très actif) et leur IMC a été mesuré : activité 1, 2, 2, 3, 3, 3, 4, 5 ; IMC 33,5 ; 31,0 ; 29,4 ; 30,2 ; 27,8 ; 28,6 ; 26,1 ; 24,9 kg/m². Calculez le coefficient de Spearman et comparez le résultat de la formule abrégée avec le résultat exact.

Solution. L'activité est ordinale, nous choisissons donc le coefficient de Spearman. Rangs de l'activité : la valeur 1 reçoit le rang 1 ; les deux valeurs 2 occuperaient les places 2 et 3, elles reçoivent donc 2,5 chacune ; les trois valeurs 3 occuperaient les places 4, 5 et 6, elles reçoivent donc 5 chacune ; 4 reçoit 7 et 5 reçoit 8. Rangs de l'IMC (dans l'ordre croissant) : 24,9 reçoit 1 ; 26,1 reçoit 2 ; 27,8 reçoit 3 ; 28,6 reçoit 4 ; 29,4 reçoit 5 ; 30,2 reçoit 6 ; 31,0 reçoit 7 ; 33,5 reçoit 8. Par patient, dans l'ordre de l'énoncé : les rangs de l'activité sont 1 ; 2,5 ; 2,5 ; 5 ; 5 ; 5 ; 7 ; 8, et les rangs de l'IMC 8 ; 7 ; 5 ; 6 ; 3 ; 4 ; 2 ; 1.

Les différences di sont -7 ; -4,5 ; -2,5 ; -1 ; 2 ; 1 ; 5 ; 7, et $\sum {d_i}^2 = 49 + 20{,}25 + 6{,}25 + 1 + 4 + 1 + 25 + 49 = 155{,}5$. Avec la formule abrégée : $r_s = 1 - \frac{6 \times 155{,}5}{8 \times 63} = 1 - \frac{933}{504} = 1 - 1{,}851 = -0{,}851$.

Le calcul exact (Pearson sur les rangs, comme le fait n'importe quel logiciel) donne rs = -0,908, p = 0,002. La différence (-0,85 contre -0,91) provient des cinq valeurs égales de l'activité, qui réduisent la variation des rangs ; la formule abrégée suppose que les rangs sont 1, 2,..., n sans répétition. Conclusion : corrélation inverse forte dans ce petit groupe, les patients plus actifs ont un IMC plus faible. Sur mg2026, la même analyse portant sur 160 patients a donné rs = -0,012, pratiquement zéro. L'exemple à huit patients a été construit pour le calcul, et ce contraste nous rappelle qu'un petit échantillon peut produire des corrélations spectaculaires.

Problème résolu 9.3. La taille et le poids, au format des devoirs de la discipline

Énoncé. Sur le jeu de données mg2026, étudiez l'association entre la taille et le poids. Répondez au format a-g et précisez : quelle proportion de la variation du poids est due à des facteurs autres que la taille ? Quel poids le modèle prédit-il pour une personne de 1,78 m et pour une personne de 2,35 m ?

(a) Hypothèse de recherche : les personnes plus grandes ont, en moyenne, un poids plus élevé.

(b) Hypothèses statistiques : H0 : ρ = 0, la taille et le poids ne sont pas corrélés linéairement dans la population ; H1 : ρ ≠ 0, les deux variables sont corrélées.

(c) Méthode : deux variables numériques mesurées chez les mêmes patients, donc une analyse bivariée de corrélation et de régression. La taille est approximativement normale (Shapiro-Wilk p = 0,078) ; le poids est asymétrique (asymétrie 1,29 ; p < 0,001), c'est pourquoi nous rapportons Pearson et, à titre de vérification, Spearman. Le prédicteur X est la taille, la réponse Y est le poids.

(d) Résolution : JASP (Regression > Correlation et Regression > Linear Regression) ou Google Sheets (=CORREL, =SLOPE, =INTERCEPT, =RSQ).

(e) Résultats : r = 0,652 (IC à 95% 0,552-0,733), t = 10,79, ddl 158, p < 0,001 ; Spearman rs = 0,693, p < 0,001. Droite : poids = -100,40 + 1,087 x taille (cm) ; IC à 95% de la pente 0,888-1,286 kg/cm ; $R^2 = 0{,}424$; SEE = 11,49 kg. Le graphique est la figure 9.6.

(f) Conclusion statistique : p < 0,001, on rejette H0 ; la corrélation est positive et modérée selon Colton (0,50 ≤ r < 0,75).

(g) Conclusion de recherche : les personnes plus grandes pèsent davantage, en moyenne 1,09 kg de plus pour chaque centimètre. $R^2 = 0{,}424$, donc 42,4% de la variation du poids est associée à la taille, et 1 - 0,424 = 0,576, soit 57,6%, relève d'autres facteurs (sexe, composition corporelle, alimentation, activité). Pour 178 cm, le poids prédit est de -100,40 + 1,087 x 178 = 93,1 kg (intervalle de prédiction à 95% pour une personne : 70,2-115,9 kg). Pour 235 cm, la formule donne 155,0 kg, mais il s'agit d'une extrapolation : la plus grande taille de l'échantillon est de 192 cm. La réponse correcte à la seconde question est que le modèle ne peut pas fournir une prédiction crédible.

Problème résolu 9.4. Vérification de corrélations publiées

Énoncé. (a) Une étude portant sur 867 patients rapporte r = -0,088 et p = 0,01 [55]. Vérifiez la valeur p. (b) Une étude portant sur 121 étudiants rapporte R = 0,107, p = 0,246 [16]. Vérifiez. (c) Une étude portant sur 157 patients rapporte ρ = 0,826 avec un IC à 95% de 0,769-0,870 [107]. Vérifiez l'intervalle par la transformation de Fisher.

Solution. $(a)\ t = \frac{0{,}088 \times \sqrt{865}}{\sqrt{1 - 0{,}088^2}} = \frac{2{,}588}{0{,}996} = 2{,}60$ (en valeur absolue) ; avec 865 degrés de liberté, p = 0,0095, soit environ 0,01. Le résultat est confirmé ; la corrélation est significative, mais négligeable ($R^2 = 0{,}8\%$).

(b) $t = \frac{0{,}107 \times \sqrt{119}}{\sqrt{1 - 0{,}0114}} = \frac{1{,}167}{0{,}994} = 1{,}17$; avec 119 degrés de liberté, p ≈ 0,24, très proche de la valeur rapportée de 0,246 (la différence provient de l'arrondi de R). On ne rejette pas H0.

(c) $z = 0{,}5 \times \ln\left(\frac{1{,}826}{0{,}174}\right) = 0{,}5 \times \ln(10{,}494) = 1{,}175$. ES $= \frac{1}{\sqrt{157-3}} = \frac{1}{\sqrt{154}} = 0{,}0806$. $z \pm 1{,}96 \times 0{,}0806 = 1{,}175 \pm 0{,}158$, soit 1,017 et 1,333. Nous retransformons avec $r = \frac{e^{2z}-1}{e^{2z}+1}$. Pour z = 1,017 : $e^{2{,}034} = 7{,}64$, donc $r = \frac{6{,}64}{8{,}64} = 0{,}769$. Pour z = 1,333 : $e^{2{,}666} = 14{,}38$, donc $r = \frac{13{,}38}{15{,}38} = 0{,}870$. L'intervalle publié est reproduit exactement. La leçon : lorsqu'un article donne r et n, vous pouvez reconstruire vous-mêmes l'intervalle de confiance.

Problème résolu 9.5. La droite de régression à partir de statistiques résumées

Énoncé. Un article rapporte pour le tour de taille (X) et la PAS (Y) : mx = 88,5 cm, sx = 12,1 cm, my = 131,1 mmHg, sy = 13,9 mmHg, r = 0,32, n = 160. (a) Écrivez la droite de régression. (b) Prédisez la PAS pour un tour de taille de 100 cm. (c) Quelle proportion de la variation de la PAS est associée au tour de taille ? (d) La corrélation est-elle significative ?

Solution. $(a)\ b = \frac{r \times s_y}{s_x} = \frac{0{,}32 \times 13{,}9}{12{,}1} = 0{,}368$ mmHg/cm ; a = 131,1 - 0,368 x 88,5 = 98,6 mmHg. PAS = 98,6 + 0,368 x tour de taille.

(b) $\text{PAS}(100) = 98{,}6 + 0{,}368 \times 100 = 135{,}3$ mmHg. La valeur de 100 cm se situe dans le domaine habituel du tour de taille chez les adultes en surpoids ; la prédiction est donc une interpolation.

(c) $R^2 = 0{,}32^2 = 0{,}102$; environ 10% de la variation de la PAS est associée linéairement au tour de taille, et 90% relève d'autres facteurs.

(d) $t = \frac{0{,}32 \times \sqrt{158}}{\sqrt{1 - 0{,}1024}} = \frac{4{,}022}{0{,}947} = 4{,}25$; ddl 158 ; p < 0,001. La corrélation est significative et faible selon Colton. Sur mg2026, le calcul direct donne r = 0,322 et PAS = 98,49 + 0,369 x tour de taille, très proche du résultat obtenu à partir des statistiques arrondies.

Problème résolu 9.6. Lecture d'un tableau de régression multiple

Énoncé. À l'aide du tableau 9.6, répondez : (a) Quelle est la différence moyenne de PAS entre deux patients de même sexe, de même statut tabagique et de même IMC, dont l'un a 50 ans et l'autre 70 ? (b) Le sexe est-il un prédicteur significatif ? (c) Pourquoi le coefficient de l'IMC (1,389) diffère-t-il légèrement de la pente de la régression simple (1,435) ? (d) Interprétez le R² ajusté.

Solution. (a) 20 ans $\frac{x\ 0{,}449\ \text{mmHg}}{\text{an}} = 9{,}0$ mmHg ; le patient de 70 ans a, en moyenne, une PAS plus élevée d'environ 9 mmHg. L'intervalle pour 20 ans s'obtient en multipliant les bornes : 20 x 0,235 = 4,7 et 20 x 0,662 = 13,2 mmHg.

(b) Non : le coefficient -1,481 a un IC à 95% allant de -5,19 à 2,23, qui contient 0, et p = 0,43. Nous n'avons pas de preuve que les hommes et les femmes diffèrent quant à la PAS après ajustement sur les autres facteurs.

(c) En régression simple, la pente de l'IMC capte aussi la part de l'association qui se superpose à d'autres facteurs corrélés à l'IMC (âge, tabagisme). Après ajustement, le coefficient se rapporte à des patients comparables en âge, en sexe et en tabagisme. Le changement très faible (de 1,435 à 1,389) montre que ces facteurs n'exercent pas d'effet de confusion important sur la relation IMC-PAS dans mg2026.

(d) R^2 ajusté $= 0{,}286$: après pénalisation pour les 5 prédicteurs, le modèle explique environ 29% de la variation de la PAS. La faible différence avec R² (0,308) montre qu'il n'y a pas de prédicteurs inutiles en excès.

9.11 Exercices proposés

1. Chez six patients, on a mesuré l'âge (années) et la PAS (mmHg) : (35 ; 118), (42 ; 121), (50 ; 130), (58 ; 128), (63 ; 140), (70 ; 145). Calculez r à la main, testez-le à α = 0,05, écrivez la droite de régression de la PAS en fonction de l'âge et prédisez la PAS à 60 ans.
2. Une étude rapporte r = 0,30 entre la consommation quotidienne de sel et la PAS. Testez la signification pour n = 40 et pour n = 200. Quelle conclusion générale tirez-vous sur la relation entre n et la signification ? Comment classez-vous la corrélation selon Colton dans les deux cas ?
3. Pour r = 0,62 obtenu sur 50 patients, calculez l'IC à 95% de ρ par la transformation de Fisher. La corrélation dans la population peut-elle être "forte" selon Colton ?
4. Sept patients ont un score de douleur sur l'échelle visuelle analogique (EVA) de 2, 5, 3, 8, 6, 9, 4 et un score d'anxiété de 10, 14, 9, 20, 18, 17, 12. Calculez à la main le coefficient de Spearman et interprétez-le.
5. Sur mg2026, la glycémie à jeun et l'HbA1c ont r = 0,738. Calculez R² et 1 - R² et formulez une phrase pour chacun, dans le style des devoirs de la discipline. Classez la corrélation selon Colton.
6. Classez selon la règle de Colton les coefficients suivants et précisez leur sens : 0,81 ; -0,573 ; 0,256 ; -0,088 ; 0,729 ; -0,832 ; 0,107 ; 0,50.
7. Les différences entre un tensiomètre automatique et l'appareil de référence, chez huit patients, sont (mmHg) : 4, -2, 6, 3, 8, 1, 5, -1. Calculez le biais moyen et les limites de concordance de Bland-Altman. Pourquoi le coefficient de corrélation entre les deux appareils n'aurait-il pas répondu à la question "peut-on les utiliser de façon interchangeable ?"

8. Un modèle logistique du diabète donne pour l'âge OR = 1,11 par an. Calculez l'OR approximatif pour une différence de 5 ans et de 10 ans. Est-il correct de dire que le diabète est "1,11 fois plus probable" à chaque année supplémentaire ? Expliquez la différence entre cote et probabilité.
9. À l'aide du tableau 9.6, calculez la PAS prédite pour une femme non-fumeuse de 45 ans avec un IMC de 27 et pour un homme fumeur de 65 ans avec un IMC de 35. Quelle variable contribue le plus à la différence entre eux ?
10. Une régression du poids (kg) en fonction de la taille (cm) chez 50 étudiants donne poids = -95 + 0,95 x taille, avec $R^2 = 0{,}38$, les tailles étant comprises entre 155 et 195 cm. (a) Interprétez la pente. (b) Quelle proportion de la variation du poids relève d'autres facteurs ? (c) Prédisez le poids pour 1,70 m et discutez la prédiction pour 1,20 m.

Pièges fréquents

- Rapporter un coefficient de corrélation sans le diagramme de dispersion ; une seule valeur extrême peut créer ou masquer une corrélation (figure 9.3).
- Confondre signification et intensité : r = -0,088 avec p = 0,01 reste une corrélation négligeable.
- Interpréter la corrélation comme une preuve de causalité, en ignorant les facteurs de confusion, la causalité inverse et le biais écologique.
- Utiliser r pour démontrer la concordance entre deux méthodes de mesure, au lieu de la méthode de Bland-Altman.
- Lire R² comme la proportion "causée" par X ; R² est la proportion de la variation associée linéairement à X.
- Interpréter cliniquement l'ordonnée à l'origine lorsque X = 0 n'a pas de sens (IMC nul, taille nulle).
- Extrapoler la droite de régression en dehors du domaine des données (le poids à 2,35 m).
- Vérifier la normalité sur la variable X au lieu des résidus ; en régression, la condition porte sur les résidus.
- Utiliser la formule abrégée de Spearman en présence de nombreuses valeurs égales.
- Écrire "on accepte H0" lorsque p > 0,05 ; la formulation correcte est "on ne rejette pas H0".

À retenir

- La corrélation mesure le sens et l'intensité de l'association linéaire (Pearson) ou monotone (Spearman) ; la régression décrit comment Y varie en moyenne avec X et permet la prédiction.
- $r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}}$, entre -1 et 1 ; le test utilise $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ avec n - 2 degrés de liberté ; l'IC se calcule par la transformation de Fisher.
- Règle de Colton : |r| < 0,25 pas de corrélation ; 0,25-0,50 faible ; 0,50-0,75 modérée ; ≥ 0,75 forte.
- R² = proportion de la variation de Y expliquée par le modèle ; 1 - R² = proportion due à d'autres facteurs.
- $b = \frac{S_{xy}}{S_{xx}} = \frac{r \times s_y}{s_x}$; a = my - b x mx ; la droite passe par le point des moyennes.
- La SEE (dans les unités de Y) indique la précision de la prédiction individuelle ; l'intervalle de prédiction pour un patient est beaucoup plus large que l'intervalle de confiance de la moyenne.
- Les conditions de la régression (LINE) se vérifient sur les résidus : linéarité, indépendance, normalité, variance constante.
- En régression multiple, chaque coefficient est ajusté sur les autres prédicteurs ; en régression logistique, e^b est un OR, et la valeur sans effet est 1.

- Sur mg2026 : IMC-PAS r = 0,405, rs = 0,372, $R^2 = 0{,}164$, PAS = 89,47 + 1,435 x IMC (IC de la pente 0,925-1,944), SEE 12,73 mmHg.

Questions d'autoévaluation

1. (QRU) Pour deux variables numériques, normalement distribuées, mesurées chez les mêmes 50 patients, la mesure la plus adaptée de l'intensité de l'association linéaire est :
    a) le test t pour échantillons indépendants
    b) le coefficient de corrélation de Pearson
    c) le test du khi-deux
    d) le coefficient kappa
    e) le rapport de cotes
2. (QRU) Un coefficient de corrélation r = -0,62 indique, selon la règle de Colton :
    a) une corrélation inverse forte
    b) l'absence de corrélation, car r est négatif
    c) une corrélation inverse modérée
    d) une corrélation directe modérée
    e) une corrélation inverse faible
3. (QRU) Entre la taille et le poids, on a obtenu r = 0,60. La proportion de la variation du poids due à des facteurs autres que la taille est de :
    a) 40%
    b) 36%
    c) 60%
    d) 64%
    e) 16%
4. (QRU) La droite de régression PAS = 89,47 + 1,435 x IMC a été obtenue chez des patients dont l'IMC était compris entre 23,5 et 48,3. L'affirmation correcte est :
    a) un patient avec un IMC de 0 a une PAS de 89,47 mmHg
    b) la PAS prédite pour un IMC de 30 est de 132,5 mmHg et représente la moyenne estimée pour les patients ayant un IMC de 30
    c) la pente montre que l'IMC provoque une augmentation de la PAS de 1,435 mmHg chez chaque patient
    d) la prédiction pour un IMC de 70 est aussi crédible que celle pour un IMC de 30
    e) r est négatif, car l'ordonnée à l'origine est positive
5. (QRU) Pour vérifier si un glucomètre portable peut remplacer le dosage de laboratoire, la méthode adaptée est :
    a) le coefficient de Pearson entre les deux dosages
    b) la régression logistique
    c) la méthode de Bland-Altman, avec le biais moyen et les limites de concordance
    d) le coefficient de Spearman
    e) le test du khi-deux d'indépendance
6. (QRM) À propos du coefficient de corrélation de Pearson, sont vraies les affirmations suivantes :
    a) il n'a pas d'unité de mesure
    b) il est sensible aux valeurs extrêmes
    c) une valeur proche de zéro exclut toute relation entre les variables

d) il change si le poids est exprimé en grammes au lieu de kilogrammes

e) il a le même signe que la pente de la droite de régression

7. (QRM) Le coefficient de corrélation de Spearman est préféré à celui de Pearson lorsque :

a) l'une des variables est ordinale (par exemple l'activité physique sur une échelle de 1 à 5)

b) les distributions sont fortement asymétriques ou il existe des valeurs extrêmes

c) les deux variables sont normales et la relation est linéaire

d) la relation est monotone, mais non linéaire

e) les variables sont nominales à trois catégories

8. (QRM) Sur un échantillon de 867 patients, on a obtenu, entre l'âge et la durée de l'intervention chirurgicale, r = -0,088, p = 0,01. Sont correctes les affirmations suivantes :

a) la corrélation est statistiquement significative

b) selon Colton, la corrélation est négligeable

c) l'âge explique environ 8,8% de la variation de la durée

d) R^2 est inférieur à 1%

e) le résultat démontre qu'un âge avancé raccourcit l'intervention

9. (QRM) Les conditions de la régression linéaire simple comprennent :

a) la relation linéaire entre X et Y

b) l'indépendance des observations

c) la distribution normale des résidus

d) la distribution normale de la variable X

e) la variance constante des résidus sur tout le domaine de X

10. (QRM) Dans une régression multiple de la PAS, le coefficient pour "fumeur, par rapport à ancien fumeur" est de 6,738 (IC à 95% 1,53-11,95 ; p = 0,012), et $R^2 = 0{,}308$, R^2 ajusté $= 0{,}286$. Sont correctes les affirmations suivantes :

a) à âge, IMC et sexe identiques, les fumeurs ont une PAS en moyenne plus élevée d'environ 6,7 mmHg que les anciens fumeurs

b) le modèle explique environ 31% de la variation de la PAS

c) le R^2 ajusté est inférieur au R^2 parce qu'il pénalise le nombre de prédicteurs

d) le tabagisme provoque une augmentation de la PAS d'exactement 6,738 mmHg chez chaque patient

e) l'intervalle de confiance qui ne contient pas 0 correspond à un coefficient significatif à $\alpha = 0{,}05$

Chapitre 10. Applications statistiques en épidémiologie : types d'études observationnelles et analyse du risque

Objectifs

À la fin de ce chapitre, l'étudiant sera capable de :

- définir le facteur de risque, le classer et décrire les types d'exposition (ponctuelle, chronique, dose, durée) ;
- calculer et interpréter la prévalence, l'incidence cumulée et le taux d'incidence en personnes-temps ;
- reconnaître d'après l'énoncé le type d'étude (expérimentale, transversale, de cohorte prospective ou rétrospective, cas-témoins, écologique) et choisir la mesure d'association adaptée ;
- construire le tableau de contingence 2 x 2 avec la notation a, b, c, d et calculer à la main le risque relatif (RR) et le rapport de cotes (OR) ;
- calculer les intervalles de confiance à 95% du RR et de l'OR par la méthode logarithmique et les utiliser pour conclure sur la signification ;
- expliquer pourquoi l'OR dépasse le RR lorsque la maladie est fréquente et pourquoi les deux sont proches pour les maladies rares ;
- calculer le risque attribuable, la fraction attribuable et le nombre de sujets à traiter (NST) ou le nombre nécessaire pour produire un effet indésirable (NNH) ;
- reconnaître un facteur de confusion, expliquer le principe de la stratification (Mantel-Haenszel) et distinguer la confusion de la modification de l'effet ;
- appliquer les critères de Bradford Hill et les recommandations STROBE à la lecture critique d'une étude observationnelle.

10.1 Le facteur de risque

10.1.1 Ce que nous appelons facteur de risque

Dans les chapitres précédents, nous avons comparé des moyennes et des proportions entre groupes. L'épidémiologie met ces mêmes outils au service d'une question précise : qu'est-ce qui augmente la probabilité qu'une personne en bonne santé développe une maladie donnée ? Un facteur de risque est une caractéristique, un comportement, une exposition ou une condition associée à une probabilité plus élevée de survenue de la maladie. Le mot clé est "probabilité" : le tabagisme ne garantit pas la BPCO, et de nombreux patients atteints de BPCO n'ont jamais fumé, mais la BPCO apparaît plus souvent chez les fumeurs. Le facteur de risque est une cause hypothétique, que l'étude tente de confirmer ou d'infirmer.

En médecine générale, nous travaillons chaque jour avec des facteurs de risque : lorsque nous calculons le score SCORE2 pour le risque cardiovasculaire, lorsque nous appliquons le questionnaire FINDRISC pour le diabète ou lorsque nous demandons à un patient combien de cigarettes il fume par jour. Chacun de ces outils synthétise les résultats de dizaines d'études épidémiologiques du type de celles décrites dans ce chapitre [64,111].

10.1.2 La classification des facteurs de risque

Les cours de la discipline regroupent les facteurs de risque selon leur nature. Les facteurs environnementaux comprennent les polluants atmosphériques, les toxines, les rayonnements, les micro-organismes infectieux et les expositions professionnelles (poussières, solvants, bruit). Les facteurs comportementaux relèvent des habitudes de la personne : le tabagisme, la consommation d'alcool ou de drogues, l'alimentation, la sédentarité, le non-respect des mesures de protection au travail. Les facteurs sociaux comprennent la pauvreté, le niveau d'études, l'isolement et les événements de vie traumatisants (le décès d'un membre de la famille, le divorce). Les facteurs génétiques sont héréditaires : l'hypercholestérolémie familiale, les mutations BRCA dans le cancer du sein et de l'ovaire, les antécédents familiaux de diabète.

Une seconde classification, utile en clinique, sépare les facteurs modifiables (tabagisme, obésité, hypertension, sédentarité) des facteurs non modifiables (âge, sexe, génétique). Les premiers sont la cible de la prévention ; les seconds aident à identifier les personnes qui doivent être suivies plus attentivement. Dans le jeu de données mg2026, l'obésité et le tabagisme sont modifiables, alors que l'âge et le sexe ne le sont pas.

10.1.3 L'exposition au facteur de risque

L'exposition peut être ponctuelle (un seul événement, limité dans le temps : un accident, une piqûre accidentelle avec une aiguille contaminée, une intoxication alimentaire lors d'un mariage) ou chronique (prolongée, le type le plus fréquent dans les maladies non transmissibles). Pour l'exposition chronique, on décrit trois dimensions : la dose actuelle (combien de cigarettes par jour le patient fume actuellement), la dose cumulée (par exemple le nombre de paquets-années, c'est-à-dire le nombre de paquets par jour multiplié par le nombre d'années de tabagisme) et la durée de l'exposition (depuis combien d'années). Un patient qui fume un paquet par jour depuis 30 ans totalise 30 paquets-années ; un patient qui a fumé deux paquets par jour pendant 15 ans totalise également 30 paquets-années. La mesure correcte de l'exposition est l'une des plus grandes difficultés des études observationnelles, car elle repose souvent sur la mémoire du patient.

10.1.4 Facteur causal et marqueur de risque

Tout facteur associé à la maladie n'en est pas pour autant la cause. Un facteur causal participe effectivement au mécanisme de la maladie : son élimination réduit l'incidence (l'arrêt du tabac réduit le risque de cancer du poumon). Un marqueur de risque (facteur favorisant, indicateur) est associé à la maladie et nous aide à repérer les personnes à risque élevé, mais il ne fait pas nécessairement partie de la chaîne causale. Les doigts jaunis par la nicotine sont associés au cancer du poumon, mais personne n'envisagerait de prévenir le cancer en se lavant les mains. Une protéine C réactive élevée est associée aux événements coronariens, mais sa diminution artificielle ne garantit pas la diminution du risque. Cette différence a des conséquences pratiques : un marqueur est utile pour le dépistage, un facteur causal l'est aussi pour la prévention.

10.2 La mesure de la fréquence de la maladie

Avant de comparer exposés et non-exposés, nous devons savoir comment mesurer la fréquence d'une maladie. Il existe deux familles de mesures : la prévalence, qui décrit combien de personnes sont malades, et l'incidence, qui décrit combien de personnes tombent malades.

10.2.1 La prévalence

La prévalence est la proportion de personnes d'une population qui présentent la maladie à un moment donné ou au cours d'une période donnée :

Prévalence = nombre de personnes atteintes de la maladie / nombre total de personnes de la population

La prévalence ponctuelle se rapporte à un moment (par exemple le jour de l'examen), la prévalence de période à un intervalle (par exemple la proportion de personnes ayant eu au moins un épisode de lombalgie au cours de la dernière année). La prévalence inclut aussi bien les nouveaux cas que les cas anciens. Dans le jeu de données mg2026, à l'inclusion, 29 des 160 patients avaient un diabète confirmé par HGPO : la prévalence est de 29 / 160 = 18,1%, avec un IC à 95% de Wilson de 12,9-24,8% (chap. 4). La prévalence est la mesure dont a besoin un gestionnaire de santé : combien de lits, combien de consultations, combien de doses de médicament sont nécessaires maintenant.

10.2.2 L'incidence cumulée (le risque)

L'incidence cumulée est la proportion de personnes initialement indemnes de la maladie qui développent la maladie au cours d'une période définie :

Incidence cumulée = nombre de nouveaux cas pendant la période / nombre de personnes indemnes de la maladie au début de la période

Elle équivaut à la probabilité (au risque) de développer la maladie au cours de cette période. La période doit toujours être précisée : un risque de 8% "sur 5 ans" est tout autre chose que 8% "par an". Si, sur 1000 personnes en bonne santé suivies pendant 5 ans, 80 développent un diabète, l'incidence cumulée à 5 ans est de 80 / 1000 = 8%. L'incidence cumulée exige que tous les participants soient suivis pendant toute la période ; pour les perdus de vue, nous ne savons pas s'ils sont tombés malades ou non.

10.2.3 Le taux d'incidence et les personnes-temps

Dans les études réelles, les participants entrent dans l'étude à des moments différents, certains sont perdus de vue, d'autres décèdent d'autres causes. La solution consiste à comptabiliser le temps effectif d'observation. Une personne suivie pendant 3 ans contribue pour 3 personnes-années ; une personne suivie pendant 6 mois contribue pour 0,5 personne-année. Le taux d'incidence (densité d'incidence) est :

Taux d'incidence = nombre de nouveaux cas / somme des personnes-temps d'observation

Exemple : 200 patients atteints de diabète de type 2 sont suivis pour la survenue d'une rétinopathie. En raison des entrées échelonnées et des pertes de vue, le temps total d'observation s'élève à 850 personnes-années, et 34 nouveaux cas apparaissent pendant ce temps. Le taux d'incidence est de 34 / 850 = 0,040 cas par personne-année, soit 4,0 cas pour 100 personnes-années. Le taux n'est pas une proportion : son unité est "par unité de temps", et sa valeur peut théoriquement dépasser 1 (une maladie très fréquente, avec des épisodes répétés). Le rapport de deux taux d'incidence s'appelle le rapport des taux d'incidence (IRR, angl. incidence rate ratio) et s'interprète de la même façon que le risque relatif.

10.2.4 Le lien entre prévalence et incidence

Pour une maladie à l'état d'équilibre (incidence et durée approximativement constantes) et de faible prévalence, il existe une relation simple :

$$P \approx I \times D$$

où P est la prévalence, I le taux d'incidence et D la durée moyenne de la maladie. Une maladie chronique incurable, mais avec une longue survie (le diabète, l'hypertension), a une prévalence élevée même si son incidence est modérée : une incidence de 2 cas pour 1000 personnes-années et une durée moyenne de 10 ans donnent une prévalence d'environ 0,002 x 10 =

0,02, soit 2%. Une maladie aiguë de courte durée (la grippe, une infection respiratoire d'une semaine) peut avoir une incidence très élevée et une prévalence ponctuelle faible. Cette relation explique aussi un paradoxe apparent : un nouveau traitement qui prolonge la survie sans guérir augmente la prévalence de la maladie, alors qu'il constitue un progrès. C'est pourquoi la comparaison des prévalences entre régions ou entre périodes demande de la prudence.

10.3 Les types d'études épidémiologiques

10.3.1 Études expérimentales et études observationnelles

La différence fondamentale entre les types d'études tient à qui décide de l'exposition. Dans l'étude expérimentale (d'intervention), le chercheur attribue l'exposition, idéalement par randomisation : il décide qui reçoit le nouveau traitement et qui reçoit le placebo ou le traitement standard. La randomisation répartit de manière égale, en moyenne, tous les facteurs connus et inconnus entre les groupes, si bien que la différence de résultat peut être attribuée à l'intervention. Les essais cliniques randomisés fournissent les preuves causales les plus solides, mais ils ne peuvent pas être utilisés pour des facteurs de risque nocifs : personne ne peut randomiser des personnes pour qu'elles fument ou travaillent dans la poussière. Le jeu de données mg2026 correspond à une telle étude expérimentale simulée (bras A, B, C), et le chapitre 13 l'examine en détail.

Dans l'étude observationnelle, l'exposition s'est produite indépendamment du chercheur (le patient a choisi de fumer, a travaillé dans une usine, est né avec un gène donné). Le chercheur se contente d'observer et de mesurer. Les principaux inconvénients sont la précision limitée de la mesure de l'exposition et le fait que les exposés et les non-exposés peuvent différer aussi par d'autres facteurs. Les études observationnelles se classent selon le moment où l'on mesure l'exposition et la maladie. La figure 10.1 présente schématiquement les trois principaux schémas d'étude.

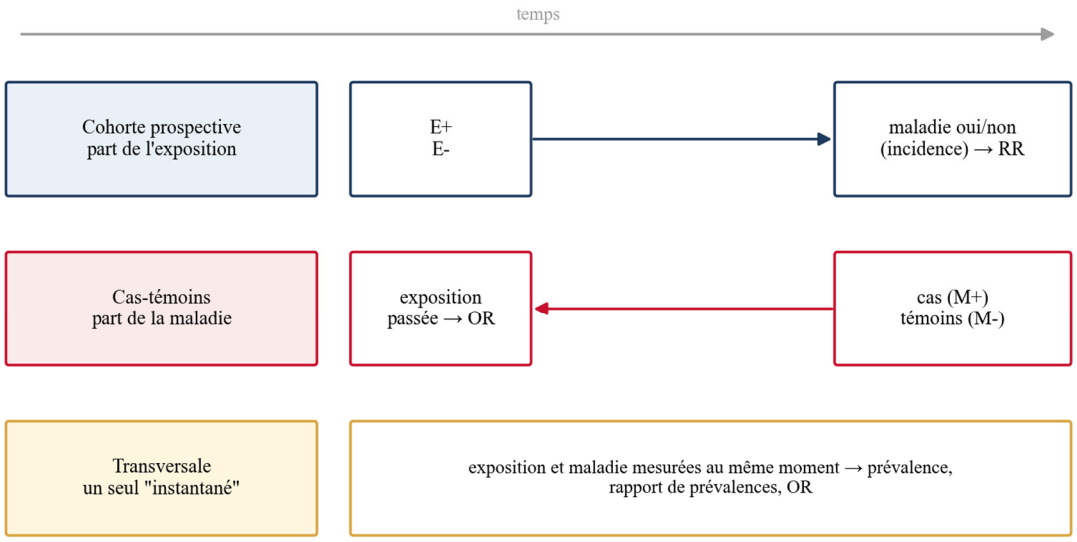


Figure 10.1. Les trois principaux schémas d'études observationnelles. L'étude de cohorte part de l'exposition (E+ et E-) et suit dans le temps la survenue de la maladie, ce qui permet de calculer l'incidence et le risque relatif. L'étude cas-témoins part de la maladie (cas et témoins) et recherche rétrospectivement l'exposition, ce qui ne permet de calculer que le rapport de cotes. L'étude transversale mesure l'exposition et la maladie au même moment et estime la prévalence, le rapport de prévalences et l'OR.

10.3.2 L'étude transversale

L'étude transversale (angl. cross-sectional) est une "photographie" de la population à un moment donné : pour chaque participant, on enregistre, au même moment, la présence ou l'absence de l'exposition et de la maladie. C'est une étude de prévalence et le modèle de recherche le plus simple. L'analyse à l'inclusion de mg2026, dans laquelle nous avons mis en relation l'obésité et le diabète confirmé par HGPO, est transversale : les deux ont été mesurés lors de la première consultation.

La mesure d'association naturelle est le rapport de prévalences (RP, angl. prevalence ratio) : la prévalence de la maladie chez les exposés divisée par la prévalence chez les non-exposés. La formule est identique à celle du risque relatif, mais le nom diffère, car nous ne mesurons pas des cas nouveaux, mais des cas existants. On rapporte aussi fréquemment l'OR, surtout lorsque l'on utilise la régression logistique.

Les avantages sont le faible coût et la rapidité, la possibilité d'étudier simultanément plusieurs facteurs et plusieurs maladies et l'utilité pour la planification des services de santé. L'inconvénient majeur est que nous ne pouvons pas établir la succession temporelle : si les obèses ont plus souvent un diabète, nous ne savons pas, à partir de cette seule étude, si l'obésité a précédé le diabète. Une enquête transversale portant sur 2976 étudiants du département de Timiș, dont 62,5% de filles, a identifié par régression logistique des facteurs associés au tabagisme intensif [22] ; le schéma permet de décrire des associations, mais non d'établir ce qui est apparu en premier, le tabagisme ou les comportements associés. Un second inconvénient tient à la survie sélective : les maladies d'évolution rapide (décès ou guérison) sont sous-représentées, car les malades concernés ne sont plus présents au moment de la "photographie". L'estimation de la prévalence est faussée pour les maladies d'évolution rapide.

10.3.3 L'étude de cohorte

Dans une étude de cohorte, les participants sont sélectionnés au départ indemnes de la maladie, classés selon l'exposition (exposés E+ et non-exposés E-) et suivis dans le temps pour la survenue de la maladie [112]. Le terme vient des cohortes des légions romaines : un groupe qui avance ensemble. Comme nous partons de personnes en bonne santé et que nous comptons les nouveaux cas, nous pouvons calculer directement l'incidence chez les exposés et chez les non-exposés, et donc le risque relatif.

La cohorte prospective commence maintenant et suit les participants dans le futur (par exemple 921 étudiants suivis pendant leur première année d'université [5]). La cohorte rétrospective (historique) utilise des enregistrements du passé : le chercheur établit à partir des documents qui était exposé il y a 10 ans et vérifie qui est tombé malade depuis. Les deux conservent la logique "de l'exposition vers la maladie" ; seul diffère le moment où le chercheur entre en scène. Les cours de la discipline, suivant une tradition plus ancienne, retiennent que, dans la cohorte rétrospective, on calcule seulement l'OR ; en réalité, si les enregistrements permettent d'identifier complètement les exposés et les non-exposés ainsi que les nouveaux cas, le RR peut également être calculé dans une cohorte rétrospective. Ce qui compte, c'est que les proportions de malades dans chaque groupe d'exposition reflètent le risque réel.

Les avantages de la cohorte sont une succession temporelle claire (l'exposition précède la maladie), le calcul direct de l'incidence et du RR, la possibilité d'étudier plusieurs maladies produites par la même exposition et la protection contre les biais de mémorisation, car l'exposition est enregistrée avant la survenue de la maladie. Les inconvénients sont le coût, la durée (les maladies à longue latence, comme le cancer, demandent des dizaines d'années), les pertes de vue et l'inefficacité pour les maladies rares : pour observer quelques cas d'une maladie rare, il faudrait suivre des dizaines de milliers de personnes.

10.3.4 L'étude cas-témoins

L'étude cas-témoins inverse la logique : elle part de la maladie [113]. Le chercheur sélectionne un groupe de personnes malades (les cas, B+) et un groupe de personnes indemnes de la maladie (les témoins, B-), puis recherche dans le passé de chacun l'exposition au facteur de risque, par entretien, questionnaire ou dossiers médicaux.

Comme le chercheur décide du nombre de cas et de témoins inclus (par exemple un cas pour deux témoins), la proportion de malades dans l'échantillon ne reflète pas le risque dans la population. C'est pourquoi, dans une étude cas-témoins, on ne peut calculer ni l'incidence ni le risque relatif. On calcule seulement le rapport de cotes, qui possède une propriété remarquable (section 10.6) : l'OR de l'exposition chez les cas par rapport aux témoins est égal à l'OR de la maladie chez les exposés par rapport aux non-exposés.

Le choix des témoins est le point le plus délicat. Les témoins doivent provenir de la même population que les cas : si un témoin avait développé la maladie, il aurait été inclus dans l'étude comme cas. Pour des cas recrutés dans un hôpital, les témoins sont choisis en règle générale dans le même hôpital, parmi des patients atteints d'autres maladies sans lien avec l'exposition, ou dans la population desservie par l'hôpital. Des témoins choisis dans une population différente (par exemple des étudiants en bonne santé pour des cas âgés hospitalisés) produisent un biais de sélection.

L'appariement (angl. matching) consiste à choisir les témoins de façon qu'ils ressemblent aux cas pour des variables qui, sinon, introduiraient une confusion (âge, sexe). L'appariement peut être individuel (on trouve pour chaque cas un ou plusieurs témoins de même âge et de même sexe) ou en fréquence (la distribution par âge et par sexe est la même dans les deux groupes). L'appariement élimine la confusion due aux variables d'appariement, mais ces variables ne peuvent plus être étudiées comme facteurs de risque, et l'analyse doit tenir compte des paires.

Le biais de mémorisation (angl. recall bias) est propre aux études cas-témoins : les malades, qui ont cherché des explications à leur maladie, se souviennent mieux de leurs expositions ou les surestiment, tandis que les personnes en bonne santé les minimisent. Une mère dont l'enfant présente une malformation congénitale se souviendra en détail de chaque médicament pris pendant la grossesse ; une mère dont l'enfant est en bonne santé peut oublier. Le résultat est un OR artificiellement augmenté.

Les avantages de l'étude cas-témoins sont le faible coût, la rapidité, l'efficacité pour les maladies rares et pour celles à longue latence, ainsi que la possibilité d'étudier simultanément plusieurs facteurs de risque pour une même maladie. Les inconvénients sont l'impossibilité de calculer l'incidence et le RR, les biais de sélection et de mémorisation et la difficulté d'établir avec certitude que l'exposition a précédé la maladie.

Une variante importante est l'étude cas-témoins "nichée" (angl. nested case-control), dans laquelle les cas et les témoins sont extraits d'une cohorte déjà définie. Cette combinaison conserve les avantages de la cohorte (l'exposition a été documentée de manière prospective) et l'efficacité de l'étude cas-témoins.

10.3.5 L'étude écologique

Dans l'étude écologique, l'unité d'analyse n'est pas la personne, mais le groupe : le pays, le département, la ville, le mois. On compare la fréquence de l'exposition à la fréquence de la maladie au niveau du groupe, par exemple la consommation moyenne de sel par département et la mortalité par accident vasculaire cérébral par département. Les études écologiques sont peu coûteuses, utilisent des données déjà existantes et sont utiles pour générer des hypothèses et pour évaluer les interventions de santé publique. Leur limite majeure est le biais écologique, discuté au

chapitre 9 : une association au niveau du groupe ne se transpose pas automatiquement au niveau individuel, et de nombreuses variables externes peuvent agir comme facteurs de confusion. L'analyse de séries temporelles interrompues, dans laquelle on compare la tendance d'un taux avant et après un événement, est une forme moderne d'étude écologique.

10.3.6 La comparaison des schémas d'étude

Tableau 10.1. Comparaison des principaux types d'études épidémiologiques.

Caractéristique	Transversale	Cohorte	Cas-témoins	Écologique	Expérimentale
Point de départ	la population, à un moment donné	l'exposition	la maladie	des groupes	l'attribution par le chercheur
Direction	simultanée	vers l'avant (vers le futur)	vers l'arrière (vers le passé)	simultanée, en général	vers l'avant
Mesure de fréquence	la prévalence	l'incidence	ne peut pas être calculée	taux par groupe	l'incidence
Mesure d'association	RP, OR	RR, OR, IRR, HR	OR seulement	corrélation, IRR	RR, différence de risques, NST
Succession temporelle	non	oui	incertaine	non	oui
Maladies rares	inadaptée	inadaptée	très adaptée	possible	inadaptée
Coût et durée	faibles	élevés	faibles	très faibles	très élevés
Biais typiques	survie sélective	pertes de vue	sélection, mémorisation	biais écologique	écarts au protocole

Ce tableau ne doit pas être mémorisé comme une liste ; il découle de la question "d'où part l'étude ?". Si elle part de l'exposition et compte de nouveaux cas, nous pouvons calculer le risque. Si elle part de la maladie, nous ne le pouvons pas.

Étude de cas : cinq schémas d'étude dans les articles des collègues

Une étude observationnelle de suivi portant sur 110 patientes présentant des lésions cervicales de haut grade traitées par excision (LEEP), avec un test HPV à 6, 12 et 18 mois, est une cohorte : les patientes ont été classées selon l'âge et le type d'HPV et suivies pour la persistance de l'infection [114]. Le taux de persistance a diminué de 40,9% à 6 mois à 20% à 12 mois et 11,8% à 18 mois, et l'âge, dichotomisé à 36,5 ans, a été associé à la persistance : RR = 2,75, IC à 95% 1,34-5,64, p = 0,0027. L'intervalle ne contient pas 1, l'association est donc significative ; sa largeur (d'un risque 1,3 fois plus élevé à un risque 5,6 fois plus élevé) reflète l'effectif de 110 patientes.

Une étude transversale portant sur 101 patients séropositifs pour le VIH sous traitement antirétroviral, dans l'ouest de la Roumanie, a analysé l'association entre la sévérité de l'infection par le VIH et les troubles psychiatriques au moyen de rapports de cotes : OR entre 2,81 (p = 0,038) et 8,72 (p < 0,001) selon les troubles [84]. Le schéma transversal justifie l'OR, mais ne permet pas d'établir le sens de la relation (l'infection sévère par le VIH conduit-elle aux troubles psychiatriques, ou les troubles psychiatriques conduisent-ils à une moins bonne observance et donc à une maladie plus sévère ?). Le fait que le résumé donne les OR avec p, mais sans intervalles de confiance, et que de nombreux troubles soient testés simultanément constituent deux points de lecture critique : avec sept comparaisons, la probabilité qu'au moins une ressorte significative par hasard augmente.

Une étude portant sur 656 patients atteints d'infarctus du myocarde avec sus-décalage du segment ST traités par angioplastie primaire a comparé les 96 patients présentant un phénomène de no-reflow au reste de la cohorte, quant aux complications à court terme [115]. Il s'agit d'un schéma cas-témoins construit à l'intérieur d'une cohorte. Une étude rétrospective portant sur 435 patients atteints de cholécystite aiguë lithiasique a utilisé un appariement sur

> l'âge et le sexe, avec un rapport de 1:2 (145 cas de forme gangreneuse, 290 témoins de forme chronique), et a rapporté pour un indice inflammatoire OR = 1,976, p < 0,001, dans un modèle de régression logistique [116]. L'appariement sur l'âge et le sexe signifie que ces deux variables ne peuvent plus être analysées comme facteurs de risque, mais qu'elles ne peuvent plus non plus agir comme facteurs de confusion. Enfin, une étude clinique transversale avec analyse cas-témoins nichée a porté sur 136 patients (52 cas présentant des troubles de l'interaction intestin-cerveau et 84 témoins asymptomatiques) et a trouvé un OR protecteur de 0,36 (IC à 95% 0,18-0,74 ; p = 0,009) [71]. Un OR inférieur à 1, avec un intervalle qui ne contient pas 1, signifie que l'exposition en question était plus rare chez les cas que chez les témoins.

10.4 Le tableau de contingence 2 x 2

Tous les calculs de l'analyse du risque partent du même tableau. La convention utilisée dans la discipline (et dans la plupart des manuels) place l'exposition en lignes et la maladie en colonnes, avec les exposés sur la première ligne et les malades dans la première colonne.

Tableau 10.2. Notation du tableau de contingence 2 x 2.

	Maladie présente (B+)	**Maladie absente (B-)**	**Total**
Exposés (E+)	a	b	a + b
Non-exposés (E-)	c	d	c + d
Total	a + c	b + d	$N = a + b + c + d$

La cellule a contient les exposés malades, b les exposés sains, c les non-exposés malades, d les non-exposés sains. Les totaux marginaux des lignes (a + b et c + d) sont les nombres d'exposés et de non-exposés ; les totaux marginaux des colonnes (a + c et b + d) sont les nombres de malades et de sujets sains. À l'examen, l'erreur la plus fréquente dans l'analyse du risque n'est pas une formule fausse, mais un tableau mal construit : lignes et colonnes inversées, ou un total placé à la place d'une cellule. Vérifiez toujours que a + b + c + d = N et que chaque total de l'énoncé se retrouve dans le tableau.

Dans une étude de cohorte ou expérimentale, les totaux des lignes sont fixés par le schéma (tant d'exposés, tant de non-exposés). Dans une étude cas-témoins, ce sont les totaux des colonnes qui sont fixés (tant de cas, tant de témoins). Dans une étude transversale, seul N est fixé. Cette différence détermine quelles proportions ont un sens : dans une étude cas-témoins, a / (a + b) n'est pas le risque de maladie chez les exposés, car le nombre de malades a été choisi par le chercheur.

10.5 Le risque relatif

10.5.1 La formule et l'interprétation

Le risque de maladie chez les exposés est $R_1 = \frac{a}{a+b}$, et chez les non-exposés $R_0 = \frac{c}{c+d}$. Le risque relatif (RR, angl. relative risk ou risk ratio) est leur rapport :

$$\mathrm{RR} = \frac{R_1}{R_0} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Le RR indique combien de fois le risque de maladie est plus élevé chez les exposés que chez les non-exposés. L'interprétation distingue trois situations. RR > 1 : l'exposition est un facteur de risque (les exposés tombent malades plus souvent). RR = 1 : l'exposition n'est pas associée à la maladie (facteur indifférent). RR < 1 : l'exposition est un facteur protecteur (les exposés tombent malades plus rarement) ; par exemple, RR = 0,6 signifie un risque inférieur de 40% chez les

exposés. Cette dernière interprétation est importante pour les traitements et les vaccins, où l'"exposition" est l'intervention.

Bien entendu, la décision finale ne se prend pas seulement d'après la valeur ponctuelle, mais d'après l'intervalle de confiance : un RR de 1,8 avec un IC à 95% de 0,9-3,6 ne montre pas un facteur de risque significatif, car l'intervalle contient 1.

10.5.2 L'exemple classique : le tabagisme et la BPCO

Une étude suit pendant 5 ans 1200 personnes initialement en bonne santé, dont 300 sont fumeurs. À la fin, 50 des fumeurs et 70 des non-fumeurs ont une BPCO. La première étape est le tableau. Fumeurs : a = 50 malades, b = 300 - 50 = 250 sujets sains. Les non-fumeurs sont 1200 - 300 = 900 : c = 70 malades, d = 900 - 70 = 830 sujets sains.

Tableau 10.3. Tabagisme et BPCO, étude de cohorte portant sur 1200 participants.

	BPCO oui	**BPCO non**	**Total**
Fumeurs	50	250	300
Non-fumeurs	70	830	900
Total	120	1080	1200

R1 = 50 / 300 = 0,1667 (16,7% sur 5 ans) ; $R_0 = \frac{70}{900} = 0{,}0778 (7{,}8\%)$; RR $= \frac{0{,}1667}{0{,}0778} = 2{,}14$. Les fumeurs ont un risque de BPCO 2,14 fois plus élevé que les non-fumeurs au cours des 5 ans de suivi.

Une erreur que nous voyons chaque année : certains étudiants divisent 70 par 1200 (le total des participants) au lieu de 900 (le nombre de non-fumeurs) et obtiennent R0 = 0,0583 et RR = 2,86. Le risque chez les non-exposés se calcule uniquement parmi les non-exposés. Un RR de 2,86 n'a pas de sens dans ce tableau, et un problème mal résolu dès la première étape entraîne des erreurs dans tous les autres résultats.

10.6 Le rapport de cotes

10.6.1 La cote

La cote (angl. odds) d'un événement est le rapport entre la probabilité que l'événement se produise et la probabilité qu'il ne se produise pas :

$$\text{cote} = \frac{p}{1-p}$$

Le terme vient des paris : une cote de "1 contre 4" signifie que, pour chaque personne qui tombe malade, 4 ne tombent pas malades, c'est-à-dire $p = \frac{1}{5} = 20\%$ et cote $= \frac{0{,}20}{0{,}80} = 0{,}25$. Pour les événements rares, la cote et la probabilité sont proches (p = 2% donne une cote de 0,0204) ; pour les événements fréquents, elles diffèrent beaucoup (p = 50% donne une cote de 1 ; p = 80% donne une cote de 4).

Dans le tableau 2 x 2, la cote de la maladie chez les exposés est a / b (malades sur sujets sains, parmi les exposés), et chez les non-exposés c / d.

10.6.2 La formule et l'interprétation

Le rapport de cotes (OR, angl. odds ratio) est :

$$\text{OR} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \times d}{b \times c}$$

La seconde forme, "le produit de la diagonale principale divisé par le produit de la diagonale secondaire", se retient facilement. L'interprétation suit les mêmes règles que pour le RR : OR > 1 facteur de risque, OR = 1 facteur indifférent, OR < 1 facteur protecteur. Pour l'exemple tabagisme-BPCO : $\text{OR} = \frac{50 \times 830}{250 \times 70} = \frac{41\,500}{17\,500} = 2{,}37$. La cote de développer une BPCO est 2,37 fois plus élevée chez les fumeurs.

L'OR possède une propriété de symétrie qui le rend indispensable dans les études cas-témoins. La cote de l'exposition chez les cas est a / c, et chez les témoins b / d. Leur rapport est $\frac{\frac{a}{c}}{\frac{b}{d}} = \frac{a \times d}{b \times c}$, exactement le même OR. Par conséquent, bien que nous ne puissions pas calculer le risque de maladie dans une étude cas-témoins, nous pouvons calculer l'OR de l'exposition, qui est numériquement égal à l'OR de la maladie. L'interprétation correcte dans une étude cas-témoins est : "la cote d'avoir été exposé est OR fois plus élevée chez les cas que chez les témoins" et, par symétrie, "la cote de la maladie est OR fois plus élevée chez les exposés".

10.6.3 Pourquoi l'OR est plus grand que le RR lorsque la maladie est fréquente

Dans l'exemple tabagisme-BPCO, OR = 2,37 et RR = 2,14 : des valeurs proches, mais non égales. Les cours de la discipline notent "habituellement OR > RR". La règle exacte est la suivante : lorsque RR > 1, l'OR est toujours plus éloigné de 1 que le RR ; lorsque RR < 1, l'OR est plus petit que le RR (lui aussi plus éloigné de 1). La différence dépend de la fréquence de la maladie. Le tableau 10.4 montre, pour un RR fixé à 2, comment l'OR augmente à mesure que le risque chez les non-exposés augmente.

Tableau 10.4. OR correspondant à un RR = 2, pour différents risques chez les non-exposés.

Risque chez les non-exposés (R0)	Risque chez les exposés (R1)	RR	OR
1%	2%	2	2,02
5%	10%	2	2,11
10%	20%	2	2,25
20%	40%	2	2,67
30%	60%	2	3,50
40%	80%	2	6,00

L'explication apparaît dans les formules : $\text{OR} = \frac{\text{RR} \times (1 - R_0)}{1 - R_1}$. Lorsque les deux risques sont faibles, le rapport (1 - R0) / (1 - R1) est proche de 1, et OR ≈ RR. C'est l'"hypothèse de la maladie rare" (angl. rare disease assumption) : pour des maladies dont la fréquence est inférieure à environ 10%, l'OR obtenu dans une étude cas-témoins est une bonne approximation du RR. Pour les maladies fréquentes, l'OR exagère l'effet s'il est lu comme un RR. Dans le jeu de données mg2026, la prévalence du diabète chez les obèses est de 33,3% : l'OR (4,32) est nettement plus grand que le RP (3,21). Un article qui rapporte un OR de 4,32 et le traduit par "le risque est multiplié par plus de 4" exagère.

10.7 Intervalles de confiance et tests statistiques

10.7.1 L'intervalle de confiance du RR par la méthode logarithmique

La distribution d'échantillonnage du RR est asymétrique : le RR ne peut pas être négatif, mais peut augmenter indéfiniment. Le logarithme népérien du RR a en revanche une distribution approximativement normale. Le calcul comporte trois étapes : on passe à l'échelle logarithmique, on construit l'intervalle habituel (estimation $\pm 1{,}96 \times$ ES), puis on revient par exponentiation.

$$\text{ES}[\ln(\text{RR})] = \sqrt{\left[\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}\right]} \quad \text{IC 95\% pour RR} = \exp[\ln(\text{RR}) \pm 1{,}96 \times \text{ES}]$$

Pour tabagisme-BPCO : ln(2,1429) = 0,762. $\text{ES} = \sqrt{\frac{1}{50} - \frac{1}{300} + \frac{1}{70} - \frac{1}{900}} = \sqrt{0{,}02 - 0{,}00333 + 0{,}01429 - 0{,}00111} = \sqrt{0{,}02984} = 0{,}1727$. Marge : 1,96 x 0,1727 = 0,339. Sur l'échelle logarithmique : 0,762 - 0,339 = 0,424 et 0,762 + 0,339 = 1,101. Par exponentiation : $e^{0{,}424} = 1{,}53$ et $e^{1{,}101} = 3{,}01$. IC à 95% du RR : 1,53-3,01.

Remarquez que l'intervalle n'est pas symétrique autour de 2,14 (2,14 - 1,53 = 0,61, mais 3,01 - 2,14 = 0,87). Cette asymétrie est normale pour les rapports, et l'intervalle est symétrique sur l'échelle logarithmique ; c'est pourquoi les diagrammes en forêt (forest plots) utilisent un axe logarithmique.

10.7.2 L'intervalle de confiance de l'OR (méthode de Woolf)

$$\text{ES}[\ln(\text{OR})] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \quad \text{IC 95\% pour OR} = \exp[\ln(\text{OR}) \pm 1{,}96 \times \text{ES}]$$

Pour tabagisme-BPCO : ln(2,3714) = 0,8635. $\text{ES} = \sqrt{\frac{1}{50} + \frac{1}{250} + \frac{1}{70} + \frac{1}{830}} = \sqrt{0{,}02 + 0{,}004 + 0{,}01429 + 0{,}00120} = \sqrt{0{,}03949} = 0{,}1987$. Marge : 1,96 x 0,1987 = 0,3895. Échelle logarithmique : 0,474 et 1,253. Par exponentiation : 1,61 et 3,50. IC à 95% de l'OR : 1,61-3,50.

La formule montre pourquoi les petites cellules produisent des intervalles larges : si l'une des cellules vaut 5, le terme 1/5 = 0,2 domine la somme. Si une cellule vaut 0, l'OR et son ES ne peuvent pas être calculés ; les logiciels ajoutent alors 0,5 à chaque cellule (correction de Haldane) ou utilisent des méthodes exactes.

10.7.3 Le test du khi-deux et le test de Fisher

Le test de signification de l'association dans le tableau 2 x 2 est le test du khi-deux d'indépendance vu au chapitre 7. Les hypothèses sont formulées en termes de mesure d'association :

H0 : RR = 1 (ou OR = 1), l'exposition n'est pas associée à la maladie H1 : RR ≠ 1 (ou OR ≠ 1), l'exposition est associée à la maladie

Pour tabagisme-BPCO, les fréquences attendues sous H0 sont : pour la cellule a, $\frac{300 \times 120}{1200} = 30$; b : $\frac{300 \times 1080}{1200} = 270$; c : $\frac{900 \times 120}{1200} = 90$; d : $\frac{900 \times 1080}{1200} = 810$. Toutes les différences observé-attendu ont la même valeur absolue, 20 (50 - 30 = 20 ; 250 - 270 = -20, etc.), ce qui se produit toujours dans un tableau 2 x 2. $\chi^2 = \frac{20^2}{30} + \frac{20^2}{270} + \frac{20^2}{90} + \frac{20^2}{810} = 13{,}33 + 1{,}48 + 4{,}44 + 0{,}49 = 19{,}75$, avec 1 degré de liberté, p < 0,001. On rejette H0.

Lorsqu'une fréquence attendue est inférieure à 5, on utilise le test exact de Fisher. La correction de Yates (de continuité) réduit la valeur du χ^2 et rend le test plus conservateur ; les logiciels la rapportent à côté du test non corrigé.

10.7.4 Le lien entre l'intervalle et le test

Un IC à 95% du RR ou de l'OR qui ne contient pas 1 correspond, presque toujours, à un test significatif à α = 0,05 ; un intervalle qui contient 1 correspond à un résultat non significatif. L'équivalence n'est pas parfaite, car l'intervalle logarithmique et le test du khi-deux sont calculés différemment et peuvent, à la limite, conduire à des conclusions différentes (le problème résolu 10.5 montre un tel cas). Rapporter l'intervalle est préférable à rapporter seulement la valeur p, car il montre à la fois l'ampleur et la précision de l'effet [70].

10.8 Mesures d'impact : le risque attribuable, la fraction attribuable, le NST et le NNH

10.8.1 Le risque attribuable (la différence de risques)

Le RR est une mesure relative : il nous dit combien de fois le risque augmente, mais pas de combien en termes absolus. Un RR de 2 peut signifier une augmentation de 1 à 2 cas par million (négligeable pour la santé publique) ou de 20% à 40% (énorme). Le risque attribuable (RA), appelé aussi différence de risques (DR, angl. risk difference), est :

$$\text{RA} = R_1 - R_0$$

Pour tabagisme-BPCO : RA = 0,1667 - 0,0778 = 0,0889, soit 8,9 points de pourcentage sur 5 ans. Sur 100 fumeurs, environ 9 cas de BPCO en 5 ans sont dus au tabagisme (si la relation est causale), en plus des 7-8 cas qui seraient survenus de toute façon. L'IC à 95% de la différence de risques se calcule comme pour la différence de deux proportions (chap. 4) : $\text{RA} \pm 1{,}96 \times \sqrt{\left[\frac{R_1(1-R_1)}{a+b} + \frac{R_0(1-R_0)}{c+d}\right]}$. Pour notre exemple : $0{,}0889 \pm 1{,}96 \times \sqrt{\frac{0{,}1667 \times 0{,}8333}{300} + \frac{0{,}0778 \times 0{,}9222}{900}} = 0{,}0889 \pm 1{,}96 \times 0{,}0233 = 0{,}043 - 0{,}135$.

10.8.2 La fraction attribuable

La fraction attribuable chez les exposés (FAe, appelée aussi fraction étiologique chez les exposés) indique quelle proportion des cas survenus chez les exposés est due à l'exposition :

$$\text{FAe} = \frac{R_1 - R_0}{R_1} = \frac{\text{RR} - 1}{\text{RR}}$$

Pour tabagisme-BPCO : FAe = (2,14 - 1) / 2,14 = 0,533. Environ 53% des cas de BPCO survenus chez les fumeurs pourraient être attribués au tabagisme.

La fraction attribuable dans la population (FAP) indique quelle proportion de l'ensemble des cas de la population pourrait être évitée si l'exposition disparaissait. Elle dépend à la fois du RR et de la diffusion de l'exposition (pe = proportion d'exposés). La formule de Levin est :

FAP = pe x (RR - 1) / [1 + pe x (RR - 1)]

Pour tabagisme-BPCO, pe = 300 / 1200 = 0,25, donc $\text{FAP} = \frac{0{,}25 \times 1{,}14}{1 + 0{,}25 \times 1{,}14} = \frac{0{,}286}{1{,}286} = 0{,}222$. Environ 22% de l'ensemble des cas de BPCO de cette population seraient évitables par l'élimination du tabagisme. De façon équivalente, FAP = (Rtotal - R0) / Rtotal = (0,100 - 0,0778) / 0,100 = 0,222, où Rtotal = 120 / 1200 = 0,100 est le risque dans la population totale. La différence Rtotal - R0 = 2,2 points de pourcentage s'appelle le risque attribuable dans la population. Ces calculs correspondent à ce que la fiche de la discipline appelle "l'estimation du risque au niveau de la population". Ils supposent que l'association est causale et qu'il n'existe pas de confusion ; sur des données observationnelles, ce sont des estimations optimistes du bénéfice de la prévention.

10.8.3 Le nombre de sujets à traiter (NST) et le nombre nécessaire pour produire un effet indésirable (NNH)

L'inverse de la différence absolue de risques a une interprétation clinique très intuitive. Dans une étude d'intervention avec un résultat favorable, le NST (angl. number needed to treat, NNT) est le nombre de patients qu'il faut traiter pour obtenir un résultat favorable de plus qu'avec le traitement de comparaison :

$$\text{NNT} = \frac{1}{|R_1 - R_0|}$$

et il s'arrondit toujours vers le haut, à l'entier suivant (on ne peut pas traiter 5,5 patients). Pour un facteur nocif, le même calcul s'appelle NNH (angl. number needed to harm) : combien de personnes doivent être exposées pour qu'apparaisse un cas supplémentaire. Pour tabagisme-BPCO, $\text{NNH} = \frac{1}{0{,}0889} = 11{,}25$, soit 12 : pour 12 fumeurs suivis pendant 5 ans, il apparaît un cas de BPCO de plus que chez les non-fumeurs.

Dans l'étude mg2026, à 6 mois, 16 des 54 patients du bras C (conseil, application et téléconsultation) ont perdu au moins 5% de leur poids, contre 6 des 53 du bras A (conseil standard). Les proportions sont de 29,6% et 11,3% ; RR = 2,62 (IC à 95% 1,11-6,17) ; la différence absolue est de 18,3 points de pourcentage (IC à 95% 3,4-33,2) ; $\text{NNT} = \frac{1}{0{,}183} = 5{,}5$, soit 6 (IC à 95% 3,0-29,1). Sur le plan clinique : pour 6 patients qui reçoivent le programme complet au lieu du conseil standard, un patient de plus atteint une perte de poids pertinente. L'intervalle du NST est large (de 3 à 29), car la différence de risques est estimée de façon imprécise avec seulement 107 patients ; la borne supérieure (29) correspond à la borne inférieure de la différence (3,4 points de pourcentage). L'OR pour la même comparaison est de 3,30, là encore plus grand que le RR, car le résultat n'est pas rare.

La figure 10.2 rassemble dans un diagramme en forêt (forest plot) les mesures d'association calculées dans ce chapitre. Chaque carré est l'estimation ponctuelle, et la ligne horizontale est l'intervalle de confiance à 95% ; la ligne verticale pointillée marque la valeur 1 (absence d'effet). Un intervalle qui franchit la ligne pointillée correspond à un résultat non significatif.

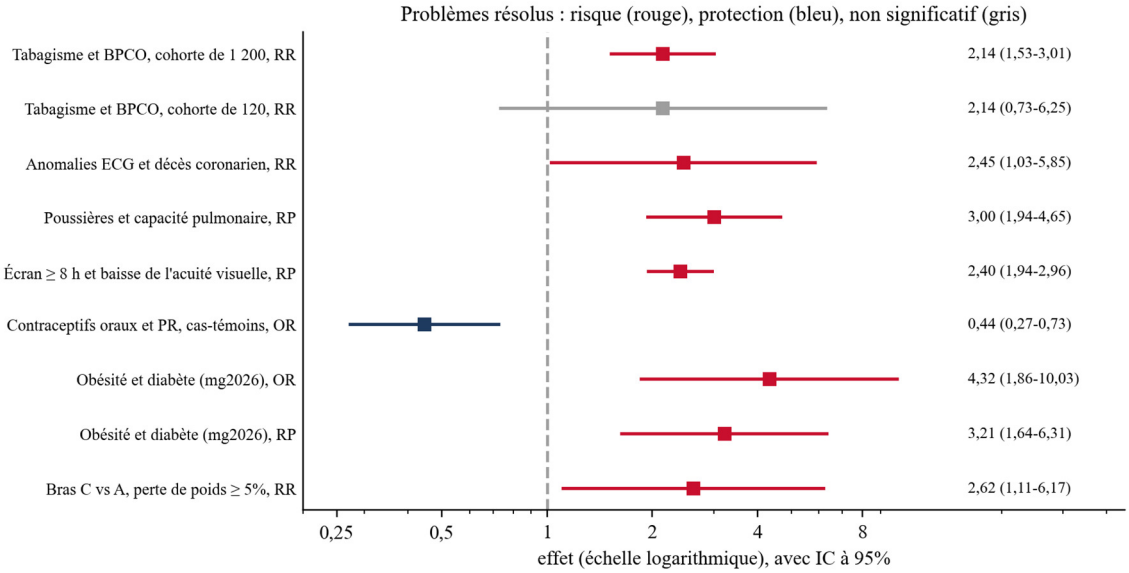


Figure 10.2. Les mesures d'association des exemples et des problèmes résolus du chapitre, avec l'IC à 95% calculé par la méthode logarithmique, sur un axe logarithmique. Rouge : facteur de risque significatif ; bleu : facteur protecteur significatif ; gris : intervalle contenant 1 (non significatif). Le même RR = 2,14 pour le tabagisme et la BPCO est significatif avec 1200 participants et non significatif avec 120. Exercices classiques de la discipline et jeu de données synthétique mg2026.

10.9 La confusion, la stratification et la modification de l'effet

10.9.1 Le facteur de confusion

Un facteur de confusion (angl. confounder) est une troisième variable qui remplit simultanément trois conditions : il est associé à l'exposition, il est un facteur de risque de la maladie (indépendamment de l'exposition) et il ne se trouve pas sur le chemin causal entre l'exposition et

la maladie. Lorsqu'il est présent, l'association brute entre l'exposition et la maladie mêle l'effet de l'exposition à celui du facteur de confusion.

L'exemple didactique classique est le lien apparent entre la consommation de café et une maladie, produit en réalité par la confusion due au tabagisme. Supposons que, parmi 1000 personnes, les fumeurs boivent beaucoup plus souvent du café (400 sur 500 fumeurs, contre 100 sur 500 non-fumeurs) et que le tabagisme augmente le risque de maladie (20% chez les fumeurs, 5% chez les non-fumeurs), alors que le café n'a aucun effet. La figure 10.3 montre ce qui se passe.

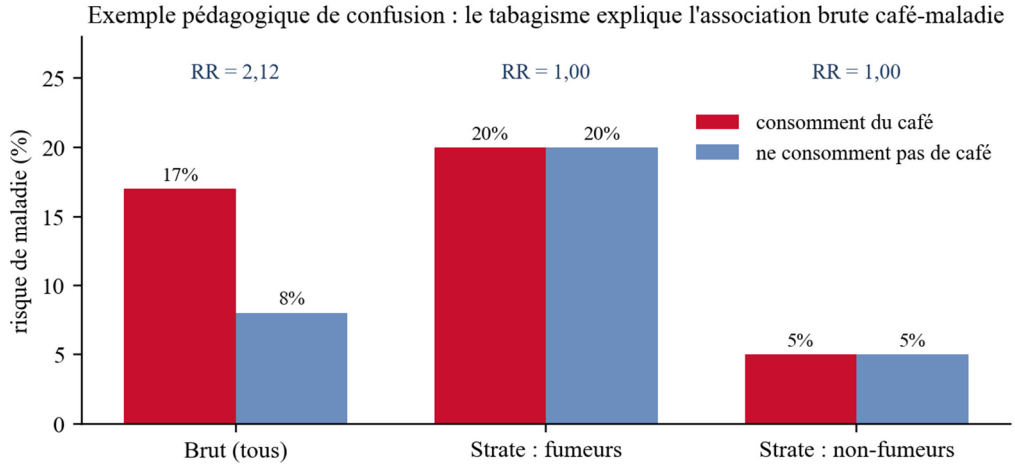


Figure 10.3. Exemple didactique de confusion (données construites). Dans l'analyse brute, les consommateurs de café ont un risque de 17%, contre 8% chez les autres (RR = 2,12). Dans chaque strate de tabagisme, le risque est identique chez les consommateurs et les non-consommateurs (20% chez les fumeurs, 5% chez les non-fumeurs ; RR = 1,00). L'association brute provient entièrement du fait que les fumeurs boivent davantage de café.

Analyse brute : parmi les 500 consommateurs de café, il y a 80 cas chez les fumeurs (20% de 400) et 5 chez les non-fumeurs (5% de 100), donc 85 cas, soit un risque de 17%. Parmi les 500 non-consommateurs, il y a 20 cas chez les fumeurs (20% de 100) et 20 chez les non-fumeurs (5% de 400), donc 40 cas, soit un risque de 8%. RR brut = 17% / 8% = 2,125, et OR brut = $\frac{85\times460}{415\times40}$ = 2,36. Un chercheur pressé conclurait que le café double le risque. L'analyse stratifiée sur le tabagisme montre pourtant RR = 1 dans chaque strate : chez les fumeurs, 20% contre 20% ; chez les non-fumeurs, 5% contre 5%. Le café n'a aucun effet ; le tabagisme, associé au café, crée l'illusion.

10.9.2 La stratification et l'estimateur de Mantel-Haenszel

La stratification consiste à répartir les données en sous-groupes (strates) selon les valeurs du facteur de confusion et à calculer la mesure d'association dans chaque strate. Si les mesures des strates sont semblables entre elles, mais diffèrent de la mesure brute, il y a confusion. Pour obtenir une seule estimation ajustée, on utilise la moyenne pondérée proposée par Mantel et Haenszel :

$$\mathrm{OR(MH)} = \frac{\sum\left(\frac{a_i \times d_i}{n_i}\right)}{\sum\left(\frac{b_i \times c_i}{n_i}\right)}$$

où l'indice i désigne la strate i, et ni est l'effectif total de la strate. Pour l'exemple du café, la strate des fumeurs a a = 80, b = 320, c = 20, d = 80, n = 500, et la strate des non-fumeurs a = 5, b = 95, c = 20, d = 380, n = 500. Le numérateur vaut $\frac{80\times80}{500} + \frac{5\times380}{500} = 12{,}8 + 3{,}8 = 16{,}6$; le dénominateur vaut $\frac{320\times20}{500} + \frac{95\times20}{500} = 12{,}8 + 3{,}8 = 16{,}6$. OR(MH) = 1,00, contre un OR brut de 2,36.

Sur mg2026, nous pouvons vérifier si le sexe est un facteur de confusion de l'association entre l'obésité et le diabète. Chez les femmes (n = 86) : a = 10, b = 19, c = 5, d = 52, OR = 5,47. Chez les hommes (n = 74) : a = 8, b = 17, c = 6, d = 43, OR = 3,37. $\mathrm{OR(MH)} = \frac{\frac{10\times52}{86}+\frac{8\times43}{74}}{\frac{19\times5}{86}+\frac{17\times6}{74}} = \frac{6{,}047+4{,}649}{1{,}105+1{,}378} = \frac{10{,}695}{2{,}483} = 4{,}31$, pratiquement identique à l'OR brut de 4,32. Le sexe n'est pas un facteur de confusion pour cette relation, ce qui concorde avec l'OR du sexe dans la régression logistique (0,97, chapitre 9).

La stratification fonctionne bien pour un ou deux facteurs comportant peu de catégories. Lorsque nous voulons ajuster simultanément sur plusieurs facteurs, dont certains numériques, nous utilisons la régression multiple (linéaire, logistique, de Cox), qui réalise, pour l'essentiel, la même opération au sein d'un modèle. C'est pourquoi les articles rapportent fréquemment un "OR brut" (ou "non ajusté") et un "OR ajusté" (ORa) ; leur comparaison montre dans quelle mesure l'ajustement a compté.

10.9.3 La modification de l'effet

Parfois, les mesures d'association des différentes strates diffèrent réellement entre elles : l'exposition a un effet plus important dans un sous-groupe que dans un autre. Ce phénomène s'appelle la modification de l'effet (ou interaction) ; ce n'est pas une erreur à corriger, mais une information à rapporter. Si un médicament réduit de 50% le risque d'accident vasculaire cérébral chez les personnes âgées et pas du tout chez les jeunes, une estimation commune masquerait un résultat clinique important. Dans de telles situations, on rapporte la mesure séparément pour chaque strate.

Les différences entre strates doivent cependant être testées, car sur de petits échantillons elles peuvent apparaître du seul fait des fluctuations d'échantillonnage. Sur mg2026, si nous stratifions l'association obésité-diabète selon l'âge (moins de 56 ans contre 56 ans et plus), nous obtenons OR = 11,67 chez les plus jeunes (seulement 6 cas de diabète) et OR = 3,99 chez les plus âgés. La différence paraît importante, mais le test d'homogénéité de Breslow-Day donne p = 0,37 : nous n'avons pas de preuve d'une modification réelle de l'effet, mais seulement des estimations imprécises dans une strate comportant très peu de cas. L'OR(MH) ajusté sur l'âge est de 5,02 (IC à 95% 2,04-12,35), proche de l'OR ajusté de la régression logistique (6,01). Une étude transversale portant sur 693 étudiants a testé formellement un terme d'interaction dans la régression logistique et a rapporté une modification de l'effet significative pour l'HbA1c (p = 0,023) [6] ; c'est la manière correcte d'affirmer qu'un effet diffère entre sous-groupes.

Étude de cas : ce que devient une association après ajustement

Quatre articles de collègues montrent, chiffres à l'appui, pourquoi l'association brute et l'association ajustée doivent être lues ensemble. Dans une cohorte prospective de 278 patients hospitalisés pour insuffisance cardiaque décompensée, suivis pendant 12 mois, le rapport hémoglobine/indice de distribution des globules rouges (médiane 0,89) a été associé à la survie sans événement dans l'analyse de Kaplan-Meier (log-rank p = 0,027), mais l'association n'était plus significative après ajustement sur l'âge, le sexe, le NT-proBNP, la fraction d'éjection et le débit de filtration glomérulaire dans un modèle de Cox (p = 0,240) [23]. L'événement principal est survenu chez 167 patients (60,1%). L'explication la plus probable est la confusion : un marqueur hématologique est associé à la sévérité de la maladie et, une fois la sévérité prise en compte par le NT-proBNP et la fraction d'éjection, l'information supplémentaire disparaît. Un résultat "perdu après ajustement" est un résultat précieux, et non un échec.

Dans une cohorte prospective de 133 patientes atteintes d'un carcinome séreux de l'ovaire de haut grade, les mutations BRCA (présentes chez 39,1% d'entre elles) ont été associées à un

risque de progression plus faible en analyse univariée : HR = 0,52, IC à 95% 0,27-0,99, p = 0,048 [117]. Le rapport des risques instantanés (hazard ratio, HR) se lit comme un RR : 0,52 signifie un risque instantané de progression inférieur de 48%. Après ajustement, HR = 0,57, p = 0,124. L'estimation ponctuelle a peu changé (de 0,52 à 0,57), mais l'intervalle s'est élargi et en est venu à contenir 1. Ici, le problème n'est pas nécessairement la confusion, mais la précision : avec 133 patientes et plusieurs variables d'ajustement, l'incertitude augmente. L'intervalle brut, qui touchait presque 1 (0,99), annonçait déjà la fragilité du résultat.

Dans une cohorte rétrospective de 213 patients pris en charge aux urgences pour suspicion d'infection bactérienne, les femmes étaient plus âgées que les hommes (médiane 77,0 contre 67,0 ans, p < 0,001), et l'origine urinaire de l'infection était plus fréquente chez les femmes (51,0% contre 34,5%, p = 0,015) [9]. Le résumé rapporte un OR = 2,14 (IC à 95% 1,20-3,81 ; p = 0,010), qui reste significatif après la correction de Benjamini-Hochberg pour tests multiples (p ajusté 0,039). Pour la mortalité hospitalière, 7 femmes (7,0%) et 9 hommes (8,0%), OR = 0,87 (IC à 95% 0,31-2,43) ; le résumé contient aussi un OR = 0,80 (IC à 95% 0,27-2,35 ; p = 0,690), également avec un intervalle large qui contient 1. La différence d'âge entre les sexes est exactement la situation dans laquelle l'âge peut introduire une confusion dans une comparaison brute entre femmes et hommes. Les intervalles très larges pour la mortalité (de 0,27 à 2,35) montrent qu'avec 16 décès l'étude ne peut exclure ni un effet protecteur important, ni un effet nocif important.

Une enquête transversale en ligne menée auprès de 753 jeunes adultes (93,5% de femmes) a rapporté les prévalences des comportements alimentaires avec des intervalles de Wilson et des OR ajustés sur l'âge et le sexe : ORa de 2,34, 2,49 et 2,73 pour différentes catégories d'IMC, tous avec p < 0,001 [20]. L'ajustement sur le sexe a cependant une portée limitée lorsque 93,5% des participants sont des femmes, et l'échantillon de convenance (qui choisit de répondre en ligne à un questionnaire sur l'alimentation) limite la généralisation, quelle que soit la justesse du calcul de l'OR.

10.10 De l'association à la causalité

10.10.1 Les erreurs systématiques des études observationnelles

Une association observée peut avoir quatre explications : le hasard, l'erreur systématique (angl. bias), la confusion et la causalité. Le hasard est évalué par les intervalles de confiance et les tests statistiques. La confusion a été discutée plus haut. Les erreurs systématiques se regroupent en deux grandes catégories. Le biais de sélection apparaît lorsque le mode de recrutement des participants fait que les groupes comparés diffèrent systématiquement : des témoins choisis dans une population différente de celle des cas, des volontaires en meilleure santé que la population générale, des pertes de vue liées à la fois à l'exposition et à la maladie. Le biais d'information (de mesure) apparaît lorsque l'exposition ou la maladie sont mesurées différemment dans les groupes comparés : le biais de mémorisation des études cas-témoins, un médecin qui recherche plus attentivement la maladie chez les patients qu'il sait exposés, une mauvaise classification de l'exposition. Contrairement à l'erreur aléatoire, l'erreur systématique ne diminue pas lorsque l'on augmente l'échantillon : une grande étude biaisée est précise autour d'une valeur fausse.

10.10.2 Les critères de Bradford Hill

En 1965, Austin Bradford Hill, qui avait démontré avec Richard Doll le lien entre le tabagisme et le cancer du poumon par des études cas-témoins et de cohorte, a proposé neuf aspects à prendre en considération avant de tirer une conclusion causale d'une association observationnelle [108]. Nous les appliquons à l'exemple tabagisme-BPCO.

La force de l'association : plus le RR est élevé, moins il est probable que l'association soit produite entièrement par un facteur de confusion inconnu (dans notre exemple, RR = 2,14). La constance : l'association a été observée dans plusieurs études, sur des populations différentes, avec des méthodes différentes. La spécificité : l'exposition conduit à une maladie donnée (le critère le plus faible, car le tabagisme provoque de nombreuses maladies). La temporalité : l'exposition précède la maladie ; c'est le seul critère indispensable, et il est garanti par le schéma de cohorte. Le gradient biologique : le risque augmente avec la dose (plus de paquets-années, risque plus élevé). La plausibilité biologique : il existe un mécanisme connu (l'inflammation chronique des voies aériennes). La cohérence : l'association ne contredit pas ce que l'on sait de l'histoire naturelle de la maladie. Les preuves expérimentales : l'arrêt du tabac ralentit le déclin de la fonction pulmonaire. L'analogie : d'autres expositions par inhalation (les poussières professionnelles) produisent des maladies semblables.

Bradford Hill ne les a pas présentés comme une liste à cocher, mais comme des points de vue qui, pesés ensemble, rendent une explication causale plus ou moins crédible. Une étude qui rapporte un OR ajusté avec un gradient dose-réponse, par exemple un OR de 1,43 pour chaque point d'un score de risque alimentaire (IC à 95% 1,23-1,66), apporte un argument plus solide qu'une étude qui ne rapporte qu'une comparaison entre exposés et non-exposés [4].

10.10.3 La rédaction des études observationnelles : STROBE

Pour que le lecteur puisse juger tous ces aspects, les études observationnelles doivent être rapportées de manière complète. Les recommandations STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) comprennent 22 points pour les études de cohorte, cas-témoins et transversales [25]. Parmi les exigences : préciser le schéma d'étude dans le titre ou le résumé, décrire les sources de participants et les critères de sélection (et, dans une étude cas-témoins, le mode de choix des témoins), définir clairement l'exposition, le critère de jugement et les facteurs de confusion, expliquer la manière dont les données manquantes ont été traitées, rapporter le nombre de participants à chaque étape, rapporter les estimations brutes et ajustées avec leurs intervalles de confiance et discuter les limites, y compris le sens probable des erreurs systématiques. Un résumé qui donne un OR sans intervalle de confiance ne respecte pas l'esprit de ces recommandations.

Étude de cas : des rapports de cotes dans les études des collègues, lus de façon critique

Une étude rétrospective comparative portant sur 1418 femmes enceintes (279 avec un diabète gestationnel, 74 avec un diabète de type 1, 107 avec un diabète de type 2 et 958 témoins) a analysé les facteurs de risque de complications, avec des OR et leurs IC à 95% [118]. L'âge a un OR = 1,03 (IC à 95% 1,01-1,05 ; p = 0,002), l'obésité un OR = 2,37 (IC à 95% 1,42-3,94 ; p < 0,001) et l'hypertension chronique un OR = 2,51 (IC à 95% 1,26-5,01 ; p = 0,009). Un OR de 1,03 paraît petit à côté de 2,37, mais la comparaison est trompeuse : l'OR de l'âge se rapporte à une année, et celui de l'obésité à la présence par rapport à l'absence d'une condition. Pour 10 ans, l'OR de l'âge devient 1,03^10 ≈ 1,34. Il faut toujours lire l'unité du prédicteur avant de comparer des OR.

Une étude transversale portant sur 921 étudiants a trouvé 24,1% de sujets en surpoids ou obèses et, pour un profil comportemental, un ORa = 7,20 (IC à 95% 4,26-12,17), ainsi qu'un ORa = 1,98 (IC à 95% 1,26-3,10) pour un autre facteur et un ORa de 1,43 pour chaque point d'un score de risque alimentaire (IC à 95% 1,23-1,66 ; p < 0,001) [4]. Un ORa de 7,20 indique une association forte, mais, avec une prévalence du résultat de 24%, il ne doit pas être traduit par "un risque 7 fois plus élevé" (tableau 10.4). Le schéma transversal laisse ouverte la question du sens : le comportement alimentaire a-t-il conduit au surpoids, ou les personnes en

surpoids ont-elles modifié leur comportement ? Dans la même population, une autre étude transversale portant sur 693 étudiants a rapporté une prévalence du prédiabète de 21,1% et un OR = 1,78 pour chaque catégorie d'un profil alimentaire ordonné (IC à 95% 1,38-2,30 ; p < 0,001), avec des prévalences de 15,4%, 20,0% et 27,3% dans les trois catégories (khi-deux p = 0,003) [6]. L'augmentation monotone de la prévalence d'une catégorie à l'autre est un gradient, l'un des arguments de Bradford Hill.

Une étude de cohorte rétrospective portant sur 486 patients atteints de COVID-19 (200 sans diabète, 62 avec un diabète de type 1 et 224 avec un diabète de type 2) a rapporté pour l'admission en réanimation un OR de 2,24 et pour la ventilation mécanique un OR de 2,46 [8]. Le résumé ne donne ni les intervalles de confiance ni les valeurs p de ces OR. En tant qu'exercice de lecture critique, la question naturelle est : quelle est la précision de ces estimations ? Avec 62 patients atteints de diabète de type 1, un OR de 2,24 pourrait avoir un intervalle large, incluant éventuellement 1 ; sans intervalle, le lecteur ne peut pas juger. C'est exactement la situation que les recommandations STROBE cherchent à prévenir.

Enfin, une analyse rétrospective des hospitalisations pour carcinome épidermoïde de la cavité buccale et de l'oropharynx entre 2018 et 2024 (248 hospitalisations) a utilisé une analyse de séries temporelles interrompues : le nombre mensuel moyen d'hospitalisations a diminué de 4,9 avant la pandémie à 2,6 pendant la pandémie et à 1,6 après, avec un rapport des taux d'incidence IRR = 0,973 (IC à 95% 0,950-0,996 ; p = 0,024) pour la tendance, et le curage ganglionnaire cervical a été associé à la période par un OR ajusté de 1,87 (IC à 95% 1,06-3,29 ; p = 0,030) [119]. L'IRR s'interprète comme un RR pour des taux : 0,973 par mois signifie une diminution d'environ 2,7% par mois. L'étude combine une approche écologique (taux par période) et une approche individuelle (régression logistique sur les patients).

10.11 Problèmes résolus

Tous les résultats numériques ci-dessous ont été calculés avec des logiciels et vérifiés à la main ; les IC du RR et de l'OR sont calculés par la méthode logarithmique (Katz pour le RR, Woolf pour l'OR), la même que celle utilisée par défaut par Epi Info et MedCalc.

Problème résolu 10.1. Le tabagisme et la BPCO : étude de cohorte portant sur 1200 participants

Énoncé. On inclut 1200 personnes initialement en bonne santé (fumeurs et non-fumeurs), suivies pendant 5 ans. Sur les 300 fumeurs, 50 développent une BPCO ; parmi les non-fumeurs, 70. (a) Formulez les hypothèses de recherche. (b) Formulez les hypothèses statistiques. (c) Construisez le tableau de contingence. (d) Le tabagisme est-il un facteur de risque ou un facteur protecteur ? Calculez le RR et l'OR. (e) Le risque est-il statistiquement significatif ? Quel test utilisons-nous ? (f) Quelle est l'interprétation clinique ?

(a) Hypothèse de recherche : le tabagisme augmente le risque de survenue d'une BPCO.

(b) Hypothèses statistiques : H0 : RR = 1, le risque de BPCO est le même chez les fumeurs et chez les non-fumeurs (le tabagisme n'est pas associé à la BPCO) ; H1 : RR ≠ 1, le risque diffère.

(c) Type d'étude : cohorte prospective (on part de l'exposition, chez des personnes en bonne santé, et on suit la survenue de la maladie). Le tableau est le tableau 10.3 : a = 50, b = 250, c = 70, d = 830, N = 1200.

(d) $R_1 = \frac{50}{300} = 16{,}7\%$; $R_0 = \frac{70}{900} = 7{,}8\%$. RR = 2,14 (IC à 95% 1,53-3,01, calcul à la section 10.7.1). $\text{OR} = \frac{50\times830}{250\times70} = 2{,}37(;\ \text{IC}\,95\%\ 1{,}61 - 3{,}50)$. RR > 1 et l'intervalle ne contient pas

1 : le tabagisme est un facteur de risque. S'agissant d'une étude de cohorte, la mesure principale est le RR ; l'OR n'est rapporté qu'à titre d'information complémentaire.

(e) Test du khi-deux (toutes les fréquences attendues sont bien supérieures à 5) : $\chi^2 = 19{,}75$, ddl 1, $p < 0{,}001$ (avec la correction de Yates 18,78, $p < 0{,}001$). On rejette H0 : l'association est statistiquement significative. La formulation "nous acceptons H1" est tolérée dans les cours, mais la formulation correcte est "nous rejetons H0".

(f) Interprétation clinique : au cours des 5 ans, les fumeurs ont eu un risque de BPCO environ 2,1 fois plus élevé que les non-fumeurs (plausiblement entre 1,5 et 3 fois). La différence absolue est de 8,9 points de pourcentage (IC à 95% 4,3-13,5), NNH = 12 : pour 12 fumeurs suivis pendant 5 ans, il apparaît un cas de BPCO supplémentaire. La fraction attribuable chez les exposés est de 53%, et la fraction attribuable dans la population de 22%. Attention au calcul erroné avec le dénominateur 1200 pour les non-fumeurs, qui donne un faux RR de 2,86.

Problème résolu 10.2. Le même risque, un échantillon dix fois plus petit

Énoncé. Supposons que seuls 120 participants soient inclus, dont 30 fumeurs. Tombent malades 5 fumeurs et 7 non-fumeurs. (a) Calculez le RR et l'OR. (b) Le risque est-il significatif ? (c) Justifiez la différence par rapport au problème 10.1.

Solution. Le tableau : a = 5, b = 25, c = 7, d = 83 (non-fumeurs 120 - 30 = 90 ; sujets sains 90 - 7 = 83).

(a) $R_1 = \frac{5}{30} = 16{,}7\%$; $R_0 = \frac{7}{90} = 7{,}8\%$; RR = 2,14, exactement comme dans le problème 10.1. OR $= \frac{5\times83}{25\times7} = \frac{415}{175} = 2{,}37$, également identique.

(b) $\mathrm{ES}[\ln(\mathrm{RR})] = \sqrt{\frac{1}{5} - \frac{1}{30} + \frac{1}{7} - \frac{1}{90}} = \sqrt{0{,}2 - 0{,}0333 + 0{,}1429 - 0{,}0111} = \sqrt{0{,}2984} = 0{,}546$. IC à 95% : $\exp(0{,}762 \pm 1{,}96 \times 0{,}546) = \exp(0{,}762 \pm 1{,}071)$, soit $e^{-0{,}309} = 0{,}73$ et $e^{1{,}833} = 6{,}25$. IC à 95% du RR : 0,73-6,25. $\mathrm{ES}[\ln(\mathrm{OR})] = \sqrt{\frac{1}{5} + \frac{1}{25} + \frac{1}{7} + \frac{1}{83}} = \sqrt{0{,}3950} = 0{,}628$; IC à 95% de l'OR : 0,69-8,13. Les deux intervalles contiennent 1. Les fréquences attendues sont 3, 27, 9 et 81 ; l'une est inférieure à 5, le test adapté est donc le test de Fisher : p = 0,17 (khi-deux non corrigé 1,98, p = 0,16). On ne rejette pas H0.

(c) Les mesures d'association sont identiques, mais avec 12 cas au total l'information est insuffisante pour distinguer un RR de 2 d'un RR de 1. La conclusion correcte n'est pas "le tabagisme n'est pas un facteur de risque", mais "l'étude n'a pas une puissance statistique suffisante pour démontrer l'association". L'absence de preuve n'est pas la preuve de l'absence. La figure 10.2 montre les deux intervalles l'un sous l'autre : la même estimation, des précisions très différentes.

Problème résolu 10.3. L'exposition à l'écran et l'acuité visuelle (étude transversale)

Énoncé. Dans une étude transversale portant sur 1000 personnes, 200 travaillent au moins 8 heures par jour devant un écran, dont 90 ont une acuité visuelle diminuée ; parmi les 800 qui travaillent moins longtemps, 150 ont une acuité diminuée. (a) Quel est le facteur de risque et quelle pourrait être son influence ? (b) Calculez la mesure d'association adaptée. (c) Quelle est la cote (odds) d'acuité diminuée chez ceux qui travaillent moins de 8 heures par jour ? (d) Existe-t-il une association significative ? (e) Calculez l'IC à 95% de l'OR.

Solution. (a) Le facteur de risque étudié est l'exposition prolongée à l'écran ; l'influence possible est une augmentation de la fréquence de l'acuité visuelle diminuée. Le tableau : a = 90, b = 110, c = 150, d = 650.

(b) L'étude est transversale, nous mesurons donc des prévalences : 90 / 200 = 45,0% chez les exposés et 150 / 800 = 18,75% chez les non-exposés. Le rapport de prévalences RP = $\frac{45{,}0}{18{,}75} = 2{,}40(;\ \text{IC}\,95\%\ 1{,}94 - 2{,}96)$. OR = $\frac{90\times650}{110\times150} = \frac{58\,500}{16\,500} = 3{,}55$.

(c) La cote chez les sujets exposés moins de 8 heures : 150 / 650 = 0,231 (environ 1 personne à l'acuité diminuée pour 4,3 à l'acuité normale). Chez les exposés, la cote est de 90 / 110 = 0,818. Si l'énoncé demande l'OR des sujets à exposition courte par rapport à ceux à exposition longue, c'est l'inverse : 1 / 3,55 = 0,28.

(d) χ^2 = 60,44, ddl 1, p < 0,001 ; l'association est significative.

(e) ln(3,545) = 1,266 ; ES = $\sqrt{\frac{1}{90}+\frac{1}{110}+\frac{1}{150}+\frac{1}{650}} = \sqrt{0{,}01111+0{,}00909+0{,}00667+0{,}00154} = \sqrt{0{,}02841} = 0{,}1685$; 1,266 ± 0,330 donne 0,935 et 1,596 ; IC à 95% de l'OR : 2,55-4,93.

Interprétation : la prévalence de l'acuité diminuée est 2,4 fois plus élevée chez ceux qui travaillent au moins 8 heures par jour devant un écran. L'OR (3,55) dépasse largement le RP (2,40) parce que l'affection est fréquente (24% au total) ; l'OR ne doit donc pas être lu comme un risque. Le schéma transversal ne permet pas de conclure quant au sens de la relation : il est possible, par exemple, que les personnes exerçant certaines professions présentent aussi d'autres caractéristiques (l'âge) associées à une acuité diminuée.

Problème résolu 10.4. Les contraceptifs oraux et la polyarthrite rhumatoïde (étude cas-témoins)

Énoncé. Un chercheur a sélectionné 100 femmes atteintes d'une polyarthrite rhumatoïde confirmée et 200 femmes traitées dans le même établissement pour d'autres maladies musculo-squelettiques. Les dossiers médicaux ont permis d'établir l'utilisation de contraceptifs oraux avant le diagnostic : 40 des cas et 120 des témoins en avaient utilisé. (a) De quel type d'étude s'agit-il ? (b) Calculez l'OR. (c) Calculez l'IC à 95% de l'OR. Peut-on calculer le RR ? (d) Existe-t-il une association significative ?

Solution. (a) Étude cas-témoins : la sélection s'est faite d'après la maladie (cas atteints de polyarthrite, témoins indemnes), et l'exposition a été recherchée dans le passé. Les témoins proviennent du même établissement, ce qui constitue un choix raisonnable. Le tableau : a = 40 (cas exposés), b = 120 (témoins exposés), c = 60 (cas non exposés), d = 80 (témoins non exposés).

(b) OR = $\frac{40\times80}{120\times60} = \frac{3200}{7200} = 0{,}44$. La cote d'avoir utilisé des contraceptifs chez les cas vaut 0,44 fois celle des témoins (40% des cas contre 60% des témoins ont été exposés).

(c) $\ln(0{,}444) = -0{,}811$; ES = $\sqrt{\frac{1}{40}+\frac{1}{120}+\frac{1}{60}+\frac{1}{80}} = \sqrt{0{,}025+0{,}00833+0{,}01667+0{,}0125} = \sqrt{0{,}0625} = 0{,}250$; -0,811 ± 0,490 donne -1,301 et -0,321 ; IC à 95% de l'OR : 0,27-0,73. Le RR ne peut pas être calculé : la proportion de malades dans l'échantillon (100 sur 300) a été fixée par le chercheur et ne reflète pas le risque dans la population. Un calcul mécanique du "RR" (40/160 contre 60/140, soit 0,58) n'a pas d'interprétation en termes de risque. Si la polyarthrite rhumatoïde est rare dans la population, l'OR = 0,44 donne néanmoins une approximation du RR.

(d) $\chi^2 = 10{,}71$, ddl 1, p = 0,001 ; l'intervalle de l'OR ne contient pas 1. L'association est significative. Comme OR < 1, l'utilisation de contraceptifs oraux apparaît comme un facteur protecteur dans cette étude. L'interprétation causale demande de la prudence : les témoins atteintes d'autres maladies musculo-squelettiques pourraient avoir, pour des raisons liées à leurs maladies, une consommation de contraceptifs différente de celle de la population générale (biais de sélection).

Problème résolu 10.5. Les anomalies de l'ECG et le décès d'origine coronarienne : un résultat limite

Énoncé. On a suivi pendant 20 ans 47 hommes de 40 à 64 ans présentant des anomalies de l'ECG à l'examen initial et 144 hommes du même âge sans anomalies. Sont décédés de maladie coronarienne 8 hommes du premier groupe et 10 du second. (a) Le type d'étude. (b) Testez l'hypothèse selon laquelle les anomalies de l'ECG sont un facteur de risque. (c) Calculez et interprétez le RR. (d) Calculez l'OR et son IC à 95%.

Solution. (a) Cohorte prospective : les groupes sont définis selon l'exposition (anomalies de l'ECG) et suivis pour le décès. Le tableau : a = 8, b = 39, c = 10, d = 134, N = 191.

(b) H0 : RR = 1 ; H1 : RR ≠ 1. Fréquences attendues : $\frac{a\,47\times18}{191} = 4{,}43$ (inférieure à 5), b 42,57, c 13,57, d 130,43. Khi-deux non corrigé : 4,22, p = 0,040 ; avec la correction de Yates : 3,12, p = 0,077 ; test exact de Fisher : p = 0,0485. Comme une fréquence attendue est inférieure à 5, le test recommandé est celui de Fisher, et p = 0,0485 se situe tout juste sous le seuil de 0,05.

(c) $R_1 = \frac{8}{47} = 17{,}0\%$; $R_0 = \frac{10}{144} = 6{,}9\%$; RR = 2,45. ES $= \sqrt{\frac{1}{8} - \frac{1}{47} + \frac{1}{10} - \frac{1}{144}} = \sqrt{0{,}1968} = 0{,}444$; IC 95% $= \exp(0{,}897 \pm 0{,}870) = 1{,}03 - 5{,}85$.

(d) OR $= \frac{8\times134}{39\times10} = \frac{1072}{390} = 2{,}75$; ES $= \sqrt{\frac{1}{8} + \frac{1}{39} + \frac{1}{10} + \frac{1}{134}} = \sqrt{0{,}2581} = 0{,}508$; IC 95% $= \exp(1{,}011 \pm 0{,}996) = 1{,}02 - 7{,}44$.

Interprétation : les anomalies de l'ECG sont associées à un risque de décès coronarien environ 2,5 fois plus élevé, mais la preuve est limite. Les intervalles de confiance commencent juste au-dessus de 1 (1,03 et 1,02), le test non corrigé et le test de Fisher sont à peine significatifs, et le test avec correction de Yates ne l'est pas. Avec 18 décès au total, une conclusion honnête est : "données compatibles avec un risque accru, mais imprécises ; le résultat doit être confirmé sur un échantillon plus grand". L'étudiant qui écrit seulement "p < 0,05, donc les anomalies de l'ECG sont un facteur de risque" perd la moitié de l'information.

Problème résolu 10.6. L'exposition aux poussières et l'âge : deux facteurs de risque

Énoncé. Partie I : on inclut 320 retraités de 70 à 75 ans, dont 80 ont travaillé au moins 5 ans dans un environnement poussiéreux. Aux examens, dans le groupe exposé comme dans le groupe non exposé, 30 personnes ont une capacité pulmonaire diminuée. Partie II : pour étudier l'âge comme facteur de risque indépendant, on examine aussi 200 personnes de 30 à 50 ans, non exposées aux poussières, dont 12 ont une capacité pulmonaire diminuée. Analysez les deux parties.

Solution, partie I. Type d'étude : transversale, avec une exposition professionnelle reconstituée à partir du passé (elle peut aussi être vue comme une cohorte rétrospective). Le tableau : a = 30, b = 50 (80 - 30), c = 30, d = 210 (240 - 30). Remarquez le piège de l'énoncé : "30 dans les deux groupes" ne signifie pas un risque égal, car les groupes ont des tailles différentes. $R_1 = \frac{30}{80} = 37{,}5\%$; $R_0 = \frac{30}{240} = 12{,}5\%$. RP = 3,00 (IC à 95% 1,94-4,65). OR $= \frac{30\times210}{50\times30} = \frac{6300}{1500} = 4{,}20$(; IC 95% $2{,}32 - 7{,}60$). $\chi^2 = 24{,}62$, p < 0,001. L'exposition aux poussières est associée à une

fréquence trois fois plus élevée de la diminution de la capacité pulmonaire ; la différence absolue est de 25 points de pourcentage, et la fraction attribuable chez les exposés $\frac{RR-1}{RR} = \frac{2}{3} = 67\%$.

Partie II. Pour isoler l'effet de l'âge, nous comparons uniquement des personnes non exposées, afin que l'exposition aux poussières n'introduise pas de confusion dans la comparaison : personnes âgées non exposées 30 / 240 = 12,5% contre jeunes adultes non exposés 12 / 200 = 6,0%. Le tableau : a = 30, b = 210, c = 12, d = 188. $RR = \frac{12,5}{6,0} = 2,08(;\ IC\,95\%\ 1,10 - 3,96)$; $OR = \frac{30\times188}{210\times12} = \frac{5640}{2520} = 2,24(;\ IC\,95\%\ 1,11 - 4,50)$; χ^2 = 5,34, p = 0,021. L'âge avancé est lui aussi un facteur de risque, multipliant le risque par environ deux. Si nous comparions les personnes âgées exposées aux jeunes non exposés, nous obtiendrions $RR = \frac{37,5}{6,0} = 6,25$, mais ce chiffre mêle l'effet de l'âge à celui des poussières. L'absence de jeunes exposés nous empêche de vérifier si l'effet des poussières est le même à tous les âges (modification de l'effet). La leçon de méthodologie : le choix du groupe de comparaison décide de la question à laquelle on répond.

Problème résolu 10.7. L'obésité et le diabète dans mg2026 : RP, OR et fraction attribuable

Énoncé. À l'inclusion dans mg2026, parmi les 54 patients obèses (IMC ≥ 30), 18 ont un diabète confirmé par HGPO ; parmi les 106 non obèses, 11. Calculez le RP, l'OR, leurs IC à 95%, le test de signification et la fraction attribuable dans la population. Commentez la différence entre l'OR et le RP.

Solution. Le tableau : a = 18, b = 36, c = 11, d = 95, N = 160. L'étude est transversale (analyse à l'inclusion), les prévalences sont donc de 18 / 54 = 33,3% chez les obèses et de 11 / 106 = 10,4% chez les non-obèses.

$RP = \frac{33,3}{10,4} = 3,21.$ $\quad ES[\ln(RP)] = \sqrt{\frac{1}{18} - \frac{1}{54} + \frac{1}{11} - \frac{1}{106}} = \sqrt{0,0556 - 0,0185 + 0,0909 - 0,0094} = \sqrt{0,1185} = 0,344;\ IC\,95\% = \exp(1,167 \pm 0,675) = 1,64 - 6,31.$

$OR = \frac{18\times95}{36\times11} = \frac{1710}{396} = 4,32.$ $\quad ES[\ln(OR)] = \sqrt{\frac{1}{18} + \frac{1}{36} + \frac{1}{11} + \frac{1}{95}} = \sqrt{0,1847} = 0,430;\ IC\,95\% = \exp(1,463 \pm 0,842) = 1,86 - 10,03.$

Les fréquences attendues sont 9,8 ; 44,2 ; 19,2 ; 86,8 (toutes supérieures à 5). χ^2 = 12,70, p < 0,001 (Yates 11,20, p = 0,0008 ; Fisher p = 0,0008). L'association est significative.

Fraction attribuable dans la population : pe = 54 / 160 = 0,3375 ; $FAP = \frac{0,3375\times2,21}{1+0,3375\times2,21} = \frac{0,746}{1,746} = 0,43$. De façon équivalente, (18,1% - 10,4%) / 18,1% = 0,43. Si l'association était causale et que l'obésité pouvait être éliminée, environ 43% des cas de diabète de cet échantillon seraient évités. La formulation au conditionnel est obligatoire : les données sont transversales et n'établissent pas la succession temporelle.

L'OR (4,32) est beaucoup plus grand que le RP (3,21) parce que le diabète est fréquent chez les obèses (un tiers). Rapporter l'OR comme "un risque plus de quatre fois plus élevé" serait exagéré. L'ajustement sur le sexe par la méthode de Mantel-Haenszel ne modifie pas l'OR (4,31), et la régression logistique ajustée sur l'âge et le sexe donne un OR de 6,01 (chap. 9). Sur la figure 10.2, l'intervalle de l'OR est plus large que celui du RP, comme d'habitude.

10.12 Exercices proposés

1. En 1985, on a étudié le lien entre le dispositif intra-utérin (DIU) et l'infertilité. Parmi 283 femmes infertiles, 89 avaient utilisé un DIU ; parmi 3833 femmes témoins, 640. (a) De quel type d'étude s'agit-il ? (b) Construisez le tableau 2 x 2. (c) Calculez l'OR avec son IC à 95% et interprétez-le. (d) Appliquez le test du khi-deux à $\alpha = 0{,}05$.
2. Dans une étude portant sur 1024 mères, 23 des 107 mères fumeuses et 117 des 917 mères non fumeuses ont eu des enfants atteints de trisomie 21, les cas étant identifiés à partir des certificats de naissance. Calculez l'OR et son IC à 95% et discutez si l'OR peut être assimilé au RR.
3. Dans l'étude mg2026, à 6 mois, 16 des 53 patients du bras B et 6 des 53 du bras A ont perdu au moins 5% de leur poids. Calculez le RR, la différence absolue de risques et le NST. Comparez avec le résultat C contre A (RR 2,62 ; NST 6).
4. Dans mg2026, parmi les 53 fumeurs, 10 ont un diabète confirmé par HGPO ; parmi les 107 non-fumeurs et anciens fumeurs, 19. Calculez le RP et l'OR avec leurs IC à 95% et formulez la conclusion statistique et la conclusion clinique.
5. Dans mg2026, parmi les 54 obèses, 21 ont une PAS $\geq$ 140 mmHg ; parmi les 106 non-obèses, 20. Calculez le RP, l'OR, leurs IC à 95% et le χ^2. Pourquoi l'OR diffère-t-il du RP ?
6. Une cohorte d'ouvriers a été suivie pour la dermatite de contact. Les sujets exposés à un solvant ont accumulé 400 personnes-années et 10 cas ; les non-exposés, 800 personnes-années et 8 cas. Calculez le taux d'incidence dans chaque groupe (pour 100 personnes-années) et le rapport des taux d'incidence.
7. Une maladie chronique a un taux d'incidence de 5 cas pour 1000 personnes-années et une durée moyenne de 4 ans. Estimez la prévalence. Comment la prévalence changerait-elle si un nouveau traitement doublait la durée moyenne de la maladie, sans la guérir ?
8. Un médicament produit un effet indésirable chez 6% des patients traités, contre 2% sous placebo. Calculez la différence de risques, le RR et le NNH. Que signifie le NNH sur le plan clinique ?
9. Une exposition a un RR = 3 et est présente chez 40% de la population. Calculez la fraction attribuable chez les exposés et la fraction attribuable dans la population.
10. Une étude est stratifiée selon un facteur de confusion. Strate 1 : a = 20, b = 80, c = 10, d = 60. Strate 2 : a = 30, b = 20, c = 60, d = 60. Calculez l'OR dans chaque strate, l'OR brut (à partir du tableau combiné) et l'OR de Mantel-Haenszel. Quelle conclusion en tirez-vous ?
11. Dans mg2026, parmi les 65 patients du milieu rural, 10 ont un diabète ; parmi les 95 du milieu urbain, 19. Calculez l'OR avec son IC à 95%. Est-il correct de dire que le milieu rural "protège" contre le diabète ?

Pièges fréquents

- Mal construire le tableau 2 x 2 : lignes et colonnes inversées ou un total placé à la place d'une cellule ; le risque chez les non-exposés se calcule uniquement parmi les non-exposés (tabagisme-BPCO : RR correct 2,14, et non 2,86).
- Calculer un RR dans une étude cas-témoins ; la proportion de malades y est fixée par le chercheur et l'on calcule seulement l'OR.
- Lire l'OR d'une maladie fréquente comme un risque relatif : l'OR exagère l'effet lorsque le résultat est fréquent (OR 4,32 contre RP 3,21 dans mg2026).

- Conclure que "le facteur n'a pas d'effet" à partir d'un résultat non significatif sur un petit échantillon ; la formulation correcte est "nous n'avons pas pu démontrer d'effet" (problème 10.2).
- Ignorer les fréquences attendues inférieures à 5 et appliquer mécaniquement le test du khi-deux au lieu du test de Fisher.
- Interpréter un OR par unité (par an, par point) comme s'il s'agissait d'un OR pour une variable binaire.
- Confondre association et causalité dans les études transversales, où la succession temporelle ne peut pas être établie.
- Rapporter un OR ou un RR sans intervalle de confiance.
- Confondre facteur de confusion et modification de l'effet : le premier se corrige par ajustement, la seconde se rapporte par strate.

À retenir

- Notation du tableau : a = exposés malades, b = exposés sains, c = non-exposés malades, d = non-exposés sains.
- $\mathrm{RR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$; $\mathrm{OR} = \frac{a \times d}{b \times c}$; RR > 1 risque, RR = 1 indifférent, RR < 1 protection.
- Cohorte et étude expérimentale : RR (et OR) ; étude transversale : RP (et OR) ; étude cas-témoins : OR seulement.
- OR ≈ RR seulement pour les maladies rares ; lorsque la maladie est fréquente, l'OR est plus éloigné de 1 que le RR.
- IC à 95% : $\mathrm{ES}[\ln(\mathrm{RR})] = \sqrt{\left[\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}\right]}$; $\mathrm{ES}[\ln(\mathrm{OR})] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$; un intervalle qui contient 1 signifie un résultat non significatif.
- Le risque attribuable (R1 - R0) mesure l'impact absolu ; NST ou $\mathrm{NNH} = \frac{1}{|R_1 - R_0|}$, arrondi vers le haut.
- Le facteur de confusion est associé à l'exposition et à la maladie ; on le contrôle par la randomisation, l'appariement, la stratification (Mantel-Haenszel) ou la régression multiple.
- Les critères de Bradford Hill permettent de peser les arguments de causalité ; la temporalité est le seul critère indispensable. Les études observationnelles sont rapportées selon les recommandations STROBE.
- Sur mg2026 : obésité-diabète RP 3,21 (1,64-6,31), OR 4,32 (1,86-10,03) ; bras C contre A : RR 2,62 (1,11-6,17), différence de 18,3 points de pourcentage, NST 6.

Questions d'autoévaluation

1. (QRU) Une étude sélectionne 150 patients atteints d'un cancer gastrique et 300 personnes sans cancer et recherche la consommation antérieure d'aliments fumés. La mesure d'association adaptée est :
    a) le risque relatif
    b) l'incidence cumulée
    c) le rapport de cotes
    d) le rapport de prévalences
    e) le taux d'incidence
2. (QRU) Dans une cohorte, le risque de maladie est de 20% chez les exposés et de 10% chez les non-exposés. Le nombre nécessaire pour produire un cas supplémentaire (NNH) est :
    a) 2
    b) 5
    c) 10

   d) 20
   e) 0,1

3. (QRU) Un RR = 0,70 avec un IC à 95% de 0,55-0,89 indique :
   a) un facteur de risque significatif
   b) un facteur protecteur significatif, avec une réduction du risque d'environ 30%
   c) un résultat non significatif, car le RR est inférieur à 1
   d) une réduction du risque de 70%
   e) un facteur indifférent

4. (QRU) Dans une étude transversale, la prévalence de la maladie est de 45% chez les exposés et de 18,75% chez les non-exposés. L'affirmation correcte est :
   a) l'OR sera approximativement égal au rapport de prévalences, car la maladie est rare
   b) le rapport de prévalences est de 2,40, et l'OR sera supérieur à 2,40
   c) l'étude permet d'établir la succession temporelle
   d) on peut calculer l'incidence cumulée
   e) l'OR sera inférieur au rapport de prévalences

5. (QRU) L'intervalle de confiance à 95% d'un OR = 2,37 est de 0,69-8,13. La conclusion correcte est :
   a) l'exposition double à coup sûr la cote de la maladie
   b) l'association est significative à $\alpha = 0,05$
   c) les données ne permettent pas de rejeter H0 ; l'étude n'a pas démontré d'association
   d) l'exposition est un facteur protecteur
   e) l'intervalle est faux, car il n'est pas symétrique autour de 2,37

6. (QRM) À propos de l'étude de cohorte, sont vraies les affirmations suivantes :
   a) elle part de personnes indemnes de la maladie, classées selon l'exposition
   b) elle permet le calcul direct de l'incidence et du risque relatif
   c) c'est le schéma le plus efficace pour les maladies très rares
   d) elle peut être prospective ou rétrospective
   e) elle est exposée aux pertes de vue

7. (QRM) À propos du rapport de cotes, sont vraies les affirmations suivantes :
   a) il se calcule comme (a x d) / (b x c)
   b) dans une étude cas-témoins, l'OR de l'exposition est égal à l'OR de la maladie
   c) pour les maladies rares, l'OR donne une approximation du risque relatif
   d) l'OR est toujours égal au RR
   e) un OR < 1 indique un facteur protecteur

8. (QRM) Un facteur de confusion :
   a) est associé à l'exposition
   b) est un facteur de risque de la maladie, indépendamment de l'exposition
   c) se trouve obligatoirement sur le chemin causal entre l'exposition et la maladie
   d) peut être contrôlé par la stratification et l'estimateur de Mantel-Haenszel
   e) peut être contrôlé au stade de la conception par la randomisation ou l'appariement

9. (QRM) À propos de l'étude cas-témoins, sont vraies les affirmations suivantes :
   a) elle est adaptée aux maladies rares et à celles à longue latence

b) elle est sensible au biais de mémorisation

c) les témoins doivent provenir de la même population que les cas

d) elle permet de calculer l'incidence de la maladie chez les exposés

e) l'appariement sur l'âge empêche d'étudier l'âge comme facteur de risque

10. (QRM) Parmi les critères de Bradford Hill pour l'évaluation de la causalité figurent :

a) la temporalité

b) le gradient biologique (la relation dose-réponse)

c) une valeur p inférieure à 0,05

d) la constance des résultats entre les études

e) la plausibilité biologique

Chapitre 11. Méthodes statistiques dans la décision médicale : l'analyse bayésienne et la médecine fondée sur les preuves

Objectifs

À la fin de ce chapitre, l'étudiant sera capable de :

- expliquer la différence entre la probabilité comprise comme fréquence et la probabilité comprise comme degré de croyance, et reconnaître les deux sens dans le raisonnement clinique ;
- calculer des probabilités inconditionnelles et conditionnelles à partir d'un tableau maladie-symptôme et appliquer le théorème de Bayes, y compris au moyen de la formule des probabilités totales ;
- transformer un problème de diagnostic en fréquences naturelles et construire l'arbre correspondant ;
- passer de la probabilité à la cote et inversement, et utiliser les rapports de vraisemblance pour obtenir la probabilité post-test, par le calcul à la main et avec le nomogramme de Fagan ;
- juger quand deux tests peuvent être combinés en série ou en parallèle et ce que suppose l'hypothèse d'indépendance conditionnelle ;
- résoudre un tableau de diagnostic différentiel à trois maladies et plusieurs signes cliniques ;
- décrire le seuil de test et le seuil de traitement et les calculer dans un exemple simple ;
- interpréter un intervalle de crédibilité et un facteur de Bayes (BF10) rapporté par JASP ;
- appliquer les cinq étapes de la médecine fondée sur les preuves, lire un diagramme en forêt (forest plot) et reconnaître les principaux biais cognitifs du raisonnement clinique.

11.1 L'incertitude dans la décision clinique

Une femme de 58 ans entre dans le cabinet de médecine générale pour une ordonnance. Elle a un IMC de 31, son père était diabétique, elle ne fait pas d'exercice. Elle ne se plaint de rien. Le médecin se demande si elle a un diabète non diagnostiqué. Il ne peut pas le savoir avec certitude sans test, mais il ne part pas non plus de zéro : il sait que, dans la population des adultes à risque cardiométabolique, environ une personne sur cinq ou six a un diabète non diagnostiqué, et le profil de la patiente lui fait penser que le risque est même plus élevé. À partir de ce moment, chaque nouvelle information (le score FINDRISC, la glycémie à jeun, l'HbA1c) déplace cette estimation vers le haut ou vers le bas.

Tel est, en bref, le sujet de ce chapitre. Le diagnostic n'est pas un acte de reconnaissance instantanée, mais une succession de révisions d'une probabilité. Les médecins expérimentés font ces révisions intuitivement, parfois très bien, parfois avec des erreurs systématiques que nous décrirons à la section 11.11. La statistique offre l'outil qui rend la révision explicite et vérifiable : le théorème de Bayes [19,120].

11.1.1 Du symptôme au diagnostic : un problème de probabilités

Dans les chapitres précédents, nous avons utilisé la probabilité surtout pour décrire les fluctuations d'échantillonnage : quelle est la probabilité d'obtenir une différence aussi grande entre deux moyennes si H0 était vraie. Dans la décision clinique, la question change. Le médecin n'a pas besoin de la probabilité des données dans l'hypothèse où le patient est en bonne santé, mais de

la probabilité que le patient soit malade, compte tenu des données. Ces deux probabilités sont différentes, et les confondre est l'une des plus anciennes erreurs de raisonnement en médecine.

Un exemple classique : 80% des patients atteints de pneumonie ont de la fièvre. Cela ne signifie pas que 80% des patients fébriles ont une pneumonie. La première affirmation décrit p(fièvre / pneumonie), la seconde décrirait p(pneumonie / fièvre). Dans un cabinet de médecine générale, en hiver, la plupart des patients fébriles ont des viroses respiratoires banales ; la seconde probabilité est donc beaucoup plus faible que la première. Le théorème de Bayes est précisément la formule qui relie ces deux probabilités et nous dit quelle information supplémentaire est nécessaire pour passer de l'une à l'autre : la fréquence de la maladie dans la population dont provient le patient.

11.1.2 La probabilité comme fréquence et la probabilité comme degré de croyance

Il existe deux manières de comprendre le mot "probabilité". Au sens fréquentiste, la probabilité d'un événement est la limite de la fréquence relative de cet événement sur un très grand nombre de répétitions de la même expérience. Lorsque nous disons que la prévalence du diabète non diagnostiqué dans le jeu de données mg2026 est de 18,1% (29 sur 160), nous décrivons une fréquence observée, qui estime une fréquence dans la population.

Au sens subjectif ou bayésien, la probabilité exprime le degré de croyance d'une personne dans une affirmation, sur la base de l'information dont elle dispose. Lorsque le médecin dit "je pense qu'il y a une probabilité de 30% que ce patient ait un diabète", il ne se réfère pas à une série de répétitions du même patient (le patient a ou n'a pas de diabète), mais à son degré de conviction. Cette probabilité peut changer à mesure qu'apparaissent de nouvelles informations, et le théorème de Bayes indique comment elle doit changer pour rester cohérente.

En pratique, les deux sens se rejoignent. Il est souhaitable que le degré de croyance du médecin avant le test (la probabilité pré-test) soit ancré dans une fréquence : la prévalence de la maladie chez des patients semblables. Et la performance des tests (sensibilité, spécificité) est estimée de façon fréquentiste, à partir d'études d'exactitude diagnostique, comme nous le verrons au chapitre 12. L'analyse bayésienne dans la décision clinique utilise donc des fréquences pour mettre à jour des degrés de croyance [63,64].

11.2 La probabilité conditionnelle revisitée

11.2.1 Le tableau maladie-symptôme

Le cours de la discipline part d'un tableau de contingence dans lequel figurent en lignes la présence (B+) ou l'absence (B-) de la maladie et en colonnes la présence (S+) ou l'absence (S-) du symptôme. Le même tableau peut être construit pour n'importe quelle paire maladie-symptôme, et dans un système expert de diagnostic on construit des dizaines ou des centaines de tableaux de ce type.

Tableau 11.1. Le tableau maladie-symptôme, notations du cours.

	S+	**S-**	**Total**
B+	a	b	a + b
B-	c	d	c + d
Total	a + c	b + d	$N = a + b + c + d$

À partir de ce tableau, on calcule deux types de probabilités. Les probabilités inconditionnelles (marginales) se rapportent à l'ensemble de l'échantillon :

$$p(B^+) = \frac{a + b}{N}$$

$$p(S^+) = \frac{a+c}{N}$$

Les probabilités conditionnelles se calculent à l'intérieur d'une ligne ou d'une colonne. Par exemple, la probabilité de présenter le symptôme parmi les sujets indemnes de la maladie est

$$p(S^+/B^-) = \frac{c}{c+d}$$

et la probabilité d'avoir la maladie parmi les sujets présentant le symptôme est

$$p(B^+/S^+) = \frac{a}{a+c}$$

La barre oblique se lit "sachant" ou "conditionnellement à". Tout ce qui se trouve après la barre restreint la population à laquelle nous nous référons. C'est là qu'apparaît la première règle pratique : pour savoir quel dénominateur utiliser, regardez après la barre. Si après la barre figure B-, le dénominateur est le total de la ligne B-, c'est-à-dire c + d.

Prenons un exemple chiffré. Dans un cabinet de médecine générale, on a évalué, au cours d'un hiver, 200 adultes présentant une toux aiguë. Chez 30 d'entre eux, la radiographie a confirmé une pneumonie. Une fièvre supérieure à 38 °C était présente chez 24 des patients atteints de pneumonie et chez 34 des patients sans pneumonie (données fictives).

Tableau 11.2. Pneumonie (B) et fièvre (S) chez 200 patients présentant une toux aiguë (données fictives).

	S+ (fièvre)	**S-**	**Total**
B+ (pneumonie)	24	6	30
B-	34	136	170
Total	58	142	200

Nous calculons :

$$p(B^+) = \frac{30}{200} = 0{,}15$$

$$p(S^+) = \frac{58}{200} = 0{,}29$$

$$p(S^+/B^+) = \frac{24}{30} = 0{,}80$$

$$p(S^+/B^-) = \frac{34}{170} = 0{,}20$$

$$p(B^+/S^+) = \frac{24}{58} = 0{,}414$$

Remarquez l'asymétrie : la fièvre apparaît chez 80% des patients atteints de pneumonie, mais seulement 41,4% des patients fébriles ont une pneumonie. Les 58,6% restants sont des patients fébriles sans pneumonie qui, beaucoup plus nombreux parmi les patients qui toussent, "diluent" le groupe des patients fébriles.

11.2.2 La règle de multiplication et la formule des probabilités totales

La probabilité que deux événements surviennent simultanément s'obtient par la règle de multiplication :

$$p(B^+ \text{ et } S^+) = p(S^+/B^+) \times p(B^+) = p(B^+/S^+) \times p(S^+)$$

Dans notre exemple, $p(B^+ \text{ et } S^+) = \frac{24}{200} = 0{,}12$. Vérifions : 0,80 x 0,15 = 0,12 et 0,414 x 0,29 = 0,12. Les deux écritures sont égales parce qu'elles décrivent la même cellule du tableau, à savoir la cellule a.

Le second outil est la formule des probabilités totales. Le symptôme peut apparaître soit chez un malade, soit chez un non-malade, et les deux situations s'excluent mutuellement. Par conséquent :

$$p(S^+) = p(S^+/B^+) \times p(B^+) + p(S^+/B^-) \times p(B^-)$$

Pour le tableau 11.2 : $p(S^+) = 0{,}80 \times 0{,}15 + 0{,}20 \times 0{,}85 = 0{,}12 + 0{,}17 = 0{,}29$, exactement ce que nous avons obtenu à partir de la marge du tableau. La formule paraît banale lorsque nous disposons du tableau complet, mais elle devient indispensable lorsque nous ne l'avons pas : la plupart du temps, le médecin connaît la sensibilité d'un test (d'après la littérature) et la prévalence de la maladie (d'après son expérience ou des études épidémiologiques), mais il ne dispose pas d'un tableau de ses propres patients.

S'il y a plusieurs maladies possibles, par exemple B1, B2 et B3, mutuellement exclusives et couvrant toutes les possibilités, la formule se généralise naturellement :

$$p(S) = p(S/B_1) \times p(B_1) + p(S/B_2) \times p(B_2) + p(S/B_3) \times p(B_3)$$

Nous utiliserons cette forme à la section 11.6, pour le diagnostic différentiel.

11.3 Le théorème de Bayes

11.3.1 La démonstration

Nous partons des deux écritures de la règle de multiplication pour la même cellule :

$$p(B/S) \times p(S) = p(S/B) \times p(B)$$

Nous divisons les deux membres par p(S) et obtenons le théorème de Bayes, sous la forme présentée dans le cours de la discipline :

$$p(B/S) = \frac{p(S/B) \times p(B)}{p(S)}$$

En remplaçant le dénominateur par la formule des probabilités totales, nous obtenons la forme que nous utiliserons le plus souvent :

$$p(B^+/S^+) = \frac{p(S^+/B^+) \times p(B^+)}{p(S^+/B^+) \times p(B^+) + p(S^+/B^-) \times p(B^-)}$$

Chaque terme a un nom. p(B+) est la probabilité a priori ou pré-test : la probabilité de la maladie avant que nous sachions quoi que ce soit sur le symptôme ou le test. p(S+ / B+) est la vraisemblance du symptôme chez les malades ; lorsque S est le résultat positif d'un test, c'est la sensibilité. p(S+ / B-) est la fréquence du symptôme chez les non-malades ; pour un test, c'est 1 - spécificité, c'est-à-dire le taux de faux positifs. p(B+ / S+) est la probabilité a posteriori ou post-test ; pour un test positif, c'est la valeur prédictive positive.

Appliqué au tableau 11.2, sans regarder la cellule a :

$$p(B^+/S^+) = \frac{0{,}80 \times 0{,}15}{0{,}80 \times 0{,}15 + 0{,}20 \times 0{,}85} = \frac{0{,}12}{0{,}29} = 0{,}414$$

Nous retrouvons la valeur calculée directement. La force de la formule apparaît lorsque nous ne changeons que la prévalence. Supposons que la même fièvre, avec la même fréquence chez les malades et les non-malades, apparaisse chez des patients présentant une toux, hospitalisés dans un service de pneumologie, où la pneumonie a une prévalence de 50%. Alors $p(B^+/S^+) = \frac{0{,}80 \times 0{,}50}{0{,}80 \times 0{,}50 + 0{,}20 \times 0{,}50} = \frac{0{,}40}{0{,}50} = 0{,}80$. Le même signe clinique a une valeur prédictive de 41,4% au cabinet et de 80% à l'hôpital.

11.3.2 Les fréquences naturelles

La formule de Bayes est correcte, mais peu intuitive. Les études sur le raisonnement des médecins ont montré à plusieurs reprises que beaucoup d'entre eux surestiment fortement la valeur prédictive positive lorsque le problème leur est présenté sous forme de pourcentages. Le même problème devient beaucoup plus facile s'il est formulé en fréquences naturelles : des nombres entiers de patients appartenant à un groupe imaginaire [120].

La figure 11.1 montre la méthode pour un test de sensibilité 80% et de spécificité 90%, appliqué dans une population où la prévalence de la maladie est de 10%. Prenons 1000 patients. Une prévalence de 10% signifie 100 malades et 900 non-malades. Parmi les 100 malades, le test est positif chez 80 (sensibilité de 80%) et négatif chez 20. Parmi les 900 non-malades, une spécificité de 90% signifie 810 tests négatifs et 90 faux positifs. Au total, nous avons 80 + 90 = 170 tests positifs, dont seulement 80 appartiennent à des malades.

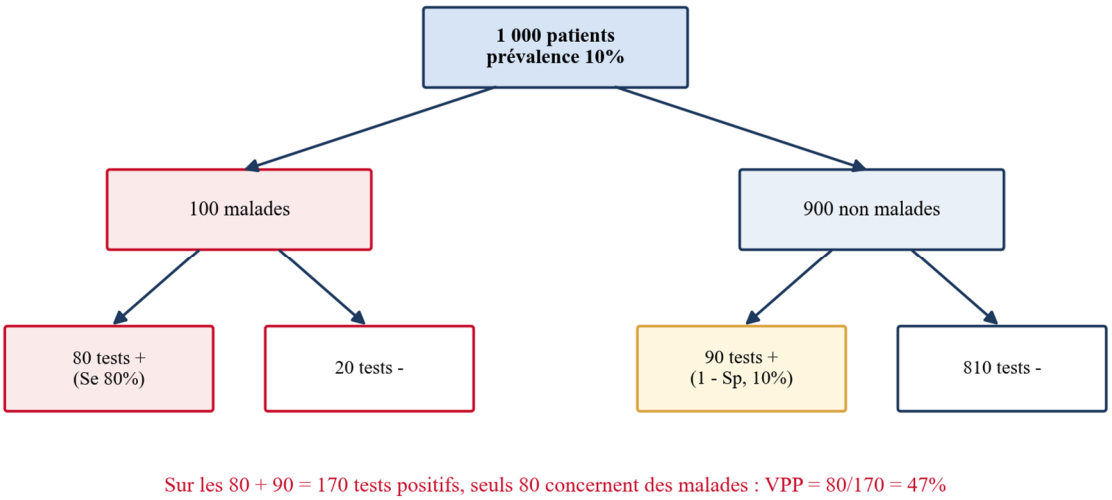


Figure 11.1. Arbre des fréquences naturelles pour un test de sensibilité 80% et de spécificité 90%, appliqué à 1000 patients avec une prévalence de la maladie de 10%. Sur 170 tests positifs, seuls 80 proviennent de malades ; la valeur prédictive positive est donc de 80/170 = 47%. Exemple didactique.

La valeur prédictive positive est de 80/170 = 47,1%. La valeur prédictive négative est de 810/(810 + 20) = 810/830 = 97,6%. Un test qui "se trompe" chez seulement 10% des sujets sains produit, dans une population où la maladie est rare, plus de faux positifs que de vrais positifs. L'explication est purement arithmétique : 10% de 900 (90) est plus que 80% de 100 (80).

En travaux pratiques, je recommande toujours la même chose : lorsque vous recevez un problème exprimé en pourcentages, réécrivez-le pour 1000 ou 10 000 patients et dessinez l'arbre. Cela prend deux minutes et élimine presque toutes les erreurs de raisonnement.

11.3.3 La probabilité pré-test : la prévalence et le contexte clinique

La probabilité pré-test n'est pas une propriété de la maladie, mais du patient, dans le contexte où il est vu. Un même diagnostic a des probabilités pré-test très différentes en médecine générale, en consultation spécialisée et à l'hôpital. Le médecin généraliste voit des patients non filtrés, avec des symptômes vagues et des maladies à des stades précoces ; la prévalence des maladies graves y est faible. À l'hôpital arrivent des patients déjà sélectionnés par d'autres médecins, si bien que la prévalence y est beaucoup plus élevée. C'est la raison pour laquelle un test aux performances excellentes dans une clinique universitaire peut générer de nombreux faux positifs lorsqu'il est transposé en soins primaires.

La probabilité pré-test s'estime à partir de plusieurs sources : la prévalence de la maladie dans la population dont provient le patient (données épidémiologiques, registres, études de prévalence), les données cliniques déjà connues (âge, sexe, facteurs de risque, symptômes, examen clinique), des scores cliniques validés et, inévitablement, l'expérience du médecin. Dans le jeu de données mg2026, la prévalence du diabète confirmé par HGPO est de 18,1% (IC à 95% de Wilson 12,9-24,8), et c'est la probabilité pré-test raisonnable pour un adulte à risque cardiométabolique d'un cabinet semblable, avant toute autre information. Pour un adulte quelconque, sans facteurs de risque, la probabilité serait beaucoup plus faible ; pour un patient obèse de plus de 60 ans, avec des antécédents familiaux, elle serait plus élevée. L'analyse transversale du jeu de données mg2026 montre, par exemple, une prévalence de 33,3% chez les obèses contre 10,4% chez les non-obèses.

Étude de cas : le dépistage du cancer colorectal et le raisonnement bayésien [101]

Iovanescu et ses collaborateurs (avec Mirela Frandeș, de notre discipline) ont évalué la fiabilité de la combinaison du test immunochimique fécal (FIT) et de la sigmoïdoscopie flexible dans le dépistage des néoplasies colorectales. Ont été inclus 2201 patients symptomatiques consécutifs, adressés pour un FIT ; un échantillon de 252 personnes a ensuite bénéficié d'une coloscopie (le standard de référence), et une sigmoïdoscopie réelle a été réalisée sur un sous-échantillon de 57 patients. La coloscopie a mis en évidence une néoplasie avancée chez 75 personnes (29,76%), et le FIT était positif dans 30,3% des cas de néoplasie avancée. Le taux d'adhésion à l'orientation vers la coloscopie a été de 52% parmi les sujets ayant un FIT positif, et la comparaison appariée des deux méthodes a été faite par le test de McNemar ($P < 0,001$). Les auteurs ont calculé les probabilités a priori et a posteriori et ont montré que, pour de faibles niveaux de prévalence, les valeurs prédictives négatives se situaient entre 81,5% et 90,12%. Que nous apprend cette étude ? Premièrement, que la prévalence de près de 30% dans l'échantillon coloscopié n'est pas la prévalence de la population de dépistage : les patients étaient symptomatiques et sélectionnés. Deuxièmement, que les valeurs prédictives doivent être recalculées pour la prévalence de la population dans laquelle le test sera utilisé, précisément au moyen du théorème de Bayes. Troisièmement, que les tests en série (FIT, puis sigmoïdoscopie) et en parallèle modifient le compromis entre sensibilité et spécificité, sujet que nous reprenons à la section 11.5. Enfin, le fait que seule la moitié des patients ayant un FIT positif aient accepté la coloscopie soulève le problème du biais de vérification, discuté au chapitre 12.

11.4 Cotes, rapports de vraisemblance et forme de Bayes avec les cotes

11.4.1 La conversion probabilité-cote

La cote (angl. odds) d'un événement est le rapport entre la probabilité que l'événement se produise et la probabilité qu'il ne se produise pas :

$$\text{cote} = \frac{p}{1 - p}$$

Inversement, la probabilité s'obtient à partir de la cote :

$$p = \frac{\text{cote}}{1 + \text{cote}}$$

Une probabilité de 0,20 correspond à une cote de 0,20/0,80 = 0,25, c'est-à-dire "1 contre 4". Une probabilité de 0,50 correspond à une cote de 1 ("un contre un"), et une probabilité de 0,90 à une cote de 9. Pour les petites probabilités, la cote et la probabilité sont presque égales (p = 0,02 donne une cote de 0,0204), mais à mesure que la probabilité augmente, la cote augmente

beaucoup plus vite et n'a pas de limite supérieure. Vous avez rencontré les cotes dans le chapitre sur l'analyse du risque, avec le rapport de cotes (OR) ; ici, nous les utilisons pour un seul patient.

Tableau 11.3. Correspondance entre probabilité et cote.

Probabilité	Cote	Lecture
0,05	0,053	environ 1 contre 19
0,10	0,111	1 contre 9
0,181	0,221	environ 1 contre 4,5
0,25	0,333	1 contre 3
0,50	1	1 contre 1
0,75	3	3 contre 1
0,90	9	9 contre 1

11.4.2 Les rapports de vraisemblance

Le rapport de vraisemblance (RV, angl. likelihood ratio) d'un résultat de test indique combien de fois ce résultat est plus probable chez un malade que chez un non-malade [121]. Pour un test dichotomique, nous avons deux rapports :

$$\mathrm{RV}^{+} = \frac{p(T^{+}/B^{+})}{p(T^{+}/B^{-})} = \frac{\mathrm{Se}}{1-\mathrm{Sp}}$$

$$\mathrm{RV}^{-} = \frac{p(T^{-}/B^{+})}{p(T^{-}/B^{-})} = \frac{1-\mathrm{Se}}{\mathrm{Sp}}$$

Un RV égal à 1 signifie que le résultat est aussi fréquent chez les malades que chez les non-malades et n'apporte donc aucune information. Un RV+ élevé (supérieur à 10) rend la maladie beaucoup plus probable ; un RV- faible (inférieur à 0,1) la rend beaucoup moins probable. Entre 2 et 5 pour le RV+, et entre 0,2 et 0,5 pour le RV-, le test déplace modérément la probabilité, parfois suffisamment, parfois non. Ces repères sont indicatifs et sont utilisés par de nombreux auteurs [120,121].

L'avantage majeur des rapports de vraisemblance par rapport aux valeurs prédictives est que, comme la sensibilité et la spécificité dont ils proviennent, ils ne dépendent pas (en principe) de la prévalence. Nous pouvons donc les transposer d'une étude publiée à notre cabinet et les appliquer à la probabilité pré-test de notre patient. La réserve "en principe" compte : si le spectre des patients diffère beaucoup (stades plus légers de la maladie, comorbidités), la sensibilité et la spécificité changent elles aussi, et donc le RV. Nous verrons dans l'étude de cas de la section 11.8 un exemple tiré de la recherche des collègues, dans lequel la spécificité d'un biomarqueur change beaucoup selon le moment du prélèvement.

Pour le score FINDRISC dans le jeu de données mg2026, au seuil ≥ 12, la sensibilité est de 75,9% (22 sur 29) et la spécificité de 74,8% (98 sur 131). Alors :

$$\mathrm{RV}^{+} = \frac{0{,}759}{1-0{,}748} = \frac{0{,}759}{0{,}252} = 3{,}01$$

$$\mathrm{RV}^{-} = \frac{1-0{,}759}{0{,}748} = \frac{0{,}241}{0{,}748} = 0{,}32$$

Un score d'au moins 12 est 3 fois plus fréquent chez les personnes diabétiques que chez les personnes non diabétiques ; un score inférieur à 12 est environ 3 fois plus rare chez les personnes diabétiques.

11.4.3 Le théorème de Bayes sous la forme des cotes

Si nous écrivons le théorème de Bayes une fois pour B+ et une fois pour B-, avec le même résultat T du test, et que nous divisons les deux relations, le dénominateur p(T) se simplifie et il reste une formule remarquablement simple :

$$\text{cote post-test} = \text{cote pré-test} \times \text{RV}$$

L'algorithme de calcul comporte trois étapes : (1) vous transformez la probabilité pré-test en cote ; (2) vous multipliez par le RV du résultat obtenu ; (3) vous retransformez la cote post-test en probabilité.

Pour une personne du jeu de données mg2026 ayant un FINDRISC ≥ 12, la probabilité pré-test est de 0,181. La cote pré-test est de 0,181/0,819 = 0,221. La cote post-test est de 0,221 x 3,01 = 0,667. La probabilité post-test est de 0,667/1,667 = 0,40, soit 40,0%. C'est exactement la valeur prédictive positive calculée directement à partir du tableau, 22/55 = 40,0%, comme on pouvait s'y attendre. Pour une personne ayant un FINDRISC inférieur à 12 : 0,221 x 0,32 = 0,071, donc la probabilité post-test est de 0,071/1,071 = 6,7%.

La figure 11.2 montre la relation entre la probabilité pré-test et la probabilité post-test pour quelques valeurs du RV. La ligne diagonale (RV = 1) correspond au test inutile. Les courbes à RV élevé s'écartent rapidement de la diagonale vers le haut, surtout dans la zone des probabilités pré-test faibles et moyennes ; les courbes à RV faible descendent nettement sous la diagonale, surtout pour les probabilités pré-test élevées.

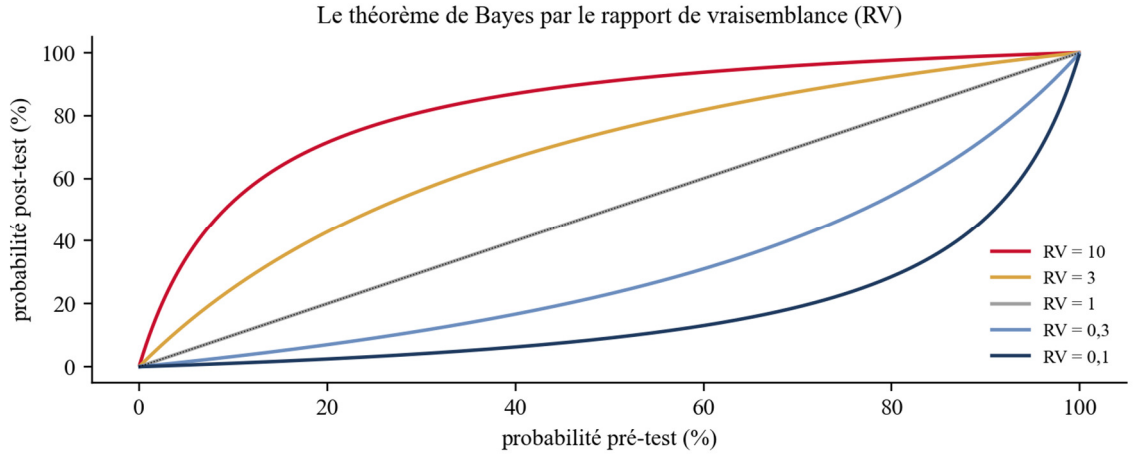


Figure 11.2. Probabilité post-test en fonction de la probabilité pré-test, pour des rapports de vraisemblance compris entre 0,1 et 10. Le même test déplace beaucoup plus la probabilité dans la zone intermédiaire qu'aux extrêmes. Exemple didactique.

La figure montre aussi quelque chose d'important pour la pratique quotidienne : un test apporte le plus lorsque la probabilité pré-test est intermédiaire. Si le patient a une probabilité pré-test de 2%, même un RV+ de 10 ne porte pas la probabilité beaucoup au-delà de 17% ; s'il a 95%, un test négatif avec un RV- de 0,1 laisse encore une probabilité d'environ 66%. C'est pourquoi les examens prescrits sans question clinique claire et sans estimation de la probabilité pré-test produisent souvent plus de confusion que d'information.

11.4.4 Le nomogramme de Fagan

Avant que chaque médecin ait une calculatrice en poche, Fagan a proposé un nomogramme qui effectue le calcul graphiquement [122]. Il comporte trois axes verticaux : à gauche la probabilité pré-test, au milieu le rapport de vraisemblance (sur une échelle logarithmique), à droite la probabilité post-test. On trace une ligne droite depuis le point pré-test, à travers la valeur du RV, et on lit la probabilité post-test là où la ligne rencontre l'axe de droite. La construction fonctionne parce que le logarithme de la cote post-test est la somme du logarithme de la cote pré-test et du logarithme du RV, et qu'une somme de logarithmes peut se représenter avec une règle droite.

La figure 11.3 montre le nomogramme avec l'exemple du FINDRISC : à partir de 18,1%, la ligne qui passe par RV+ = 3,01 aboutit à environ 40%, et celle qui passe par RV- = 0,32 aboutit à environ 6,7%.

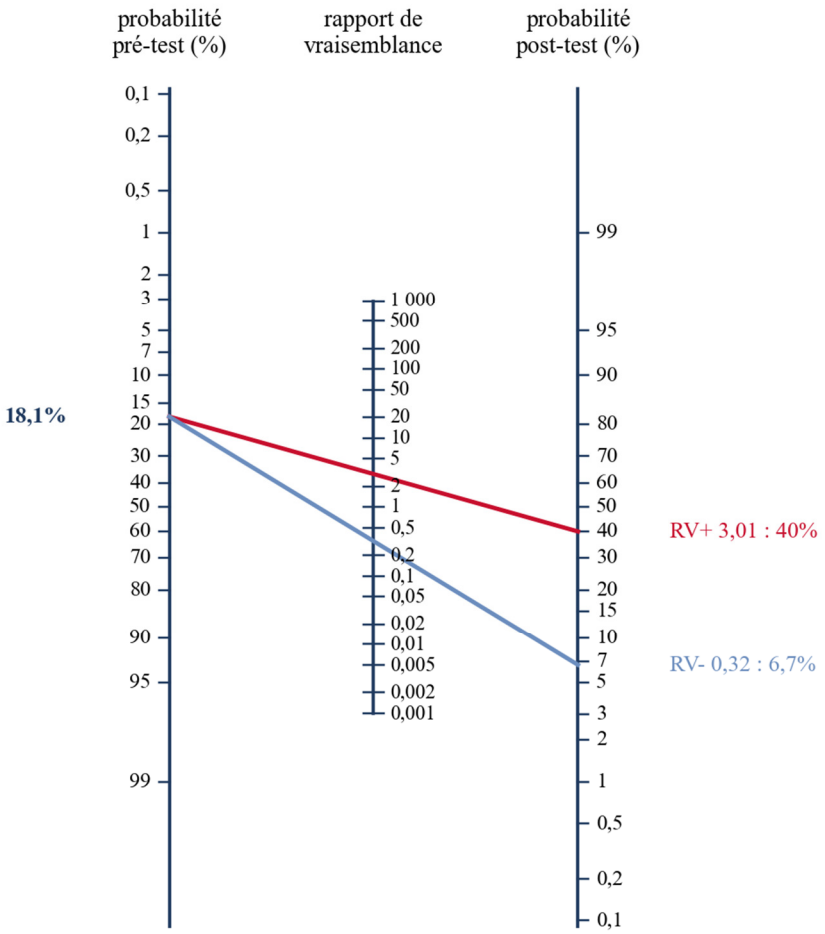


Figure 11.3. Nomogramme de Fagan appliqué au score FINDRISC au seuil ≥ 12 dans le jeu de données mg2026 : à partir de la probabilité pré-test de 18,1%, un résultat positif (RV+ 3,01) conduit à 40%, et un résultat négatif (RV- 0,32) à 6,7%. Données synthétiques (jeu de données mg2026).

Le nomogramme a aujourd'hui surtout une valeur didactique, mais il est utile aussi au lit du malade pour une estimation rapide. La précision de lecture est de quelques pour cent, ce qui suffit pour la décision. Pour des calculs exacts, nous construirons lors des travaux pratiques 11 un calculateur bayésien dans Google Sheets.

> **Étude de cas : deux seuils pour la même élastographie, l'un pour exclure, l'autre pour confirmer [3]**
>
> Moga et ses collaborateurs (avec R. Lupuşoru, de la discipline) ont évalué l'élastographie time-harmonic pour la fibrose hépatique, avec l'élastographie impulsionnelle à vibration contrôlée (VCTE) comme méthode de référence, chez 165 patients. Dans l'échantillon, 28,1% avaient une cirrhose (F4) selon la VCTE. Les auteurs n'ont pas rapporté un seul seuil, mais deux. Le seuil d'exclusion de la cirrhose (angl. rule-out), inférieur à 1,61 m/s, avait une Se de 90,7%, une Sp de 66,6%, une VPP de 55,7% et une VPN de 93,9%. Le seuil de confirmation (angl. rule-in), supérieur à 1,83 m/s, avait une Se de 65,1%, une Sp de 96,7%, une VPP de 90,3% et une VPN de 85,7%. L'AUROC était de 0,90 (IC à 95% 0,82-0,93). Traduisons en rapports de vraisemblance. Pour le seuil d'exclusion, $\mathrm{RV}^{-} = \frac{1-0,907}{0,666} = 0,14$, c'est-à-dire qu'un résultat inférieur au seuil diminue fortement la cote de cirrhose. Pour le seuil de confirmation, $\mathrm{RV}^{+} = \frac{0,651}{1-0,967} = 19,7$, c'est-à-dire qu'un résultat supérieur au seuil multiplie la cote par près de 20. Si nous appliquons ces RV à la prévalence de l'échantillon (28,1%), nous obtenons, par la forme avec les cotes, une VPP d'environ 88,5% au seuil de confirmation et une VPN d'environ 94,8% au seuil d'exclusion, des valeurs proches de celles rapportées (les petites différences proviennent des arrondis et du nombre exact de patients analysés à chaque seuil). Pour une prévalence de 10%, comme dans une population de dépistage, les mêmes seuils donneraient une VPP d'environ 68,7% pour le seuil de confirmation. Entre les deux seuils subsiste une zone grise dans laquelle le test ne tranche pas, et le patient a besoin d'un autre examen. L'étude a également rapporté la reproductibilité (ICC entre 0,92 et 0,97), une information souvent oubliée, mais décisive pour un test utilisé par des examinateurs différents.

11.5 Les tests séquentiels et la combinaison des tests

11.5.1 Tests en série et en parallèle

Il est rare qu'un seul test tranche un diagnostic. Nous combinons habituellement des tests, et le mode de combinaison modifie la performance globale. Il existe deux stratégies de base.

Dans les tests en parallèle, on réalise les deux tests, et le patient est considéré comme positif si au moins l'un des tests est positif (règle "OU"). Cette stratégie augmente la sensibilité (nous manquons moins de malades), mais diminue la spécificité (le nombre de faux positifs augmente). On l'utilise lorsque manquer un malade est grave, par exemple aux urgences, quand le patient ne reviendra pas.

Dans les tests en série, le second test n'est réalisé que chez les sujets positifs au premier, et le patient est considéré comme positif seulement si les deux tests sont positifs (règle "ET"). Cette stratégie augmente la spécificité et la VPP, mais diminue la sensibilité. Elle est économique, car le second test (habituellement plus coûteux ou plus invasif) n'est appliqué qu'à une partie des patients. Le dépistage du cancer colorectal (FIT, puis coloscopie) et le diagnostic de l'infection par le VIH (test de dépistage, puis test de confirmation) en sont des exemples classiques.

Si les deux tests sont indépendants conditionnellement au statut de maladie (nous expliquerons tout de suite ce que cela signifie), la performance combinée se calcule simplement :

en série : Se = Se1 x Se2 ; Sp = 1 - (1 - Sp1) x (1 - Sp2)

en parallèle : Se = 1 - (1 - Se1) x (1 - Se2) ; Sp = Sp1 x Sp2

Et, sous la forme des cotes, les tests séquentiels reviennent à multiplier successivement par les rapports de vraisemblance :

cote post-test finale = cote pré-test x RV1 x RV2

La probabilité post-test après le premier test devient la probabilité pré-test du second. C'est la propriété la plus élégante du raisonnement bayésien : l'information s'accumule.

11.5.2 L'hypothèse d'indépendance conditionnelle

Deux tests sont conditionnellement indépendants si, parmi les malades, le résultat d'un test ne nous dit rien sur le résultat de l'autre, et de même parmi les non-malades. Cette hypothèse est souvent violée, surtout lorsque les tests mesurent des aspects proches de la même physiopathologie. Le score FINDRISC et la glycémie à jeun, par exemple, dépendent tous deux de l'obésité, de l'âge et de la sédentarité. Une personne non diabétique, mais présentant une obésité abdominale, a plus de chances d'être "positive" aux deux.

Le jeu de données mg2026 nous permet de le vérifier. Considérons FINDRISC ≥ 12 (Se 75,9%, Sp 74,8%) et glycémie à jeun ≥ 110 mg/dL (Se 82,8%, Sp 70,2% ; 24 malades positifs sur 29, 92 non-malades négatifs sur 131). Si les tests étaient conditionnellement indépendants, la combinaison en série aurait :

Se = 0,759 x 0,828 = 0,628 ; Sp = 1 - 0,252 x 0,298 = 0,925

Les données réelles (patients positifs aux deux tests) montrent pourtant une Se de 69,0% (20 sur 29) et une Sp de 85,5% (112 sur 131). La spécificité réelle est inférieure à la spécificité prédite, car parmi les 131 non-malades, 19 sont positifs aux deux tests, beaucoup plus qu'on ne l'attendrait si les tests étaient indépendants $\left(\frac{33 \times 39}{131} = 9{,}8\right)$. La conséquence pratique se voit sur la probabilité post-test. En multipliant successivement les rapports de vraisemblance (RV+ 3,01 pour le FINDRISC, RV+ 2,78 pour la glycémie), nous estimerions pour un patient positif aux deux tests une probabilité post-test de 64,9%. La valeur observée dans les données n'est que de 20/39 = 51,3%. L'hypothèse d'indépendance a surestimé le gain apporté par le second test.

Tableau 11.4. Combinaison du FINDRISC ≥ 12 et de la glycémie à jeun ≥ 110 mg/dL pour le diabète confirmé par HGPO (données synthétiques, jeu de données mg2026, n = 160, dont 29 diabétiques).

Stratégie	VP	FP	FN	VN	Se (%)	Sp (%)	VPP (%)	VPN (%)
FINDRISC ≥ 12	22	33	7	98	75,9	74,8	40,0	93,3
Glycémie ≥ 110	24	39	5	92	82,8	70,2	38,1	94,8
En série (les deux +)	20	19	9	112	69,0	85,5	51,3	92,6
En parallèle (au moins un +)	26	53	3	78	89,7	59,5	32,9	96,3

Le tableau confirme la règle générale : la série gagne en spécificité et en VPP, le parallèle gagne en sensibilité et en VPN. La leçon méthodologique est que la multiplication des rapports de vraisemblance n'est correcte que pour des tests conditionnellement indépendants. Pour des tests corrélés, la performance de la combinaison doit être mesurée directement, chez des patients ayant bénéficié des deux tests, ou estimée par un modèle multivarié, par exemple la régression logistique du chapitre 9.

Étude de cas : des règles cliniques qui combinent plusieurs mesures [110,123]

Deux études du groupe d'hépatologie de Timişoara, avec R. Lupuşoru de la discipline, montrent comment combiner plusieurs mesures dans une décision. Barbulescu et ses collaborateurs ont étendu et validé les critères de Baveno VI pour l'exclusion des varices œsophagiennes à haut risque, chez 839 patients (cohorte de dérivation 728, cohorte de validation 111). Les critères de Baveno VI ont eu une exactitude de 86,2% dans la première cohorte. Les auteurs ont dérivé une règle de confirmation (plaquettes inférieures à 150 000/mm³ et rigidité hépatique supérieure à 35,3 kPa) et une règle d'exclusion (plaquettes supérieures à 150 000/mm³ et rigidité hépatique inférieure à 19,6 kPa), et ont proposé pour la zone grise un seuil supplémentaire d'albumine, inférieur à 3,4 g/dl. Dans la cohorte de validation, les seuils d'exclusion ont eu une exactitude de 100%. La règle "ET" de chaque critère est une combinaison en série ; le maintien d'une zone grise reconnaît que, entre les deux seuils, la probabilité post-test reste trop proche du milieu pour décider sans endoscopie. Fofiu et ses collaborateurs ont abordé le même problème autrement : chez 124 patients analysés atteints de cirrhose compensée, ils ont combiné la rigidité splénique (seuil 13,2 kPa, AUROC 0,84), la rigidité hépatique (12,1 kPa, AUROC 0,86) et la taille de la rate (12,9 cm, AUROC 0,71) en un score obtenu par régression : 0,053 x SS + 0,054 x LS + 0,059 x SSZ - 1,84, avec un seuil > 0,34 et une AUROC de 0,93, supérieure à celle de chaque facteur pris séparément. Le modèle de régression estime les pondérations en tenant compte de la corrélation entre les mesures, c'est-à-dire précisément du problème de l'indépendance conditionnelle. Les deux études doivent être lues avec une réserve : une exactitude de 100% dans une cohorte de validation de 111 patients a un intervalle de confiance large, et un score dérivé sur 124 patients nécessite une validation externe avant une utilisation en routine.

11.6 Le diagnostic différentiel avec plusieurs maladies

11.6.1 Le tableau de Bayes à trois diagnostics

En pratique, le médecin ne se demande pas seulement "le patient a-t-il ou non la maladie X ?", mais "laquelle des maladies X, Y, Z explique le mieux ce que je vois ?". Le théorème de Bayes s'étend directement à plusieurs hypothèses, à condition que celles-ci soient mutuellement exclusives (le patient n'en a qu'une) et exhaustives (l'une d'elles est vraie). Pour chaque maladie Di :

$$p(D_i/S) = \frac{p(S/D_i) \times p(D_i)}{\sum\left[p\left(S/D_j\right) \times p\left(D_j\right)\right]}$$

Le numérateur se calcule pour chaque maladie, et le dénominateur est la somme des numérateurs. En pratique, la probabilité a posteriori de chaque maladie est la part de son numérateur dans le total.

Prenons un exemple de médecine générale, avec des chiffres fictifs, mais plausibles. Un homme de 68 ans, ancien fumeur, consulte pour une dyspnée d'effort progressive au cours des derniers mois. Par souci de simplicité, supposons que trois diagnostics couvrent presque tous les cas de ce type dans le cabinet concerné : l'insuffisance cardiaque (IC), la bronchopneumopathie chronique obstructive (BPCO) et l'anémie. D'après l'expérience du cabinet, chez les patients ayant ce profil, les probabilités a priori sont de 0,35 pour l'IC, 0,45 pour la BPCO et 0,20 pour l'anémie. Le médecin note trois signes : des œdèmes bilatéraux des membres inférieurs (E), des sibilants à l'auscultation (W) et une pâleur cutanée (P). La fréquence de chaque signe dans chaque maladie est donnée dans le tableau 11.5.

Tableau 11.5. Probabilités a priori et fréquence des signes cliniques dans chaque diagnostic (données fictives).

Diagnostic	**p(D)**	$p(E/D)$	$p(W/D)$	$p(P/D)$
Insuffisance cardiaque	0,35	0,60	0,20	0,15
BPCO	0,45	0,15	0,60	0,10
Anémie	0,20	0,10	0,05	0,70

Le patient a des œdèmes (E+), n'a pas de sibilants (W-) et n'est pas pâle (P-). Nous supposons que, à l'intérieur de chaque maladie, les signes apparaissent indépendamment les uns des autres (hypothèse d'indépendance conditionnelle, appelée en informatique modèle "Bayes naïf"). La vraisemblance de la combinaison observée dans chaque maladie est le produit des probabilités : pour un signe présent, nous utilisons p(S / D), pour un signe absent, 1 - p(S / D).

Pour l'IC : 0,60 x (1 - 0,20) x (1 - 0,15) = 0,60 x 0,80 x 0,85 = 0,408

Pour la BPCO : 0,15 x (1 - 0,60) x (1 - 0,10) = 0,15 x 0,40 x 0,90 = 0,054

Pour l'anémie : 0,10 x (1 - 0,05) x (1 - 0,70) = 0,10 x 0,95 x 0,30 = 0,0285

Nous multiplions chaque vraisemblance par la probabilité a priori :

IC : 0,35 x 0,408 = 0,1428

BPCO : 0,45 x 0,054 = 0,0243

Anémie : 0,20 x 0,0285 = 0,0057

La somme vaut 0,1728. Les probabilités a posteriori sont :

$$p(\text{IC/données}) = \frac{0{,}1428}{0{,}1728} = 0{,}826$$

$$p(\text{BPCO/données}) = \frac{0{,}0243}{0{,}1728} = 0{,}141$$

$$p(\text{ anémie /données}) = \frac{0{,}0057}{0{,}1728} = 0{,}033$$

Tableau 11.6. Calcul des probabilités a posteriori pour le patient dyspnéique (E+, W-, P-).

Diagnostic	**A priori**	**Vraisemblance**	**Produit**	**A posteriori**
Insuffisance cardiaque	0,35	0,408	0,1428	82,6%
BPCO	0,45	0,054	0,0243	14,1%
Anémie	0,20	0,0285	0,0057	3,3%
Total	1,00		0,1728	100%

La hiérarchie des diagnostics s'est inversée. Avant l'examen, la BPCO était le diagnostic le plus probable (45%) ; après l'examen, l'insuffisance cardiaque atteint 82,6%. Ce résultat ne signifie pas que le diagnostic est établi : une probabilité de 82,6% justifie un examen ciblé (par exemple NT-proBNP et échocardiographie), et non un traitement définitif sans confirmation. Le calcul peut aussi se faire pas à pas, signe par signe, en utilisant la probabilité a posteriori de l'étape précédente comme probabilité a priori de l'étape suivante. Après les œdèmes seuls, les probabilités seraient de 70,6% / 22,7% / 6,7% ; après l'ajout de l'absence de sibilants, de 78,5% / 12,6% / 8,9% ; après l'absence de pâleur, de 82,6% / 14,1% / 3,3%. Le résultat final est le même, quel que soit l'ordre dans lequel nous ajoutons les signes, à condition que l'hypothèse d'indépendance soit valable.

11.6.2 Les limites du modèle "naïf"

Le modèle de l'exemple ci-dessus présente trois faiblesses que vous devez garder à l'esprit. La première est l'exhaustivité : si le patient a en réalité une quatrième maladie (par exemple une fibrose pulmonaire ou une valvulopathie), aucun chiffre du tableau ne la signalera ; les probabilités seront redistribuées entre les trois diagnostics listés. La deuxième est l'indépendance

des signes : les œdèmes et l'orthopnée apparaissent ensemble dans l'insuffisance cardiaque beaucoup plus souvent que ne le suggérerait le produit de leurs fréquences, et les multiplier comme s'ils étaient indépendants exagère les preuves. La troisième est la qualité des données d'entrée : les probabilités a priori et les fréquences des signes sont des estimations, souvent issues d'études menées sur des populations différentes.

Malgré cela, les modèles de type Bayes naïf ont été parmi les premiers systèmes d'aide à la décision clinique et restent étonnamment robustes pour ordonner les diagnostics, même si les probabilités absolues ne sont pas exactes [124]. Le cours de la discipline les présente, à côté de la reconnaissance de formes (angl. pattern recognition), comme des méthodes statistiques de classification. L'idée de base, la mise à jour des probabilités par la multiplication de la probabilité a priori par la vraisemblance, est aussi au fondement de nombreux classificateurs modernes, y compris certains de ceux utilisés en intelligence artificielle médicale.

11.7 Les seuils de décision : tester, traiter

La probabilité post-test n'est pas une fin en soi. Elle sert une décision : traitons-nous, poursuivons-nous les investigations ou ne faisons-nous rien ? Pour relier la probabilité à la décision, nous avons besoin du seuil de traitement, c'est-à-dire de la probabilité au-delà de laquelle le bénéfice attendu du traitement dépasse son risque.

Notons B le bénéfice net du traitement d'un malade (par exemple les complications évitées) et R le risque net du traitement d'un non-malade (effets indésirables, coûts, anxiété). Le seuil de traitement est

$$p_{\text{traitement}} = \frac{R}{R + B}$$

Si traiter inutilement un patient en bonne santé "coûte" une unité et ne pas traiter un malade "coûte" quatre unités (B = 4), le seuil est de 1 / (1 + 4) = 0,20. Au-delà de 20%, il vaut la peine de traiter ; en deçà de 20%, non. Pour un traitement très sûr et très efficace (par exemple un antibiotique oral bien toléré dans une infection aux complications graves), le seuil est bas. Pour un traitement risqué (chimiothérapie, anticoagulation chez une personne âgée à risque hémorragique), le seuil est élevé.

Dès que nous disposons d'un test, deux seuils supplémentaires apparaissent. Le seuil de test (inférieur) est la probabilité pré-test en dessous de laquelle même un résultat positif ne ferait pas passer le patient au-dessus du seuil de traitement ; en dessous de ce seuil, le test est inutile et le patient n'est pas traité. Le seuil de test-traitement (supérieur) est la probabilité pré-test au-dessus de laquelle même un résultat négatif ne ferait pas descendre le patient sous le seuil de traitement ; au-dessus de ce seuil, on traite sans test. Entre les deux seuils, le test peut changer la décision et mérite d'être réalisé [120].

Avec la forme des cotes, le calcul est direct. Soit un seuil de traitement de 20% (cote 0,25) et un test avec RV+ = 6 et RV- = 0,2.

Seuil de test : cote pré-test = 0,25 / 6 = 0,0417, donc probabilité 0,0417 / 1,0417 = 4,0%.

Seuil supérieur : cote pré-test = 0,25 / 0,2 = 1,25, donc probabilité 1,25 / 2,25 = 55,6%.

Les patients dont la probabilité pré-test se situe entre 4,0% et 55,6% bénéficient du test. En dessous de 4,0%, on ne teste pas et on ne traite pas ; au-dessus de 55,6%, on traite directement. Le modèle est simplifié (il ignore le risque et le coût du test lui-même), mais il illustre l'idée principale : la décision de demander un test dépend de la probabilité pré-test, de la performance du test et des conséquences des erreurs. Un meilleur test (RV+ plus élevé, RV- plus faible) élargit l'intervalle dans lequel le test est utile.

11.8 Le dépistage en médecine générale : l'exemple du FINDRISC

Le score FINDRISC (angl. Finnish Diabetes Risk Score) a été construit pour identifier, au moyen d'un questionnaire simple, les personnes à haut risque de diabète de type 2 [125]. Il prend des valeurs de 0 à 26 et inclut l'âge, l'IMC, le tour de taille, l'activité physique, la consommation quotidienne de légumes et de fruits, le traitement antihypertenseur, des glycémies élevées dans les antécédents et le diabète chez des parents au premier ou au deuxième degré. C'est exactement le type d'outil qu'un médecin généraliste peut utiliser chez un patient qui consulte pour autre chose.

Dans le jeu de données mg2026, le diabète a été confirmé par HGPO (le standard de référence, selon les critères de l'ADA [80]), et au seuil FINDRISC ≥ 12 nous avons RV+ 3,01 et RV- 0,32. Pour voir ce que signifient ces chiffres chez des patients différents, nous appliquons la forme des cotes à plusieurs probabilités pré-test.

Tableau 11.7. Probabilité post-test de diabète après le FINDRISC au seuil ≥ 12 (RV+ 3,01 ; RV- 0,32), pour différentes probabilités pré-test. Les rapports de vraisemblance proviennent du jeu de données mg2026 (données synthétiques) ; les probabilités pré-test sont des scénarios.

Probabilité pré-test	Contexte possible	Cote pré-test	Post-test après FINDRISC ≥ 12	Post-test après FINDRISC < 12
2%	jeune adulte, sans facteurs de risque	0,020	5,8%	0,7%
5%	population adulte générale	0,053	13,7%	1,7%
10%	adulte d'âge moyen, en surpoids	0,111	25,1%	3,5%
18,1%	échantillon mg2026 (risque cardiométabolique)	0,221	40,0%	6,7%
30%	patient obèse, prédiabète connu	0,429	56,3%	12,1%
50%	patient présentant des symptômes évocateurs	1,000	75,1%	24,4%

Le tableau nous dit plusieurs choses importantes. Un FINDRISC positif ne "pose pas le diagnostic" : même dans l'échantillon à risque élevé, seuls 4 patients positifs sur 10 ont un diabète. Un FINDRISC négatif est plus utile pour rassurer qu'un FINDRISC positif ne l'est pour alerter, du moins pour des probabilités pré-test faibles et modérées : en dessous de 10% de probabilité pré-test, un score inférieur à 12 fait descendre la probabilité sous 3,5%. Pour des probabilités pré-test élevées (50%), un score négatif laisse néanmoins une probabilité de 24,4%, beaucoup trop élevée pour renoncer au test de laboratoire. Le rôle du FINDRISC au cabinet est donc un rôle de tri : il décide qui mérite une glycémie à jeun ou une HbA1c, et non qui a un diabète. La colonne post-test pour le score positif est d'ailleurs identique au tableau de la VPP en fonction de la prévalence, que vous retrouverez au chapitre 12 : la VPP pour une prévalence p est exactement la probabilité post-test d'un patient dont la probabilité pré-test est p.

Lorsque le patient a déjà un résultat de FINDRISC et que l'on mesure aussi sa glycémie, le raisonnement se poursuit : la probabilité post-test après le FINDRISC devient la probabilité pré-test pour la glycémie. Nous avons vu à la section 11.5 que, dans ce cas particulier, l'hypothèse d'indépendance n'est pas respectée, si bien que la simple multiplication des rapports de vraisemblance surestime la probabilité finale. En pratique, les recommandations exigent de toute façon la confirmation du diagnostic de diabète par des tests de laboratoire, et non par un calcul probabiliste [80].

Étude de cas : les valeurs prédictives dépendent de la prévalence et du spectre des patients [75,126,127]

Trois études des collègues montrent la même leçon sous des angles différents. Albai et ses collaborateurs (avec Mirela Frandeș) ont validé le score ISARIC-4C chez 159 patients atteints de diabète de type 2 hospitalisés pour une infection à SARS-CoV-2, avec une mortalité de 25,15%. L'aire sous la courbe ROC était de 0,875 (IC à 95% 0,820-0,930) et, au seuil choisi, le score a correctement classé 77,36% de la cohorte, avec une Se de 87,80%, une Sp de 73,72%, une VPP de 53,73% et une VPN de 94,56%. Si nous recalculons la VPP par le théorème de Bayes pour une prévalence de 25,15%, nous obtenons environ 52,9%, une valeur proche de celle rapportée ; cette même VPP diminuerait fortement dans une population où la mortalité serait de 10%. Une observation de lecture critique : les intervalles de confiance rapportés pour la Se et la Sp (par exemple 87,66-87,93) sont inhabituellement étroits pour 159 patients, dont seulement 40 environ sont décédés ; un intervalle de Wilson pour une proportion de 88% calculée sur 40 cas aurait une largeur supérieure à 20 points de pourcentage. Sipos et ses collaborateurs (avec R. Moleriu) ont rapporté pour l'IL-6 dans le diagnostic du sepsis néonatal un seuil optimal de 181 pg/ml, avec une Se de 80,1%, une Sp de 85,7%, une VPP de 84,6% et une VPN de 81,8%. Le résumé ne donne pas le nombre de nouveau-nés, mais nous pouvons faire un exercice bayésien inverse : à partir de la Se et de la Sp, $RV^+ = \frac{0,801}{0,143} = 5,6$; la cote post-test correspondant à la VPP de 84,6% est de 0,846 / 0,154 = 5,49 ; en divisant, la cote pré-test est d'environ 0,98, soit une prévalence du sepsis d'environ 49,5% dans l'échantillon analysé. Parmi l'ensemble des nouveau-nés testés dans une maternité, la prévalence du sepsis est beaucoup plus faible ; en pratique, la VPP serait donc inférieure à 84,6%. Pavel et ses collaborateurs (avec A. Tudor) ont évalué la H-FABP chez 83 patients suspects d'infarctus aigu du myocarde, séparément selon le délai depuis le début de la douleur. À moins de 4 heures, la sensibilité et la spécificité étaient de 89,67 et 95,65, ce qui correspond à un RV+ d'environ 20,6 ; au-delà de 4 heures, la sensibilité était de 86,73, mais la spécificité de seulement 49,84, soit un RV+ d'environ 1,73, presque sans valeur de confirmation. Le même biomarqueur, dans le même service, a des performances complètement différentes selon le moment où les patients sont testés. Par conséquent, les RV ne sont pas des constantes universelles : ils s'appliquent à des patients semblables à ceux de l'étude qui les a estimés.

11.9 L'inférence bayésienne par rapport à l'inférence fréquentiste

Jusqu'ici, nous avons appliqué le théorème de Bayes au patient individuel. La même logique peut s'appliquer aux paramètres d'une population : une moyenne, une proportion, une différence entre traitements. C'est l'inférence bayésienne, une alternative aux tests d'hypothèse et aux intervalles de confiance fréquentistes des chapitres 4-10.

11.9.1 A priori, vraisemblance, a posteriori

Dans l'inférence fréquentiste, le paramètre (par exemple la prévalence du diabète dans la population, π) est une constante inconnue, et la probabilité ne décrit que la variabilité des échantillons. Dans l'inférence bayésienne, l'incertitude sur le paramètre s'exprime directement par une distribution de probabilité. Avant les données, nous disposons de la distribution a priori (angl. prior), qui résume ce que nous savons ou croyons du paramètre. Les données interviennent par la vraisemblance (la probabilité des données observées pour chaque valeur possible du paramètre). Leur combinaison donne la distribution a posteriori (angl. posterior) :

a posteriori $\propto$ vraisemblance x a priori

Le signe ∝ signifie "proportionnel à" ; une constante de normalisation fait en sorte que l'aire sous la distribution a posteriori soit égale à 1. La formule est le théorème de Bayes appliqué à une distribution entière, et non à une seule probabilité.

Pour une proportion, le calcul est particulièrement simple si l'on prend comme a priori une distribution Bêta. Une distribution Bêta(α ; β) peut être comprise comme l'équivalent de α - 1 "succès" et β - 1 "échecs" observés auparavant. Si nous observons k cas sur n, la distribution a posteriori est Bêta(α + k ; β + n - k). Pour le diabète dans mg2026, k = 29 et n = 160.

Avec un a priori uniforme, Bêta(1 ; 1), qui affirme que toute prévalence comprise entre 0 et 100% est également plausible, la distribution a posteriori est Bêta(30 ; 132). Sa moyenne est 30/162 = 18,5%, et l'intervalle qui contient 95% de la probabilité a posteriori, appelé intervalle de crédibilité (angl. credible interval), est 12,9-24,8%. Avec un a priori informatif, Bêta(10 ; 40), qui exprime la conviction antérieure que la prévalence se situe autour de 20% (avec 95% de la probabilité entre 10,2% et 32,0%), la distribution a posteriori est Bêta(39 ; 171), de moyenne 18,6% et d'intervalle de crédibilité 13,6-24,1%. La figure 11.4 illustre les deux situations.

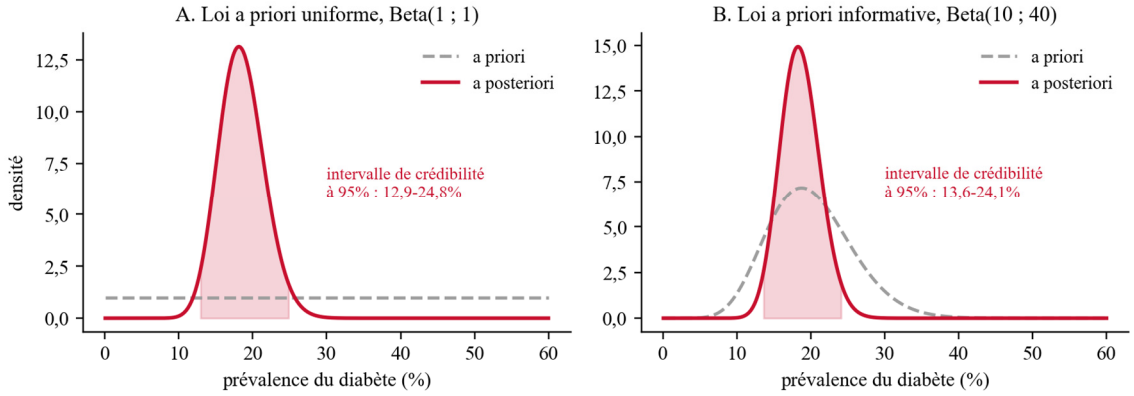


Figure 11.4. Distribution a priori (trait discontinu) et a posteriori (trait continu) de la prévalence du diabète confirmé par HGPO, avec 29 cas sur 160. A : a priori uniforme, intervalle de crédibilité à 95% 12,9-24,8%. B : a priori informatif centré sur 20%, intervalle de crédibilité 13,6-24,1%. Données synthétiques (jeu mg2026).

Deux remarques. Avec 160 patients, ce sont les données qui dominent : les deux distributions a posteriori sont très proches, même si les a priori diffèrent beaucoup. Avec 20 patients, l'a priori aurait pesé bien davantage. Ensuite, l'intervalle de crédibilité obtenu avec l'a priori uniforme (12,9-24,8%) coïncide pratiquement avec l'intervalle de confiance de Wilson calculé au chapitre 4 (12,9-24,8%). Pour de grands échantillons et des a priori peu informatifs, les méthodes bayésiennes et fréquentistes donnent des résultats numériques très proches ; c'est l'interprétation qui diffère.

11.9.2 Intervalle de crédibilité et intervalle de confiance

L'intervalle de confiance fréquentiste à 95% a une interprétation peu commode : si nous répétions l'étude un très grand nombre de fois en calculant chaque fois l'intervalle, 95% de ces intervalles contiendraient la vraie valeur. Nous ne pouvons pas dire que "la vraie valeur se trouve avec une probabilité de 95% dans l'intervalle 12,9-24,8%", car, au sens fréquentiste, le paramètre n'est pas une variable aléatoire. Cette confusion est l'une des erreurs d'interprétation les plus fréquentes en statistique [70].

L'intervalle de crédibilité bayésien a exactement l'interprétation que le clinicien attend intuitivement : compte tenu de l'a priori et des données, la probabilité que la prévalence se situe entre 12,9% et 24,8% est de 95%. Le prix à payer est le choix de l'a priori, qui introduit un élément de jugement devant être justifié et rapporté en toute transparence. Un a priori choisi pour

obtenir le résultat souhaité est aussi malhonnête qu'un test fréquentiste répété jusqu'à obtenir p < 0,05.

11.9.3 Facteur de Bayes

L'équivalent bayésien du test d'hypothèse est le facteur de Bayes. Il compare dans quelle mesure deux hypothèses concurrentes prédisent les données observées, habituellement H1 (il existe un effet) et H0 (il n'existe pas d'effet) :

$$\mathrm{BF}_{10} = \frac{p(\text{données}/H_1)}{p(\text{données}/H_0)}$$

Un BF10 de 10 signifie que les données sont 10 fois plus probables sous H1 que sous H0. Un BF10 de 0,1 signifie que les données sont 10 fois plus probables sous H0 ; de manière équivalente, $\mathrm{BF}_{01} = \frac{1}{\mathrm{BF}_{10}} = 10$. Contrairement à la valeur p, le facteur de Bayes peut apporter des preuves en faveur de l'hypothèse nulle, et pas seulement contre elle. Il s'agit en fait d'un rapport de vraisemblance appliqué aux hypothèses :

cote a posteriori de H1 = cote a priori de H1 x BF10

Vous reconnaissez la formule vue pour le diagnostic : les hypothèses jouent le rôle des maladies, les données celui du résultat du test.

Le logiciel JASP, utilisé en travaux pratiques, propose pour de nombreux tests fréquentistes une variante bayésienne (dans le menu de chaque test, à la section Bayesian) et utilise une échelle d'interprétation largement répandue [47,49].

Tableau 11.8. Échelle d'interprétation du facteur de Bayes BF10 couramment utilisée (notamment dans la documentation de JASP).

BF10	Preuves en faveur de H1	BF10	Preuves en faveur de H0
1-3	anecdotiques (faibles)	1/3-1	anecdotiques (faibles)
3-10	modérées	1/10-1/3	modérées
10-30	fortes	1/30-1/10	fortes
30-100	très fortes	1/100-1/30	très fortes
plus de 100	extrêmes	moins de 1/100	extrêmes

Sur le jeu mg2026, quelques analyses des chapitres précédents, refaites selon l'approche bayésienne (avec les a priori par défaut de JASP, par exemple la distribution de Cauchy de largeur 0,707 pour le test t), donnent les résultats suivants :

- PAS à l'inclusion, femmes versus hommes (test t pour échantillons indépendants) : BF10 = 0,18, soit BF01 = 5,7. Les données sont environ 5,7 fois plus probables s'il n'existe pas de différence entre les sexes : preuves modérées en faveur de H0. Le test fréquentiste (p = 0,81) disait seulement que nous ne pouvons pas rejeter H0 ; le facteur de Bayes indique en outre que les données soutiennent activement l'absence de différence.
- Glycémie à 0 versus 6 mois (test t apparié) : BF10 de l'ordre de 10^18, preuves extrêmes en faveur d'une modification.
- Corrélation IMC-PAS (r = 0,405, n = 160) : BF10 d'environ 116 000, preuves extrêmes.
- Corrélation HDL-triglycérides (r = 0,076) : BF10 = 0,16, soit BF01 = 6,4, preuves modérées en faveur de l'absence de corrélation.
- Obésité et diabète (tableau de contingence 2 x 2, échantillonnage multinomial conjoint, a priori par défaut) : BF10 d'environ 85, preuves très fortes en faveur d'une association.

L'analyse bayésienne ne résout pas les problèmes de conception (une étude mal conçue reste mal conçue) et ne remplace pas le jugement clinique sur la taille de l'effet. Elle présente toutefois deux

avantages didactiques : elle répond directement à la question "dans quelle mesure les données soutiennent-elles une hypothèse plutôt qu'une autre" et elle oblige à expliciter l'information antérieure. Dans la littérature médicale, les méthodes bayésiennes apparaissent de plus en plus souvent dans les essais cliniques adaptatifs et dans les méta-analyses, et l'étudiant d'aujourd'hui les lira en tant que médecin [42,63].

11.10 Médecine fondée sur les preuves

La médecine fondée sur les preuves (angl. evidence-based medicine) a été définie comme l'utilisation consciencieuse, explicite et judicieuse des meilleures preuves disponibles dans les décisions concernant la prise en charge de chaque patient, en intégrant l'expertise clinique aux preuves issues de la recherche systématique [15]. La définition comporte trois composantes, et toutes trois comptent : les preuves issues de la recherche, l'expérience du clinicien ainsi que les valeurs et les préférences du patient. La médecine fondée sur les preuves ne signifie ni des recommandations appliquées mécaniquement, ni la statistique à la place de la clinique.

11.10.1 Les cinq étapes

La pratique de la médecine fondée sur les preuves se décrit habituellement en cinq étapes, faciles à mémoriser en anglais sous la forme des cinq "A".

Étape 1, formulation de la question (angl. ask). Une question clinique bien construite a la structure PICO [24] : P (le patient ou la population), I (l'intervention, l'exposition ou le test diagnostique), C (le comparateur) et O (le critère de jugement, angl. outcome). Par exemple : "Chez les adultes à risque cardiométabolique suivis en médecine générale (P), l'utilisation d'une application mobile d'automesure, ajoutée au conseil (I), par rapport au conseil standard (C), augmente-t-elle la proportion de ceux qui perdent au moins 5% de leur poids à 6 mois (O) ?". Pour une question diagnostique : "Chez les adultes en surpoids non explorés (P), quelles sont la sensibilité et la spécificité du score FINDRISC ≥ 12 (I) pour le diabète de type 2 (O), par rapport à l'HGPO comme standard de référence (C) ?".

Étape 2, recherche des preuves (angl. acquire). On cherche dans des bases de données bibliographiques comme PubMed, dans des collections de revues systématiques et dans des recommandations, en utilisant les termes du PICO. La recherche efficace a été abordée dans le volume d'informatique médicale du premier semestre [1].

Étape 3, évaluation critique (angl. appraise). L'étude trouvée est-elle valide (schéma, randomisation, insu, perdus de vue, standard de référence) ? Quels sont les résultats (taille de l'effet, intervalle de confiance) ? S'appliquent-ils à mon patient ? Pour les études diagnostiques, l'évaluation critique suit la liste STARD, que nous discutons au chapitre 12.

Étape 4, application (angl. apply). Les preuves sont intégrées à la situation concrète du patient : comorbidités, préférences, accès au traitement, coûts. C'est ici qu'intervient le raisonnement bayésien de ce chapitre, car la probabilité pré-test de mon patient peut différer de la prévalence de l'étude.

Étape 5, évaluation du résultat (angl. assess). Le médecin évalue sa propre pratique : la décision a-t-elle conduit au résultat souhaité ? Que peut-on mieux faire la prochaine fois ?

11.10.2 Hiérarchie des preuves et système GRADE

Toutes les études n'offrent pas des preuves aussi solides. Pour les questions portant sur l'efficacité des traitements, la hiérarchie classique place à la base l'opinion d'experts et les séries de cas, suivies des études cas-témoins, des études de cohorte, des essais cliniques randomisés et, au sommet, des revues systématiques avec méta-analyse d'essais randomisés. Pour les questions

diagnostiques, le sommet est occupé par les études transversales menées sur des patients consécutifs, qui reçoivent tous à la fois le test évalué et le standard de référence ; les études de type cas-témoins (malades connus versus volontaires sains) surestiment systématiquement l'exactitude. Pour les questions pronostiques, le schéma préféré est la cohorte prospective.

La pyramide est utile comme première orientation, mais elle simplifie. Un petit essai randomisé, avec beaucoup de perdus de vue, peut fournir des preuves plus faibles qu'une grande cohorte bien menée. C'est pourquoi l'approche actuelle, GRADE (angl. Grading of Recommendations, Assessment, Development and Evaluation), évalue le niveau de confiance dans les preuves pour chaque critère de jugement séparément, et non pour chaque type d'étude [128]. Les preuves issues des essais randomisés partent d'un niveau de confiance élevé, celles des études observationnelles d'un niveau faible. Le niveau de confiance peut être abaissé en raison d'un risque de biais, d'une incohérence entre les études, du caractère indirect des preuves (la population ou le critère de jugement diffèrent de ceux qui nous intéressent), d'une imprécision (intervalles de confiance larges) et d'un biais de publication. Il peut être relevé, surtout pour les études observationnelles, lorsque l'effet est très important ou qu'il existe un gradient dose-réponse. Le résultat est l'une des quatre catégories : confiance élevée, modérée, faible ou très faible.

11.10.3 Revues systématiques et méta-analyses

Une revue systématique répond à une question précise (habituellement de type PICO), avec un protocole établi à l'avance, une recherche exhaustive dans plusieurs bases de données, des critères explicites d'inclusion et d'exclusion, l'évaluation de la qualité de chaque étude et une synthèse transparente. Elle se distingue de la revue narrative, dans laquelle l'auteur choisit les études selon des critères non déclarés. Une méta-analyse est la composante statistique d'une revue systématique : elle combine les résultats des études en une estimation globale, en pondérant chaque étude selon sa précision (en règle générale, de façon inversement proportionnelle à sa variance).

Le résultat d'une méta-analyse est présenté dans un diagramme en forêt (angl. forest plot), que vous avez déjà rencontré au chapitre 10 pour les mesures d'association du jeu mg2026. Chaque ligne correspond à une étude : un carré pour l'estimation ponctuelle (dont l'aire est proportionnelle au poids de l'étude) et une ligne horizontale pour l'intervalle de confiance. La ligne verticale marque l'absence d'effet (1 pour le RR et l'OR, 0 pour les différences). Le losange situé en bas est l'estimation combinée, sa largeur étant égale à son intervalle de confiance. Si le losange ne touche pas la ligne verticale, l'effet combiné est statistiquement significatif.

L'hétérogénéité décrit à quel point les résultats des études diffèrent entre eux au-delà de ce que le hasard expliquerait. On la teste avec le test Q de Cochran et on la quantifie par I^2, la proportion de la variabilité totale due à des différences réelles entre les études. À titre indicatif, un I^2 d'environ 25% indique une hétérogénéité faible, 50% modérée et 75% élevée. Lorsque l'hétérogénéité est élevée, la combinaison des études en un seul chiffre doit se faire avec prudence, à l'aide d'un modèle à effets aléatoires (qui admet que chaque étude a son propre effet vrai), et il faut en rechercher les explications : populations différentes, doses différentes, schémas d'étude différents.

Pour la publication des revues systématiques, on utilise la recommandation PRISMA 2020 [129], qui comprend une liste de contrôle et le diagramme de flux de la sélection des études (combien d'enregistrements ont été trouvés, combien sont restés après élimination des doublons, combien ont été exclus à la lecture du titre et du résumé, combien à la lecture du texte intégral et pour quels motifs). Dans les méta-analyses d'exactitude diagnostique, on combine des paires de sensibilité et de spécificité, et le résultat est habituellement présenté sous la forme d'une courbe ROC récapitulative, sujet qui dépasse le niveau de ce cours.

Contrairement à une revue systématique, une revue narrative ne déclare pas ses critères de sélection et ne produit pas d'estimation combinée. Un exemple tiré des travaux de nos collègues est la revue de Suciu et collaborateurs consacrée à l'insuffisance cardiaque à fraction d'éjection préservée, qui synthétise de façon narrative des études publiées entre 2009 et 2025 [130]. De tels travaux sont précieux pour comprendre un domaine, mais ils se situent plus bas dans la hiérarchie des preuves qu'une revue systématique avec méta-analyse.

11.11 Erreurs cognitives dans le raisonnement clinique

Le raisonnement clinique fait appel à deux systèmes. Le premier est rapide, intuitif, fondé sur la reconnaissance de schémas ("j'ai déjà vu ça"). Le second est lent, analytique et explicite, comme le calcul bayésien de ce chapitre. Les médecins expérimentés utilisent surtout le premier système et, la plupart du temps, avec succès. L'intuition comporte cependant des erreurs systématiques, appelées biais cognitifs, que l'on peut connaître et corriger [120].

La négligence du taux de base (angl. base-rate neglect) est l'erreur centrale de ce chapitre : nous jugeons la probabilité d'une maladie d'après la ressemblance du tableau clinique, en ignorant à quel point la maladie est rare. Un jeune homme présentant des palpitations, des sueurs et des céphalées "correspond" au phéochromocytome, mais la maladie est si rare que, même avec un tableau évocateur, la probabilité post-test reste faible. L'antidote consiste à toujours commencer par la question "quelle est la fréquence de cette maladie chez les patients comme celui-ci ?".

L'ancrage (angl. anchoring) désigne la fixation sur la première impression et l'ajustement insuffisant lorsque de nouvelles informations apparaissent. Le patient orienté par le tri avec l'étiquette "attaque de panique" sera regardé à travers ce prisme, et une troponine légèrement élevée risque d'être minimisée. En termes bayésiens, l'ancrage consiste à ne pas actualiser suffisamment la probabilité après un résultat ayant un RV élevé. La clôture prématurée lui est apparentée : l'arrêt de la recherche après le premier diagnostic plausible.

La disponibilité (angl. availability) désigne la surestimation de la probabilité des maladies dont nous nous souvenons facilement : le cas dramatique de la semaine précédente, la maladie discutée récemment lors d'un congrès, le diagnostic manqué qui nous a valu une plainte. Après une embolie pulmonaire manquée, beaucoup de médecins demandent un angioscanner chez des patients dont la probabilité pré-test est très faible. La probabilité réelle n'a pas changé ; seule a changé la facilité avec laquelle la maladie nous vient à l'esprit.

D'autres erreurs fréquentes sont le biais de confirmation (nous ne cherchons que les informations qui soutiennent notre hypothèse), l'effet de cadrage (la même information, présentée comme "90% de survie" ou comme "10% de mortalité", conduit à des décisions différentes) et l'excès de confiance. Aucune liste n'élimine ces erreurs, mais deux habitudes les réduisent beaucoup : estimer explicitement la probabilité pré-test avant de demander un test et se poser la question "qu'est-ce que cela pourrait être d'autre ?" avant de clore le diagnostic.

11.12 Communication du risque au patient

Les patients, tout comme les médecins, comprennent beaucoup mieux les fréquences naturelles que les pourcentages conditionnels. "Sur 100 personnes comme vous qui ont un score FINDRISC d'au moins 12, environ 40 ont un diabète et 60 n'en ont pas" est une phrase que tout le monde comprend. "La valeur prédictive positive du test est de 40%" ne l'est pas.

La deuxième règle consiste à communiquer le risque absolu, et pas seulement le risque relatif. Dans l'étude simulée mg2026, la proportion de patients ayant perdu au moins 5% de leur poids a été de 29,6% dans le bras C et de 11,3% dans le bras A. Le risque relatif est de 2,62 ("2,6 fois plus de patients"), mais la différence absolue est de 18,3 points de pourcentage, et le nombre

de sujets à traiter (NST) est de 5,5 : environ 6 patients doivent être inclus dans le programme C au lieu du programme A pour qu'un patient de plus atteigne l'objectif. Pour un effet indésirable rare, la différence entre relatif et absolu est encore plus importante : un médicament qui "double le risque" de thrombose peut faire passer le risque absolu de 2 à 4 pour 10 000 patients par an. Les deux affirmations sont vraies, mais seule la seconde permet au patient de peser sa décision.

La troisième règle est d'utiliser le même dénominateur pour toutes les options ("sur 1000 personnes") et, lorsque c'est possible, un support visuel : une grille de 100 silhouettes, dont certaines sont colorées. La communication de l'incertitude fait elle aussi partie de la communication du risque. Dire "le test est revenu positif, ce qui signifie que la probabilité que vous ayez la maladie est passée d'environ 1 sur 5 à 2 sur 5 ; nous avons besoin d'un test de confirmation" est plus honnête et, à long terme, plus rassurant qu'un verdict prématuré.

11.13 Problèmes résolus

Problème résolu 11.1. Le tableau maladie-symptôme et le théorème de Bayes

Dans un service d'urgences, 400 patients présentant une douleur thoracique ont été analysés rétrospectivement. Chez 60 d'entre eux, un syndrome coronarien aigu a été confirmé (B+). Le sous-décalage du segment ST sur l'ECG initial (S+) était présent chez 36 des patients atteints d'un syndrome coronarien et chez 34 des autres (données fictives). Calculez p(B+), p(S+ / B+), p(S+ / B-), p(S+), p(B+ / S+) et p(B+ / S-). Vérifiez p(B+ / S+) à l'aide du théorème de Bayes.

Solution. Nous construisons le tableau avec les notations du cours. Sur la ligne B+ : a = 36, b = 60 - 36 = 24. Sur la ligne B- : c = 34, d = 340 - 34 = 306.

	S+	**S-**	**Total**
B+	36	24	60
B-	34	306	340
Total	70	330	400

Probabilités non conditionnelles : $p(B^+) = \frac{60}{400} = 0{,}15$; $p(S^+) = \frac{70}{400} = 0{,}175$.

Probabilités conditionnelles à la maladie : $p(S^+/B^+) = \frac{36}{60} = 0{,}60$; $p(S^+/B^-) = \frac{34}{340} = 0{,}10$.

Probabilités conditionnelles au symptôme : $p(B^+/S^+) = \frac{36}{70} = 0{,}514$; $p(B^+/S^-) = \frac{24}{330} = 0{,}073$.

Vérification par le théorème de Bayes, sans la cellule a : $p(B^+/S^+) = \frac{0{,}60 \times 0{,}15}{0{,}60 \times 0{,}15 + 0{,}10 \times 0{,}85} = \frac{0{,}090}{0{,}090 + 0{,}085} = \frac{0{,}090}{0{,}175} = 0{,}514$.

Interprétation clinique. Le sous-décalage du segment ST fait passer la probabilité de syndrome coronarien de 15% à 51,4%, mais ne la rend pas certaine. Son absence l'abaisse à 7,3%, valeur encore trop élevée pour laisser sortir le patient sans autres examens (dosages répétés de la troponine). Le rapport de vraisemblance du signe est $\text{RV}^+ = \frac{0{,}60}{0{,}10} = 6$, et $\text{RV}^- = \frac{0{,}40}{0{,}90} = 0{,}44$: le signe est utile pour confirmer, beaucoup moins pour exclure.

Problème résolu 11.2. Fréquences naturelles pour un test rapide

Un test rapide de détection des anticorps anti-VHC a une sensibilité de 98% et une spécificité de 97%. On propose de l'utiliser (a) dans la population générale, où la prévalence de l'infection est de

1%, et (b) dans un centre accueillant des personnes qui s'injectent des drogues, où la prévalence est de 20% (chiffres illustratifs). Calculez la VPP et la VPN dans les deux situations en utilisant des fréquences naturelles.

Solution (a). Prenons 10 000 personnes. Une prévalence de 1% donne 100 personnes infectées et 9900 non infectées. Parmi les 100, le test est positif chez 98 (Se 98%) et négatif chez 2. Parmi les 9900, la spécificité de 97% donne 9603 négatifs et 297 faux positifs. Total des positifs : 98 + 297 = 395. VPP $= \frac{98}{395} = 24{,}8\%$. Total des négatifs : 2 + 9603 = 9605. VPN $= \frac{9603}{9605} = 99{,}98\%$.

Solution (b). Sur 10 000 personnes, 2000 sont infectées et 8000 ne le sont pas. Vrais positifs : 0,98 x 2000 = 1960 ; faux négatifs : 40. Faux positifs : 0,03 x 8000 = 240 ; vrais négatifs : 7760. VPP $= \frac{1960}{1960+240} = \frac{1960}{2200} = 89{,}1\%$. VPN $= \frac{7760}{7760+40} = 99{,}49\%$.

Interprétation. Le même test, avec une excellente spécificité de 97%, produit dans la population générale trois faux positifs pour chaque vrai positif : environ 3 résultats positifs sur 4 sont faux. Dans la population à haut risque, 9 résultats positifs sur 10 sont vrais. C'est pourquoi un test de dépistage positif dans la population générale doit toujours être confirmé par un test différent (par exemple la détection de l'ARN viral), et la communication avec le patient doit l'expliquer avant la réalisation du test.

Problème résolu 11.3. Même étude, prévalences différentes : génotypage HPV et double immunomarquage p16/Ki-67

Moleriu et collaborateurs ont suivi 184 patientes ayant une cytologie LSIL, dont 64 ont été biopsiées [131]. L'histologie (le standard de référence) a montré des lésions de haut grade chez 18 sur 64 (28,13%). Le génotypage HPV avait une Se de 94,44% (17/18) et une Sp de 58,7% (27/46) ; le double immunomarquage p16/Ki-67 avait une Se de 66,67% (12/18) et une Sp de 84,78% (39/46). Calculez le RV+ et le RV- de chaque test, puis la VPP et la VPN à la prévalence de l'étude et à une prévalence de 10%.

Solution. Nous reconstituons les tableaux 2 x 2. Pour le HPV : a = 17, b = 1, c = 46 - 27 = 19, d = 27. Pour p16/Ki-67 : a = 12, b = 6, c = 46 - 39 = 7, d = 39. Nous observons que 17 + 19 = 36 patientes étaient HPV positives (56,25% des 64), et 12 + 7 = 19 étaient p16/Ki-67 positives (29,69%), ce qui concorde avec les proportions du résumé.

Les rapports de vraisemblance :

HPV : $RV^+ = \frac{0{,}9444}{1-0{,}587} = \frac{0{,}9444}{0{,}413} = 2{,}29$; $RV^- = \frac{0{,}0556}{0{,}587} = 0{,}095$.

p16/Ki-67 : $RV^+ = \frac{0{,}6667}{0{,}1522} = 4{,}38$; $RV^- = \frac{0{,}3333}{0{,}8478} = 0{,}393$.

À la prévalence de l'étude (28,13%, cote 0,391) : pour le HPV, VPP $= \frac{17}{36} = 47{,}2\%$ et VPN $= \frac{27}{28} = 96{,}4\%$; pour p16/Ki-67, VPP $= \frac{12}{19} = 63{,}2\%$ et VPN $= \frac{39}{45} = 86{,}7\%$. On obtient les mêmes valeurs avec la forme en cotes (par exemple 0,391 x 2,29 = 0,896, et 0,896/1,896 = 47,2%).

À une prévalence de 10% (cote 0,111) : un HPV positif donne 0,111 x 2,29 = 0,254, soit une VPP de 20,3% ; un HPV négatif donne 0,111 x 0,095 = 0,0106, soit une probabilité post-test de 1,0% (VPN 99,0%). Pour p16/Ki-67 : 0,111 x 4,38 = 0,487, VPP 32,7% ; 0,111 x 0,393 = 0,0437, probabilité post-test 4,2% (VPN 95,8%).

Interprétation. Les deux tests ont des profils opposés. Le génotypage HPV, très sensible et peu spécifique, est bon pour exclure : un résultat négatif abaisse fortement la probabilité de lésion de haut grade. L'immunomarquage p16/Ki-67, plus spécifique, est meilleur pour confirmer. Leur

combinaison (HPV en premier test, p16/Ki-67 chez les patientes positives) constituerait une stratégie en série logique, mais la performance de la combinaison ne peut pas être calculée à partir du résumé, car nous ne savons pas combien de patientes ont été positives aux deux tests.

Problème résolu 11.4. Tests séquentiels avec rapports de vraisemblance

Golu et collaborateurs ont étudié l'élastographie 2D-SWE dans les lésions parathyroïdiennes, chez 22 patients atteints d'hyperparathyroïdie et 43 témoins sains [132]. Au seuil de 12,5 kPa, la sensibilité était de 93% et la spécificité de 86% (AUC 0,949). (a) Calculez le RV+ et le RV-. (b) Quelle est la probabilité post-test après un résultat positif, à la prévalence de l'étude et pour une probabilité pré-test de 5% ? (c) Chez le patient ayant une probabilité pré-test de 5% et une élastographie positive, on réalise un second test, conditionnellement indépendant, avec RV+ = 8,5 (valeur hypothétique). Quelle est la probabilité finale ?

Solution. (a) $\mathrm{RV}^{+} = \frac{0{,}93}{1-0{,}86} = \frac{0{,}93}{0{,}14} = 6{,}64$. $\mathrm{RV}^{-} = \frac{0{,}07}{0{,}86} = 0{,}081$.

(b) La prévalence de l'étude est de 22/65 = 33,8%, avec $\frac{\text{cote } 0{,}338}{0{,}662} = 0{,}512$. La cote post-test est 0,512 x 6,64 = 3,40, et la probabilité 3,40/4,40 = 77,3%. Pour une probabilité pré-test de 5%, la cote est 0,0526 ; la cote post-test 0,0526 x 6,64 = 0,350 ; la probabilité 0,350/1,350 = 25,9%.

(c) La probabilité de 25,9% devient la nouvelle probabilité pré-test, avec une cote de 0,350. Après le second test positif : 0,350 x 8,5 = 2,97, soit une probabilité de 2,97/3,97 = 74,8%. De manière équivalente, en une seule étape : 0,0526 x 6,64 x 8,5 = 2,97.

Interprétation. La prévalence de 33,8% de l'étude est artificielle, car elle résulte du schéma cas-témoins (22 patients et 43 témoins choisis). La valeur prédictive de 77,3% n'est pas transposable à la pratique. De plus, le schéma cas-témoins, avec des témoins sains, tend à surestimer la sensibilité et la spécificité (biais de spectre, discuté au chapitre 12), si bien que les RV réels peuvent être plus modestes. Le calcul séquentiel n'est valable que si les deux tests sont conditionnellement indépendants.

Problème résolu 11.5. Diagnostic différentiel d'une douleur thoracique au cabinet de médecine générale

Un patient de 55 ans consulte son médecin généraliste pour une douleur thoracique intermittente. Le médecin envisage trois diagnostics, supposés mutuellement exclusifs et exhaustifs, avec les probabilités a priori suivantes (données fictives, mais plausibles en soins primaires) : douleur musculo-squelettique 0,50 ; reflux gastro-œsophagien (RGO) 0,30 ; angor stable 0,20. Trois éléments de l'interrogatoire et de l'examen ont les fréquences indiquées dans le tableau.

Diagnostic	**p(D)**	$p(\text{ douleur à l'effort } / D)$	$p(\text{ douleur reproduite à la palpation } / D)$	$p(\text{ brûlure postprandiale } / D)$
Musculo-squelettique	0,50	0,20	0,70	0,05
RGO	0,30	0,10	0,10	0,70
Angor stable	0,20	0,80	0,15	0,15

Le patient décrit une douleur à l'effort, la douleur n'est pas reproduite à la palpation et il n'a pas de brûlures postprandiales. Calculez les probabilités a posteriori, en supposant l'indépendance conditionnelle des trois éléments.

Solution. Vraisemblance de la combinaison (effort +, palpation -, brûlures -) pour chaque diagnostic :

Musculo-squelettique : 0,20 x (1 - 0,70) x (1 - 0,05) = 0,20 x 0,30 x 0,95 = 0,0570

RGO : 0,10 x (1 - 0,10) x (1 - 0,70) = 0,10 x 0,90 x 0,30 = 0,0270

Angor : 0,80 x (1 - 0,15) x (1 - 0,15) = 0,80 x 0,85 x 0,85 = 0,5780

Produits avec les probabilités a priori : 0,50 x 0,0570 = 0,0285 ; 0,30 x 0,0270 = 0,0081 ; 0,20 x 0,5780 = 0,1156. La somme est 0,1522.

Probabilités a posteriori : musculo-squelettique 0,0285/0,1522 = 18,7% ; $\frac{\text{RGO } 0{,}0081}{0{,}1522} = 5{,}3\%$; angor 0,1156/0,1522 = 76,0%.

Interprétation. L'angor, le diagnostic le moins probable avant l'interrogatoire (20%), devient le plus probable (76,0%). Ce résultat justifie d'adresser le patient pour une évaluation cardiologique (ECG d'effort ou autre test fonctionnel) et la mise en route des mesures de prévention secondaire. Le résultat dépend cependant de l'hypothèse que les trois éléments sont indépendants dans chaque maladie et du fait que la liste des diagnostics est complète ; un syndrome coronarien aigu, par exemple, n'y figure pas et doit être exclu cliniquement avant tout calcul.

Problème résolu 11.6. Seuil de test et seuil de traitement

Pour une pharyngite aiguë chez un adolescent, le médecin estime qu'un traitement antibiotique se justifie si la probabilité d'une étiologie streptococcique dépasse 25%. Un test antigénique rapide a, dans notre scénario, une sensibilité de 86% et une spécificité de 95%. Entre quelles probabilités pré-test le test est-il utile ?

Solution. $\text{RV}^+ = \frac{0{,}86}{0{,}05} = 17{,}2$. $\text{RV}^- = \frac{0{,}14}{0{,}95} = 0{,}147$. Le seuil de traitement de 25% correspond à une cote de 0,25/0,75 = 0,333.

Seuil inférieur (seuil de test) : la cote pré-test pour laquelle un test positif atteint exactement le seuil de traitement est 0,333 / 17,2 = 0,0194, soit une probabilité de 0,0194/1,0194 = 1,9%.

Seuil supérieur (seuil test-traitement) : la cote pré-test pour laquelle un test négatif abaisse la probabilité exactement jusqu'au seuil de traitement est 0,333 / 0,147 = 2,27, soit une probabilité de 2,27/3,27 = 69,4%.

Interprétation. Pour des probabilités pré-test comprises entre 1,9% et 69,4%, le résultat du test peut modifier la décision, le test est donc utile. En dessous de 1,9%, même un test positif ne justifierait pas l'antibiotique (et la probabilité que le résultat positif soit faux est élevée). Au-dessus de 69,4%, même un test négatif laisserait la probabilité au-dessus de 25%, de sorte que le test ne modifie pas la décision. Le modèle ignore le coût et l'inconfort du test et repose sur un seuil de traitement choisi par jugement clinique ; les seuils réels sont fixés dans les recommandations.

Problème résolu 11.7. Inférence bayésienne pour une proportion sur un petit échantillon

Dans une étude pilote, 3 patients sur 20 traités par un nouveau médicament ont présenté une réaction indésirable cutanée. (a) Avec un a priori uniforme Bêta(1 ; 1), quelle est la distribution a posteriori, sa moyenne et l'intervalle de crédibilité à 95% ? (b) Une étude antérieure suggère une fréquence d'environ 10%, résumée par un a priori Bêta(2 ; 18) (moyenne 10%, intervalle a priori à 95% de 1,3% à 26,0%). Qu'est-ce qui change ?

Solution. (a) La distribution a posteriori est Bêta(1 + 3 ; 1 + 17) = Bêta(4 ; 18). La moyenne est 4/22 = 18,2%. L'intervalle de crédibilité à 95% (d'après les tables de la distribution Bêta ou le logiciel) est 5,4-36,3%.

(b) La distribution a posteriori est Bêta(2 + 3 ; 18 + 17) = Bêta(5 ; 35). La moyenne est 5/40 = 12,5%, et l'intervalle de crédibilité à 95% est 4,3-24,2%.

Interprétation. Avec seulement 20 patients, l'a priori compte : l'estimation baisse de 18,2% à 12,5% et l'intervalle se resserre, car l'a priori informatif ajoute pratiquement l'information de 18 patients supplémentaires observés auparavant. Comparez avec l'exemple de la section 11.9, où, avec 160 patients, les deux a priori donnaient des résultats presque identiques. Dans un rapport scientifique, les deux analyses doivent être présentées et le choix de l'a priori doit être justifié, par exemple par une référence à l'étude antérieure.

Problème résolu 11.8. Du facteur de Bayes à la probabilité de l'hypothèse

Dans le jeu mg2026, la comparaison de la PAS entre les sexes donne BF10 = 0,18, et l'association obésité-diabète donne un BF10 d'environ 85. Calculez la probabilité a posteriori de l'hypothèse H1 pour chaque analyse, en partant (a) d'une probabilité a priori de 50% et (b) d'une probabilité a priori de 20%.

Solution. Nous utilisons cote a posteriori = cote a priori x BF10.

PAS et sexe : (a) cote a priori 1 ; cote a posteriori 0,18 ; $p(H_1/\text{données}) = \frac{0,18}{1,18} = 15,3\%$. (b) Cote a priori 0,25 ; cote a posteriori 0,045 ; $p(H_1/\text{données}) = \frac{0,045}{1,045} = 4,3\%$.

Obésité et diabète : (a) 1 x 85 = 85 ; $p(H_1/\text{données}) = \frac{85}{86} = 98,8\%$. (b) 0,25 x 85 = 21,25 ; $p(H_1/\text{données}) = \frac{21,25}{22,25} = 95,5\%$.

Interprétation. Le facteur de Bayes décrit la force des preuves apportées par les données ; la probabilité finale de l'hypothèse dépend aussi de sa plausibilité initiale. L'association obésité-diabète est si bien étayée qu'elle reste très probable même avec un a priori sceptique. Pour la PAS et le sexe, les données rendent improbable l'hypothèse d'une différence. Remarquez l'analogie parfaite avec le diagnostic : le BF10 est le rapport de vraisemblance, et la probabilité a priori est la probabilité pré-test.

11.14 Exercices proposés

1. Dans une étude portant sur 500 patients, 60 sont atteints de la maladie B. Le symptôme S est présent chez 45 des malades et chez 88 des non-malades. Construisez le tableau maladie-symptôme et calculez p(B+), p(S+ / B+), p(S+ / B-) et p(B+ / S+). Vérifiez le dernier résultat à l'aide du théorème de Bayes.
2. Une mammographie de dépistage a une sensibilité de 90% et une spécificité de 93%. La prévalence du cancer du sein non diagnostiqué dans la population dépistée est de 0,8%. En utilisant 10 000 femmes et des fréquences naturelles, calculez la VPP et la VPN. Combien de femmes ayant une mammographie positive faut-il explorer davantage pour trouver un cancer ?
3. Transformez en cotes les probabilités 0,40 et 0,05, et transformez en probabilités les cotes 3 et 0,25.
4. Danila et collaborateurs (avec F. Cioca) ont rapporté, pour le dNLR chez les patients atteints de COVID-19 et de diabète, un seuil de 2,685 avec une sensibilité de 70,312%, une spécificité de 65,978% et une AUC de 0,624, et pour l'ALRI une sensibilité de 76,429% et une spécificité de 69,541% (AUC 0,752) [105]. (a) Calculez le RV+ et le RV- du dNLR. (b) Pour un patient ayant une probabilité pré-test de 30%, quelle est la probabilité post-test après un dNLR supérieur au seuil et après un dNLR inférieur au seuil ? (c) Commentez le nombre de décimales rapporté pour la sensibilité et la spécificité.

5. En utilisant les données du problème résolu 11.4 (2D-SWE, Se 93%, Sp 86%), calculez la probabilité post-test après un résultat négatif, pour une probabilité pré-test de 33,8% et de 5%.
6. Le test A a une Se de 0,90 et une Sp de 0,80 ; le test B a une Se de 0,85 et une Sp de 0,95. En supposant l'indépendance conditionnelle, calculez la sensibilité et la spécificité de la combinaison en série et en parallèle. Quelle stratégie choisiriez-vous pour le tri aux urgences et quelle stratégie pour la confirmation avant un traitement toxique ?
7. Dans un service d'urgences, un patient présentant une céphalée aiguë a trois diagnostics possibles, avec les probabilités a priori suivantes : migraine 0,70 ; céphalée de tension 0,25 ; hémorragie sous-arachnoïdienne (HSA) 0,05. Les fréquences des éléments cliniques sont : début brutal, "en coup de tonnerre" (0,05 / 0,02 / 0,80) ; raideur de nuque (0,05 / 0,10 / 0,60) ; nausées (0,70 / 0,10 / 0,70), dans l'ordre migraine / céphalée de tension / HSA. Le patient présente les trois éléments. Calculez les probabilités a posteriori, en supposant l'indépendance conditionnelle, et commentez la décision.
8. Le seuil de traitement d'une maladie est de 10%. Le test disponible a un RV+ = 4 et un RV- = 0,25. Calculez le seuil de test et le seuil test-traitement.
9. Ion et collaborateurs (avec F. Cioca) ont comparé des scores cliniques pour la prédiction de la COVID-19 sévère chez les patients diabétiques [106]. Le SOFA a eu une AUC de 0,771 à l'admission et de 0,873 au cinquième jour ; le NEWS2 de 0,892 à l'admission et de 0,729 au cinquième jour ; l'APACHE II de 0,698 à l'admission et de 0,806 au cinquième jour. (a) Situez chaque valeur sur l'échelle de l'AUC présentée dans le cours (chapitre 12). (b) Quel score utiliseriez-vous à l'admission et quel score après cinq jours ? (c) L'étude décrit une régression logistique, mais rapporte des rapports de risques instantanés (HR, par exemple SOFA au cinquième jour HR 3,07, IC à 95% 2,29-4,12). Quelle mesure d'effet aurait-on attendue d'une régression logistique ?
10. Une analyse bayésienne donne BF10 = 0,16 pour une corrélation. (a) Exprimez le résultat sous forme de BF01 et situez-le sur l'échelle du tableau 11.8. (b) Si la probabilité a priori de l'existence de la corrélation était de 50%, quelle est la probabilité a posteriori ?
11. Formulez selon la structure PICO une question clinique sur l'utilité de l'HbA1c par rapport à l'HGPO pour le diagnostic du diabète au cabinet de médecine générale et identifiez le type d'étude le plus approprié pour y répondre.

Pièges fréquents

- Confondre p(S / B) avec p(B / S). La sensibilité d'un test n'est pas la probabilité qu'un patient dont le test est positif soit malade.
- Ignorer la prévalence (négligence du taux de base) : un test ayant une Se et une Sp de 95% peut avoir une VPP inférieure à 50% si la maladie est rare.
- Utiliser la VPP ou la VPN rapportées dans une étude cas-témoins comme si elles étaient valables au cabinet. La prévalence d'une telle étude est fixée par le chercheur.
- Multiplier les rapports de vraisemblance de tests corrélés (par exemple FINDRISC et glycémie), ce qui surestime la probabilité finale.
- Confondre cote et probabilité : une cote de 0,25 correspond à une probabilité de 0,20, et non de 0,25.
- Interpréter l'intervalle de confiance fréquentiste comme un intervalle de crédibilité ("95% de probabilité que le paramètre se trouve ici").
- Lire un BF10 faible comme un "résultat non concluant". Un BF10 de 0,1 est une preuve en faveur de H0, et non une absence de preuves.

À retenir

- Théorème de Bayes : $p(B/S) = \frac{p(S/B) \times p(B)}{p(S)}$, avec p(S) donnée par la formule des probabilités totales.
- Forme en cotes : cote post-test = cote pré-test x RV ; $RV^+ = \frac{Se}{1-Sp}$; $RV^- = \frac{1-Se}{Sp}$.
- La probabilité pré-test dépend du patient et du contexte (cabinet, consultation externe, hôpital) ; les valeurs prédictives en dépendent.
- Les fréquences naturelles (1000 patients, arbre) sont le moyen le plus sûr de résoudre un problème de diagnostic et d'expliquer un résultat à un patient.
- En série : la Sp et la VPP augmentent ; en parallèle : la Se et la VPN augmentent. Les formules simples supposent l'indépendance conditionnelle.
- Diagnostic différentiel bayésien : produit a priori x vraisemblance pour chaque maladie, divisé par la somme des produits.
- Inférence bayésienne : a posteriori $\propto$ vraisemblance x a priori ; l'intervalle de crédibilité et le facteur de Bayes ont des interprétations directes, mais exigent un a priori justifié.
- Médecine fondée sur les preuves : cinq étapes (PICO, recherche, évaluation critique, application, évaluation), niveau de confiance dans les preuves évalué par GRADE, revues systématiques rapportées selon PRISMA 2020.

Questions d'autoévaluation

1. (QRU) Dans le tableau maladie-symptôme avec les notations du cours (a, b sur la ligne B+ ; c, d sur la ligne B-), la probabilité conditionnelle p(S+ / B-) est :
   a) $\frac{c}{N}$
   b) $\frac{c}{a+c}$
   c) $\frac{c}{c+d}$
   d) $\frac{a+c}{N}$
   e) $\frac{d}{c+d}$
2. (QRU) Un test ayant une sensibilité de 80% et une spécificité de 90% est appliqué dans une population où la prévalence de la maladie est de 10%. La valeur prédictive positive est d'environ :
   a) 80%
   b) 47%
   c) 90%
   d) 8%
   e) 72%
3. (QRU) Une probabilité pré-test de 0,20 correspond à une cote de :
   a) 0,20
   b) 0,25
   c) 0,80
   d) 4
   e) 0,16
4. (QRU) Un résultat de test dont le rapport de vraisemblance est égal à 1 :
   a) confirme la maladie
   b) exclut la maladie
   c) double la probabilité pré-test
   d) ne modifie pas la probabilité de la maladie

e) a une sensibilité de 100%

5. (QRU) Une analyse bayésienne rapporte BF10 = 0,05. L'interprétation correcte est :
   a) preuves fortes en faveur de l'hypothèse H1
   b) p = 0,05, résultat à la limite de la significativité
   c) preuves fortes en faveur de l'hypothèse H0 (BF01 = 20)
   d) résultat non concluant, on ne peut rien dire
   e) la probabilité que H0 soit vraie est de 5%
6. (QRM) À propos des tests en série (le second test n'est réalisé que chez les sujets positifs au premier, et le patient est considéré comme positif si les deux tests sont positifs) :
   a) la spécificité de la combinaison augmente par rapport à chaque test pris séparément
   b) la sensibilité de la combinaison augmente par rapport à chaque test pris séparément
   c) la valeur prédictive positive augmente
   d) cette stratégie est plus économique, car le second test n'est appliqué qu'à une partie des patients
   e) la formule Se = Se1 x Se2 n'est valable que si les tests sont conditionnellement indépendants
7. (QRM) À propos des rapports de vraisemblance :
   a) $RV^+ = \frac{Se}{1-Sp}$
   b) $RV^- = \frac{1-Se}{Sp}$
   c) ils dépendent fortement de la prévalence de la maladie, comme les valeurs prédictives
   d) cote post-test = cote pré-test x RV
   e) un RV inférieur à 1 augmente la probabilité de la maladie
8. (QRM) À propos de la médecine fondée sur les preuves :
   a) elle intègre les preuves issues de la recherche à l'expertise clinique et aux valeurs du patient
   b) dans la structure PICO, C désigne le comparateur
   c) dans le système GRADE, les preuves issues des études observationnelles partent d'un niveau de confiance élevé
   d) une revue narrative se situe au-dessus d'une revue systématique dans la hiérarchie des preuves
   e) l'hétérogénéité d'une méta-analyse peut être quantifiée par I^2
9. (QRM) À propos des erreurs cognitives dans le raisonnement clinique :
   a) la négligence du taux de base consiste à ignorer la prévalence de la maladie
   b) l'ancrage désigne l'ajustement insuffisant de la première impression lorsque de nouvelles données apparaissent
   c) la disponibilité conduit à surestimer la probabilité des maladies dont on se souvient facilement
   d) le biais de vérification est une erreur cognitive du médecin au lit du malade
   e) la clôture prématurée désigne la poursuite des examens après l'établissement du diagnostic
10. (QRM) À propos de l'inférence bayésienne :

a) la distribution a posteriori est proportionnelle à la vraisemblance multipliée par la distribution a priori

b) à mesure que l'échantillon augmente, l'influence de l'a priori sur le résultat diminue

c) un intervalle de crédibilité à 95% peut s'interpréter comme une probabilité de 95% que le paramètre se trouve dans l'intervalle, compte tenu de l'a priori et des données

d) le facteur de Bayes peut apporter des preuves en faveur de l'hypothèse nulle

e) un a priori Bêta(1 ; 1) favorise les petites valeurs d'une proportion

Chapitre 12. Évaluation de la performance des tests diagnostiques : sensibilité, spécificité, valeurs prédictives et courbe ROC

Objectifs

À la fin de ce chapitre, l'étudiant sera capable de :

- définir le standard de référence et le test à évaluer, et décrire le schéma correct d'une étude d'exactitude diagnostique ;
- construire le tableau 2 x 2 (T+/T- versus B+/B-) avec les notations a, b, c, d et calculer la prévalence, la sensibilité, la spécificité, les valeurs prédictives, l'exactitude, le taux d'erreur de classification, l'indice de Youden, les rapports de vraisemblance et le rapport de cotes diagnostique ;
- reconstituer un tableau 2 x 2 complet à partir de quelques indicateurs et de quelques cellules connues ;
- expliquer pourquoi les valeurs prédictives dépendent de la prévalence et les recalculer pour une autre population ;
- reconnaître le biais de spectre, le biais de vérification et le biais d'incorporation dans un article publié ;
- construire à la main une courbe ROC à partir d'un tableau de seuils, interpréter l'aire sous la courbe et son intervalle de confiance, et comparer deux courbes ROC ;
- choisir un seuil de décision (indice de Youden, seuil d'exclusion, seuil de confirmation) en fonction de l'objectif clinique ;
- calculer et interpréter le coefficient kappa de Cohen et reconnaître les situations dans lesquelles on utilise le coefficient de corrélation intraclasse ;
- appliquer les mêmes indicateurs à l'évaluation d'un classificateur d'intelligence artificielle et utiliser les listes STARD 2015 et STARD-AI pour la lecture critique d'une étude.

12.1 Test à évaluer et standard de référence

Tout test diagnostic, d'une question de l'interrogatoire à un biomarqueur ou à un algorithme d'intelligence artificielle, doit répondre à une question simple : dans quelle mesure sépare-t-il les patients qui ont la maladie de ceux qui ne l'ont pas ? Pour y répondre, il nous faut deux choses. La première est le test à évaluer (angl. index test), c'est-à-dire le test nouveau, plus simple, moins coûteux ou moins invasif, dont nous voulons connaître la performance. La seconde est le standard de référence (angl. reference standard, appelé traditionnellement "gold standard" ou étalon-or), c'est-à-dire la méthode considérée comme la meilleure disponible pour établir la présence ou l'absence de la maladie.

Dans le jeu mg2026, le test à évaluer est le score FINDRISC, et le standard de référence est l'hyperglycémie provoquée par voie orale (HGPO). Dans les travaux de nos collègues de Timișoara, les standards de référence sont très variés : la biopsie hépatique avec stadification Metavir pour la fibrose [107], l'histologie pour les lésions cervicales [131], l'endoscopie digestive haute pour les varices œsophagiennes [110], le diagnostic final établi sur l'ensemble des données disponibles pour l'obstruction biliaire maligne [57]. Parfois, le standard de référence est lui-même une méthode non invasive, choisie parce qu'elle est mieux validée : l'élastographie impulsionnelle

(VCTE) pour l'élastographie time-harmonic [3] ou le paramètre d'atténuation contrôlée (CAP) pour les nouvelles méthodes de quantification de la stéatose [58,65].

Le standard de référence n'est presque jamais parfait. La biopsie hépatique analyse un fragment de quelques millimètres d'un organe entier ; l'histologie dépend de l'anatomopathologiste ; l'HGPO a une reproductibilité modeste. Lorsque le standard de référence se trompe, le test à évaluer paraîtra moins bon qu'il ne l'est, car les désaccords seront automatiquement attribués au nouveau test. Il faut garder cela à l'esprit lorsqu'un nouveau test "perd" face à une référence imparfaite.

12.1.1 Schéma d'une étude d'exactitude diagnostique

Le schéma idéal est transversal : une série de patients consécutifs, représentatifs de ceux chez qui le test sera utilisé en pratique (par exemple des adultes vus au cabinet de médecine générale pour une suspicion de la maladie), chez qui l'on réalise à la fois le test à évaluer et le standard de référence, de façon indépendante et sans que l'interprète de l'un connaisse le résultat de l'autre. La prévalence de la maladie dans l'échantillon est alors une estimation de la prévalence réelle, et tous les indicateurs, y compris les valeurs prédictives, ont un sens pour la population concernée.

Le schéma cas-témoins diagnostique (malades connus comparés à des volontaires sains) est plus facile à réaliser et il est fréquemment utilisé dans les premières phases de l'évaluation d'un test. Il pose cependant deux problèmes majeurs. Premièrement, la prévalence est fixée par le chercheur (par exemple 1:2 cas et témoins), de sorte que les valeurs prédictives calculées n'ont pas de signification clinique. Deuxièmement, la comparaison entre des malades typiques, souvent à des stades avancés, et des personnes en parfaite santé exagère l'exactitude, car il manque précisément les cas difficiles (la maladie débutante, les patients atteints d'autres maladies qui imitent la maladie étudiée). Nous reviendrons sur ce biais de spectre à la section 12.5.

12.2 Le tableau 2 x 2 et les indicateurs de performance

12.2.1 Les notations

Le cours de la discipline utilise un tableau avec la maladie en lignes et le résultat du test en colonnes. Nous conservons ces notations dans tout le chapitre.

Tableau 12.1. Tableau 2 x 2 pour l'évaluation d'un test diagnostique, notations du cours.

	T+	**T-**	**Total**
B+	a (vrais positifs, VP)	b (faux négatifs, FN)	a + b
B-	c (faux positifs, FP)	d (vrais négatifs, VN)	c + d
Total	a + c	b + d	$N = a + b + c + d$

La cellule a contient les malades dont le test est positif (vrais positifs), b les malades dont le test est négatif (faux négatifs), c les personnes non atteintes de la maladie étudiée mais dont le test est positif (faux positifs), et d les personnes sans la maladie et dont le test est négatif (vrais négatifs). Dans certains manuels et logiciels (par exemple MedCalc), le tableau est transposé, avec la maladie en colonnes et le test en lignes ; les lettres b et c échangent alors leur place. C'est pourquoi, avant tout calcul, vérifiez toujours ce que représente chaque cellule, et non quelle lettre elle porte.

12.2.2 Prévalence, sensibilité et spécificité

La prévalence (dans l'échantillon) est la proportion de malades :

prévalence = (a + b) / N

La sensibilité (Se) est la probabilité qu'un malade ait un test positif, c'est-à-dire la proportion de malades correctement identifiés :

$$\text{Se} = p(T^+/B^+) = \frac{a}{a+b}$$

La spécificité (Sp) est la probabilité qu'une personne non atteinte de la maladie étudiée ait un test négatif :

$$\text{Sp} = p(T^-/B^-) = \frac{d}{c+d}$$

La sensibilité se calcule uniquement sur la ligne des malades, la spécificité uniquement sur la ligne des non-malades. Par conséquent, en principe, elles ne dépendent pas de la prévalence : si nous doublons le nombre de sujets sains dans l'étude, la spécificité reste la même, et la sensibilité ne s'en aperçoit même pas. Nous disons "en principe" parce que la sensibilité et la spécificité dépendent du spectre des patients (stade de la maladie, comorbidités), et que le spectre change souvent en même temps que la prévalence.

Deux moyens mnémotechniques anglais sont utiles au lit du malade. SnNout (angl. high Sensitivity, Negative result, rules out) : un test très sensible, s'il est négatif, exclut la maladie, car un test sensible manque peu de malades. SpPin (angl. high Specificity, Positive result, rules in) : un test très spécifique, s'il est positif, confirme la maladie, car un test spécifique donne peu de faux positifs. Ces moyens mnémotechniques sont des approximations : un test ayant une Se de 95% et une Sp de 40% n'"exclut" bien la maladie que si la probabilité pré-test n'est pas très élevée, comme nous l'avons vu au chapitre 11.

12.2.3 Les valeurs prédictives

La valeur prédictive positive (VPP) est la probabilité qu'une personne dont le test est positif ait la maladie, c'est-à-dire la proportion de malades parmi les sujets dont le test est positif :

$$\text{VPP} = p(B^+/T^+) = \frac{a}{a+c}$$

La valeur prédictive négative (VPN) est la proportion de sujets non malades parmi ceux dont le test est négatif :

$$\text{VPN} = p(B^-/T^-) = \frac{d}{b+d}$$

Les valeurs prédictives se calculent sur les colonnes. Elles répondent à la question que se pose le patient ("mon test est positif ; quelle est la probabilité que j'aie la maladie ?"), mais dépendent fortement de la prévalence, comme nous le verrons à la section 12.4. Les valeurs prédictives sont en fait les probabilités post-test du chapitre 11, calculées pour une probabilité pré-test égale à la prévalence de l'échantillon [133].

12.2.4 Exactitude et taux d'erreur de classification

L'exactitude (Ex) est la proportion de l'ensemble des personnes correctement classées par le test :

$$\text{Ex} = \frac{a+d}{N}$$

Le taux d'erreur de classification (TE) est la proportion des sujets mal classés, c'est-à-dire des faux négatifs et des faux positifs :

$$\text{TE} = \frac{b+c}{N} = 1 - \text{Ex}$$

L'exactitude paraît l'indicateur le plus intuitif, mais elle peut être trompeuse. Elle est une moyenne pondérée de la sensibilité et de la spécificité, les poids étant donnés par la prévalence :

Ex = Se x prévalence + Sp x (1 - prévalence). Lorsque la maladie est rare, l'exactitude est dominée par la spécificité. Un "test" qui déclare tout le monde sain a, pour une prévalence de 2%, une exactitude de 98%, bien qu'il ne détecte aucun malade. C'est pourquoi l'exactitude n'est jamais rapportée seule.

> **La règle qui évite l'erreur**
>
> Le taux d'erreur de classification se calcule à partir des cellules qui contiennent des erreurs : les faux négatifs (b) et les faux positifs (c). La formule correcte est $\text{TE} = \frac{b+c}{N} = \frac{\text{FN+FP}}{N}$. La formule (b + d) / N, que l'on rencontre parfois dans des notes de cours, est fausse : d contient les vrais négatifs, c'est-à-dire des personnes correctement classées. Une vérification rapide : Ex + TE doit être exactement égal à 1. Pour le tableau 150 / 30 / 4 / 256 du problème résolu 12.1, $\frac{a+d}{N} = \frac{406}{440} = 0{,}923$, et $\frac{b+c}{N} = \frac{34}{440} = 0{,}077$; la somme est 1.

12.2.5 Indice de Youden, rapports de vraisemblance et rapport de cotes diagnostique

L'indice de Youden combine la sensibilité et la spécificité en un seul chiffre [134] :

$$J = \text{Se} + \text{Sp} - 1$$

J varie entre 0 (le test ne discrimine pas mieux qu'un tirage à pile ou face) et 1 (test parfait). Il a l'avantage de ne pas dépendre de la prévalence et l'inconvénient d'accorder la même importance aux faux négatifs et aux faux positifs, ce qui est rarement vrai sur le plan clinique. Nous l'utiliserons pour choisir le seuil sur la courbe ROC.

Les rapports de vraisemblance, introduits au chapitre 11, se calculent eux aussi à partir de la Se et de la Sp [121] :

$$\text{RV}^+ = \frac{\text{Se}}{1 - \text{Sp}}$$

$$\text{RV}^- = \frac{1 - \text{Se}}{\text{Sp}}$$

Le rapport de cotes diagnostique (angl. diagnostic odds ratio, DOR) est le rapport entre la cote d'un test positif chez les malades et la cote d'un test positif chez les non-malades :

$$\text{DOR} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \times d}{b \times c} = \frac{\text{RV}^+}{\text{RV}^-}$$

C'est le même calcul que celui du rapport de cotes de l'analyse du risque, appliqué au tableau diagnostique. Un DOR de 1 correspond à un test inutile ; des valeurs de l'ordre de plusieurs dizaines ou centaines indiquent un test puissant. Le DOR est utile pour comparer des tests dans les méta-analyses, mais il est difficile à interpréter cliniquement et n'indique pas si le test est meilleur pour confirmer ou pour exclure : deux tests ayant respectivement une Se de 90% / Sp de 50% et une Se de 50% / Sp de 90% ont exactement le même DOR (9).

Tableau 12.2. Récapitulatif des indicateurs de performance diagnostique (notations du cours).

Indicateur	Formule	Dépend de la prévalence ?	Question à laquelle il répond
Prévalence	$\frac{a+b}{N}$	(c'est la prévalence)	Quelle est la fréquence de la maladie dans l'échantillon ?
Sensibilité	$\frac{a}{a+b}$	non (en principe)	Quelle proportion des malades le test détecte-t-il ?
Spécificité	$\frac{d}{c+d}$	non (en principe)	Quelle proportion des sujets sains est négative ?

Indicateur	Formule	Dépend de la prévalence ?	Question à laquelle il répond
VPP	$\frac{a}{a+c}$	oui	Un sujet positif : quelle probabilité qu'il soit malade ?
VPN	$\frac{d}{b+d}$	oui	Un sujet négatif : quelle probabilité qu'il soit sain ?
Exactitude	$\frac{a+d}{N}$	oui	Quelle proportion est correctement classée ?
Taux d'erreur	$\frac{b+c}{N}$	oui	Quelle proportion est mal classée ?
Indice de Youden	Se + Sp - 1	non	Quelle est la qualité globale de la discrimination ?
RV+	$\frac{\text{Se}}{1-\text{Sp}}$	non	De combien de fois la cote augmente-t-elle après un résultat positif ?
RV-	$\frac{1-\text{Se}}{\text{Sp}}$	non	De combien de fois la cote diminue-t-elle après un résultat négatif ?
DOR	$\frac{a \times d}{b \times c}$	non	Rapport des cotes d'un test positif

12.2.6 Les indicateurs sur le jeu mg2026 et leurs intervalles de confiance

Au seuil FINDRISC ≥ 12, le tableau pour le diabète confirmé par HGPO est : a = 22 (VP), b = 7 (FN), c = 33 (FP), d = 98 (VN), avec N = 160 et 29 malades (prévalence 18,1%).

Tableau 12.3. FINDRISC ≥ 12 versus HGPO, jeu mg2026 (données synthétiques).

	FINDRISC ≥ 12	FINDRISC < 12	Total
Diabète (HGPO)	22	7	29
Sans diabète	33	98	131
Total	55	105	160

$\text{Se} = \frac{22}{29} = 75{,}9\%$; $\text{Sp} = \frac{98}{131} = 74{,}8\%$; $\text{VPP} = \frac{22}{55} = 40{,}0\%$; $\text{VPN} = \frac{98}{105} = 93{,}3\%$; $\text{Ex} = \frac{120}{160} = 75{,}0\%$; $\text{TE} = \frac{40}{160} = 25{,}0\%$; J = 0,759 + 0,748 - 1 = 0,507 ; RV+ = 3,01 ; RV- = 0,32 ; $\text{DOR} = \frac{22 \times 98}{7 \times 33} = \frac{2156}{231} = 9{,}33$.

Chacun de ces indicateurs est une proportion estimée sur un échantillon et doit être accompagné d'un intervalle de confiance. Pour les proportions, on utilise de préférence la méthode de Wilson, présentée au chapitre 4. La sensibilité repose uniquement sur les 29 malades, son intervalle est donc large : Se 75,9% (IC à 95% 57,9-87,8). La spécificité, calculée sur 131 non-malades, est plus précise : 74,8% (IC à 95% 66,7-81,5). VPP 40,0% (IC à 95% 28,1-53,2), VPN 93,3% (IC à 95% 86,9-96,7). Pour les rapports de vraisemblance, on utilise une méthode sur échelle logarithmique : RV+ 3,01 (IC à 95% 2,10-4,31), RV- 0,32 (IC à 95% 0,17-0,62), et DOR 9,33 (IC à 95% 3,65-23,84).

C'est une leçon que vous retrouverez dans chaque article d'exactitude diagnostique : la sensibilité est presque toujours estimée de façon moins précise que la spécificité, parce que les malades sont moins nombreux. Une étude portant sur 30 malades ne peut pas estimer la sensibilité

avec une précision meilleure qu'environ ±15 points de pourcentage, quel que soit le nombre de décimales rapporté [135].

12.3 Reconstitution d'un tableau 2 x 2

Un type de problème propre à notre cours, fréquent à l'examen, demande de reconstituer le tableau entier à partir de quelques indicateurs et de quelques cellules. Ces problèmes paraissent compliqués, mais ils se résolvent toujours avec la même stratégie, en quatre étapes.

Étape 1. Écrivez le tableau vide, avec les lettres a, b, c, d et les totaux marginaux, et notez la formule de chaque indicateur donné : Se $= \frac{a}{a+b}$, Sp $= \frac{d}{c+d}$, VPP $= \frac{a}{a+c}$, VPN $= \frac{d}{b+d}$, prévalence = (a + b)/N.

Étape 2. Remplissez les cellules données directement (par exemple "VP = 90" signifie a = 90).

Étape 3. Choisissez un indicateur qui ne contient qu'une seule inconnue et résolvez l'équation. Par exemple, si VPP $= \frac{9}{12}$ et a = 90, alors $\frac{90}{90+c} = \frac{9}{12}$, d'où 9(90 + c) = 12 x 90, donc c = 30. Répétez jusqu'à ce que toutes les cellules soient connues.

Étape 4. Calculez les totaux et vérifiez les indicateurs donnés. Calculez ensuite les indicateurs demandés.

Une erreur fréquente consiste à utiliser un indicateur à deux inconnues avant d'avoir trouvé l'une d'elles. Un autre type d'erreur survient lorsque les rapports sont donnés sous forme de fractions (2/3, 6/8) : ne les transformez pas en décimales arrondies (0,66) avant la résolution, travaillez avec la fraction, sinon les cellules ne tombent plus sur des nombres entiers. Les problèmes résolus 12.2-12.5 appliquent cette stratégie pas à pas.

12.4 Dépendance des valeurs prédictives à l'égard de la prévalence

La sensibilité et la spécificité sont des propriétés du test (pour un spectre donné de patients). Les valeurs prédictives sont des propriétés du test appliqué à une population déterminée. Le théorème de Bayes, écrit avec les indicateurs de ce chapitre, montre précisément ce lien :

$$\text{VPP} = \frac{\text{Se} \times P}{\text{Se} \times P + (1 - \text{Sp}) \times (1 - P)}$$

$$\text{VPN} = \frac{\text{Sp} \times (1 - P)}{\text{Sp} \times (1 - P) + (1 - \text{Se}) \times P}$$

où P est la prévalence. À mesure que la prévalence augmente, la VPP augmente et la VPN diminue.

La figure 12.1 illustre cette dépendance pour FINDRISC ≥ 12 (Se 76%, Sp 75%), et le tableau 12.4 donne les valeurs exactes calculées sur le jeu mg2026 pour quelques prévalences.

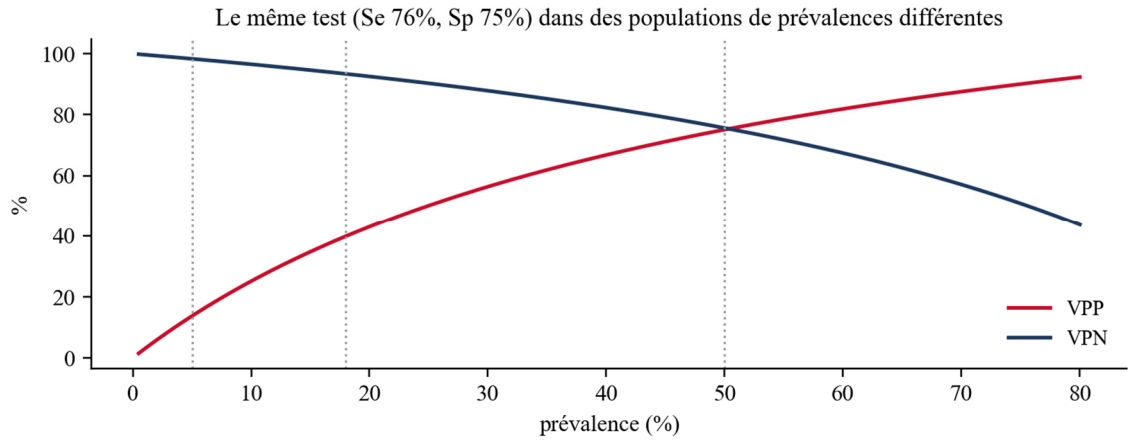


Figure 12.1. Valeur prédictive positive et négative d'un même test (Se 76%, Sp 75%, FINDRISC ≥ 12) dans des populations de prévalences différentes. Les lignes pointillées marquent les prévalences de 5%, 18% et 50%. Données synthétiques (jeu mg2026).

Tableau 12.4. VPP et VPN du FINDRISC ≥ 12 (Se 75,9%, Sp 74,8%) pour différentes prévalences du diabète non diagnostiqué. Données synthétiques (jeu mg2026).

Prévalence	2%	5%	10%	18%	30%	50%
VPP (%)	5,8	13,7	25,1	39,8	56,3	75,1
VPN (%)	99,3	98,3	96,5	93,4	87,9	75,6

Pour une prévalence de 2% (par exemple de jeunes adultes sans facteurs de risque), sur 100 personnes ayant un FINDRISC ≥ 12, seules 6 environ ont un diabète. Pour une prévalence de 50% (par exemple des patients présentant une polyurie, une polydipsie et un amaigrissement), 75 sur 100 ont un diabète. Le test est le même ; c'est la population qui change. La conclusion pratique a deux faces. Les valeurs prédictives rapportées dans un article ne s'appliquent qu'aux populations de prévalence comparable. Pour les transposer à une autre population, on les recalcule à partir de la Se, de la Sp et de la nouvelle prévalence (ou, de manière équivalente, avec les rapports de vraisemblance) [133,136].

Cette dépendance explique pourquoi le dépistage dans la population générale produit autant de résultats faux positifs et pourquoi les tests de confirmation sont indispensables. Elle explique aussi pourquoi le même biomarqueur paraît excellent dans une étude hospitalière et décevant au cabinet.

12.5 Biais dans les études d'exactitude diagnostique

Des indicateurs correctement calculés à partir d'un tableau correct peuvent néanmoins être trompeurs si le tableau provient d'une étude dont le schéma est défectueux. Trois biais sont si fréquents que tout médecin doit savoir les reconnaître.

Le biais de spectre apparaît lorsque les patients de l'étude ne ressemblent pas à ceux chez qui le test sera utilisé. Sa forme la plus fréquente est le schéma cas-témoins avec des malades à des stades avancés et des témoins en parfaite santé. Les malades à un stade avancé ont des valeurs du test très perturbées (sensibilité apparemment élevée), et les sujets sains n'ont aucune maladie susceptible de produire des faux positifs (spécificité apparemment élevée). En pratique, le test rencontrera des cas débutants et des patients atteints d'autres maladies, et sa performance diminuera. Le biais de spectre n'est pas seulement un problème de schéma d'étude : il traduit le fait que la sensibilité et la spécificité varient réellement selon le type de patient. L'étude de Pavel

et collaborateurs sur la H-FABP, discutée au chapitre 11, a montré une sensibilité et une spécificité de 89,67 et 95,65 moins de 4 heures après le début de la douleur, mais de 86,73 et 49,84 au-delà de 4 heures [127]. Sipos et collaborateurs ont observé, pour l'IL-6 dans le sepsis néonatal, une sensibilité de 90% dans un sous-groupe et de 71,4% dans un autre, selon le statut des cultures [126].

Le biais de vérification (angl. verification bias ou work-up bias) apparaît lorsque le standard de référence n'est réalisé que chez une partie des patients et que la sélection dépend du résultat du test à évaluer. Typiquement, les patients dont le test est positif sont adressés pour une biopsie ou une coloscopie, et ceux dont le test est négatif ne le sont pas. Les faux négatifs restent alors non découverts (la cellule b est sous-estimée), et la sensibilité paraît plus élevée qu'elle ne l'est. Dans l'étude de Moleriu et collaborateurs, sur 184 patientes ayant une cytologie LSIL, seules 64 ont été biopsiées [131] ; dans l'étude d'Iovanescu et collaborateurs, 252 patients sur 2201 ont eu une coloscopie, et l'adhésion à la coloscopie parmi les sujets FIT positifs a été de 52% [101]. Dans les deux situations, les critères selon lesquels on a décidé de la vérification influencent les estimations, et les auteurs doivent les décrire.

Le biais d'incorporation apparaît lorsque le résultat du test à évaluer fait partie de la définition du standard de référence. Si le diagnostic final de pancréatite est établi, entre autres, sur la base de la lipase, nous ne pouvons pas ensuite évaluer correctement l'exactitude de la lipase : le test "se confirme lui-même". Le problème est fréquent lorsque le standard de référence est un diagnostic clinique final, construit à partir de toutes les données disponibles, y compris le test évalué. L'étude de Rațiu et collaborateurs a comparé la CPRE, la cholangio-IRM et l'histologie au diagnostic final de l'obstruction biliaire maligne [57] ; face à un tel schéma, le lecteur doit vérifier si le diagnostic final a utilisé les résultats des tests comparés.

D'autres sources de biais sont l'interprétation du test en connaissant le résultat du standard de référence (biais d'interprétation, prévenu par l'insu), l'exclusion des résultats ininterprétables (par exemple les mesures élastographiques non valides) et l'utilisation du même échantillon à la fois pour choisir le seuil et pour estimer la performance à ce seuil (surestimation optimiste, corrigée par une validation sur un échantillon distinct).

Étude de cas : le schéma cas-témoins et la performance apparente des tests [2,116,132]

Trois études de nos collègues utilisent un schéma de type cas-témoins et constituent un bon exercice de lecture critique. Lupușoru et collaborateurs ont évalué l'échographie de contraste avec imagerie paramétrique du temps d'arrivée (CEUS-PAT) pour la cirrhose hépatique, chez 64 personnes : 37 atteintes de cirrhose et 27 volontaires sains. Les valeurs de CEUS-PAT ont été de 0,83 ± 0,09 chez les cirrhotiques et de 0,49 ± 0,11 chez les sujets sains ($p < 0,0001$), la corrélation avec la VCTE a été de $r = 0,81$, et au seuil > 0,7 l'aire sous la courbe ROC a été de 0,98, avec une Se de 89,19%, une Sp de 100%, une VPP de 100% et une VPN de 87,1%. Les chiffres peuvent être reconstitués exactement : 33 cirrhotiques sur 37 ont été positifs (89,19%), les 27 volontaires ont tous été négatifs (100%), soit 31 négatifs, dont 27 sains $\left(\frac{\text{VPN } 27}{31} = 87{,}1\%\right)$. Golu et collaborateurs (avec L. Moleriu et A. Tudor) ont étudié l'élastographie 2D-SWE chez 22 patients atteints d'hyperparathyroïdie et 43 témoins sains ; les valeurs moyennes de rigidité rapportées dans le résumé ont été de 10,2 ± 4,9 kPa et de 19,5 ± 7,6 kPa ($p = 0,007$), et au seuil de 12,5 kPa, Se 93%, Sp 86% et AUC 0,949. Feier et collaborateurs (avec C. Muntean) ont comparé 145 cholécystites gangréneuses à 290 cholécystites chroniques, selon un rapport fixé de 1:2, appariées sur l'âge et le sexe ; le meilleur indice inflammatoire a eu une AUC de 0,889, une Se de 79,5% et une Sp de 86,3%. Qu'ont en commun ces trois études ? La prévalence est fixée par le schéma d'étude (57,8%, 33,8% et 33,3%), de sorte que les valeurs prédictives ne s'appliquent pas directement à la

> pratique. Lorsque les témoins sont des volontaires sains (les deux premières études), les AUC très élevées (0,98 et 0,949) représentent probablement la limite supérieure de la performance ; les patients atteints d'hépatopathie chronique non cirrhotique ou de nodules thyroïdiens auraient été les comparateurs cliniquement pertinents. L'étude de Feier est plus proche de la situation clinique, car les témoins ont eux aussi une cholécystite, mais le rapport 1:2 reste artificiel. Ces trois études sont des étapes nécessaires de l'évaluation d'un test ; l'étape suivante serait une cohorte de patients consécutifs présentant une suspicion clinique.

12.6 Tests à résultat continu et choix du seuil

12.6.1 Le compromis entre sensibilité et spécificité

Beaucoup de tests n'ont pas par nature un résultat "positif" ou "négatif", mais une valeur numérique : la glycémie, la troponine, la rigidité hépatique, un score clinique. Pour les utiliser dans la décision, nous choisissons un seuil (angl. cut-off) : les valeurs supérieures au seuil sont considérées comme positives. Le choix du seuil modifie le tableau 2 x 2 et, avec lui, tous les indicateurs.

Le mécanisme apparaît dans le panneau B de la figure 12.3, plus bas : la distribution du score FINDRISC chez les personnes diabétiques (moyenne 14,76, ÉT 3,96) est décalée vers la droite par rapport à la distribution chez les personnes non diabétiques (moyenne 9,40, ÉT 3,70), mais les deux distributions se chevauchent. Où que nous placions le seuil, certains malades resteront en dessous (faux négatifs) et certains sujets sains se trouveront au-dessus (faux positifs). Si nous déplaçons le seuil vers la gauche (valeurs plus faibles), nous détectons plus de malades (la sensibilité augmente), mais nous déclarons aussi positifs davantage de sujets sains (la spécificité diminue). Si nous le déplaçons vers la droite, c'est l'inverse. Il n'existe pas de seuil qui les augmente toutes les deux ; il n'existe que des compromis. Le cours de la discipline décrit cette relation comme une relation de "proportionnalité inverse" entre sensibilité et spécificité ; plus exactement, il s'agit d'une relation monotone inverse : quand l'une augmente, l'autre ne peut que diminuer ou rester identique.

12.6.2 Le tableau des seuils pour le FINDRISC

Le tableau 12.5 présente, pour quelques seuils du score FINDRISC, les quatre cellules et les principaux indicateurs, calculés sur le jeu mg2026.

Tableau 12.5. Performance du FINDRISC pour le diabète confirmé par HGPO, à différents seuils (n = 160 ; 29 avec diabète, 131 sans). Données synthétiques (jeu mg2026).

Seuil	VP	FP	FN	VN	Se (%)	Sp (%)	1 - Sp	VPP (%)	VPN (%)	J
≥ 8	28	94	1	37	96,6	28,2	0,718	23,0	97,4	0,248
≥ 10	28	66	1	65	96,6	49,6	0,504	29,8	98,5	0,462
≥ 11	27	44	2	87	93,1	66,4	0,336	38,0	97,8	0,595
≥ 12	22	33	7	98	75,9	74,8	0,252	40,0	93,3	0,507
≥ 14	17	16	12	115	58,6	87,8	0,122	51,5	90,6	0,464
≥ 15	13	11	16	120	44,8	91,6	0,084	54,2	88,2	0,364
≥ 16	12	7	17	124	41,4	94,7	0,053	63,2	87,9	0,360

Lisez le tableau de haut en bas. À mesure que le seuil augmente, les VP diminuent et les FN augmentent (la sensibilité baisse de 96,6% à 41,4%), tandis que les FP diminuent et les VN augmentent (la spécificité passe de 28,2% à 94,7%). La VPP augmente, la VPN diminue. L'indice de Youden atteint son maximum, 0,595, au seuil ≥ 11 (Se 93,1%, Sp 66,4%). Notez qu'à tous les seuils la prévalence reste de 18,1% ; seule change la façon dont le test répartit les patients.

12.7 La courbe ROC

12.7.1 Construction pas à pas

La courbe ROC (angl. receiver operating characteristic, dénomination héritée de la théorie des signaux radar) représente graphiquement tous les compromis possibles entre sensibilité et spécificité. Sur l'axe horizontal, on place 1 - spécificité (la proportion de faux positifs parmi les sujets sains), sur l'axe vertical la sensibilité. Chaque seuil devient un point. Le cours écrit cette relation sous la forme Se = f(1 - Sp).

La construction à la main comporte cinq étapes.

Étape 1. Ordonnez les seuils du plus strict (le plus élevé, celui qui donne le moins de résultats positifs) au plus permissif.

Étape 2. Pour chaque seuil, calculez la Se et 1 - Sp à partir du tableau 2 x 2 correspondant (tableau 12.5).

Étape 3. Ajoutez les deux points extrêmes : (0 ; 0), qui correspond à un seuil si élevé que personne n'est positif, et (1 ; 1), qui correspond à un seuil si bas que tout le monde est positif.

Étape 4. Placez les points sur le graphique et reliez-les par des segments de droite, dans l'ordre des seuils.

Étape 5. Tracez la diagonale de (0 ; 0) à (1 ; 1) : elle représente un test sans valeur, pour lequel la proportion de positifs est la même chez les malades et chez les sujets sains.

Pour le FINDRISC, avec les six seuils ≥ 16, ≥ 15, ≥ 14, ≥ 12, ≥ 10 et ≥ 8, les points (1 - Sp ; Se) sont : (0 ; 0), (0,053 ; 0,414), (0,084 ; 0,448), (0,122 ; 0,586), (0,252 ; 0,759), (0,504 ; 0,966), (0,718 ; 0,966), (1 ; 1). La figure 12.2 montre la courbe obtenue, superposée à la courbe complète construite avec tous les seuils possibles du score.

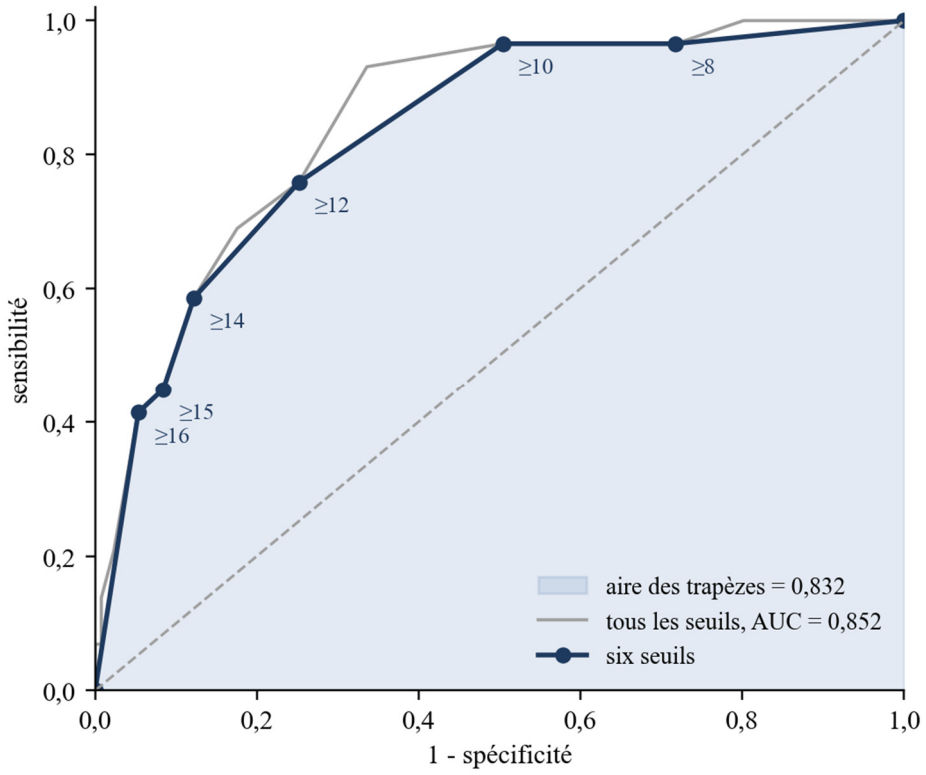


Figure 12.2. Courbe ROC du score FINDRISC pour le diabète confirmé par HGPO, construite à la main à partir de six seuils (ligne avec points), superposée à la courbe complète avec tous les seuils (ligne grise). L'aire des trapèzes

sous la courbe à six seuils est de 0,832 ; l'aire sous la courbe complète est de 0,852. Données synthétiques (jeu mg2026).

L'aire sous la courbe peut être calculée à la main par la méthode des trapèzes. Entre deux points consécutifs (x1 ; y1) et (x2 ; y2), l'aire du trapèze est (x2 - x1) x (y1 + y2) / 2. Pour les sept segments, nous obtenons 0,0111 + 0,0132 + 0,0197 + 0,0873 + 0,2172 + 0,2064 + 0,2776 = 0,832. Avec tous les seuils possibles, l'aire est de 0,852, la valeur rapportée par les logiciels. La différence vient de ce qu'en reliant par des lignes droites des points peu nombreux, nous "coupons les angles" de la courbe ; plus nous utilisons de seuils, plus l'aire des trapèzes se rapproche de l'aire réelle.

12.7.2 Que signifie l'aire sous la courbe ROC

L'aire sous la courbe ROC (angl. area under the curve, AUC), appelée aussi dans le cours coefficient c, est un indicateur global de la capacité du test à séparer les malades des sujets sains, indépendamment du seuil [137,138]. Elle a une interprétation probabiliste très concrète : l'AUC est la probabilité qu'un malade choisi au hasard ait une valeur du test plus élevée qu'un sujet sain choisi au hasard (les paires à valeurs égales comptant pour moitié).

Pour le FINDRISC dans le jeu mg2026, nous pouvons le vérifier directement. Il existe 29 x 131 = 3799 paires malade-sujet sain. Dans 3154 paires, le malade a le score le plus élevé ; dans 169 paires, les scores sont égaux. $\text{AUC} = \frac{3154+\frac{169}{2}}{3799} = \frac{3238,5}{3799} = 0,852$. Interprétation : si nous prenons au hasard un patient diabétique et un patient non diabétique, dans 85,2% des cas le patient diabétique a le score FINDRISC le plus élevé. Le calcul est identique à la statistique U du test de Mann-Whitney du chapitre 7, divisée par le produit des effectifs des deux groupes ; l'AUC est donc une mesure de la taille de l'effet pour la différence entre les distributions d'un marqueur chez les malades et chez les sujets sains.

Un test parfait a une AUC = 1 (la courbe monte verticalement jusqu'à Se = 1 pour 1 - Sp = 0, puis se poursuit horizontalement). Un test inutile a une AUC = 0,5 (la courbe se confond avec la diagonale). Une AUC inférieure à 0,5 indique habituellement une erreur de codage : le test fonctionne "à l'envers", et l'inversion du sens (les petites valeurs sont positives) donne AUC = 1 - AUC.

12.7.3 Échelles d'interprétation de l'AUC

Le cours de la discipline utilise l'échelle suivante pour le coefficient c :

Tableau 12.6. Échelle d'interprétation de l'aire sous la courbe ROC utilisée dans le cours de la discipline.

AUC	Interprétation
0,91-1	excellent
0,81-0,90	très bon
0,71-0,80	bon
0,61-0,70	satisfaisant
moins de 0,6	faible

D'autres manuels utilisent des échelles légèrement différentes. Une variante fréquente dans la littérature clinique considère 0,9-1 comme excellent, 0,8-0,9 bon, 0,7-0,8 acceptable (ou modéré), 0,6-0,7 faible et 0,5-0,6 comme un échec. Les différences entre les échelles montrent que les seuils sont des conventions, et non des lois. L'interprétation doit aussi tenir compte du contexte

clinique : une AUC de 0,75 peut être très utile pour un test de tri peu coûteux et insuffisante pour un test qui décide d'une intervention chirurgicale. Sur l'échelle du cours, le FINDRISC (0,852) et la glycémie à jeun (0,844) sont des tests "très bons" pour le diabète confirmé par HGPO dans le jeu mg2026, l'HbA1c (0,764) est un test "bon", et l'IMC (0,731) et le tour de taille (0,737) sont "bons", proches de la limite inférieure.

Les exemples tirés des travaux de nos collègues couvrent toute l'échelle : CEUS-PAT pour la cirrhose, AUC 0,98 (excellent) [2] ; score combiné pour les varices à haut risque, AUROC 0,93 [110] ; ISARIC-4C pour la mortalité, 0,875 [75] ; strain longitudinal global pour l'insuffisance cardiaque à fraction d'éjection préservée, 0,833 [90] ; ALRI chez les patients atteints de COVID-19 et de diabète, 0,752, contre seulement 0,624 pour le dNLR (satisfaisant) [105].

12.7.4 Intervalle de confiance de l'AUC

L'AUC est une estimation sur échantillon et possède une erreur standard. Hanley et McNeil ont proposé une formule qui dépend uniquement de l'AUC et du nombre de malades et de sujets sains [137] ; la méthode de DeLong, utilisée par défaut par de nombreux logiciels, estime l'erreur standard de façon non paramétrique, à partir des données elles-mêmes. Pour le FINDRISC dans mg2026, l'IC à 95% est de 0,762-0,943 avec la méthode de Hanley-McNeil et d'environ 0,782-0,923 avec la méthode de DeLong. Pour la glycémie à jeun, AUC 0,844, avec un IC à 95% de 0,752-0,937 (Hanley-McNeil). Les deux intervalles sont larges, car ils reposent sur seulement 29 malades. C'est la borne inférieure de l'intervalle qui compte pour une affirmation prudente : nous pouvons dire que le FINDRISC a au moins une performance "bonne" (0,76), pas nécessairement "très bonne".

Un intervalle de confiance qui inclut 0,5 signifie que le test ne discrimine pas significativement mieux que le hasard. Le test statistique associé (H0 : AUC = 0,5) figure dans tous les logiciels ; pour le FINDRISC, $p < 0,001$.

12.7.5 Comparaison de deux courbes ROC

Lorsque deux tests sont réalisés chez les mêmes patients, leurs AUC sont corrélées, et la comparaison doit tenir compte de cette corrélation, de même que le test t apparié tient compte de la corrélation entre mesures répétées. La méthode de DeLong fait exactement cela : elle estime la variance de la différence entre deux AUC sur les mêmes données et calcule une statistique z. Lorsque les tests sont évalués sur des groupes différents de patients, on utilise une comparaison pour échantillons indépendants.

La figure 12.3 compare le FINDRISC à la glycémie à jeun dans le jeu mg2026.

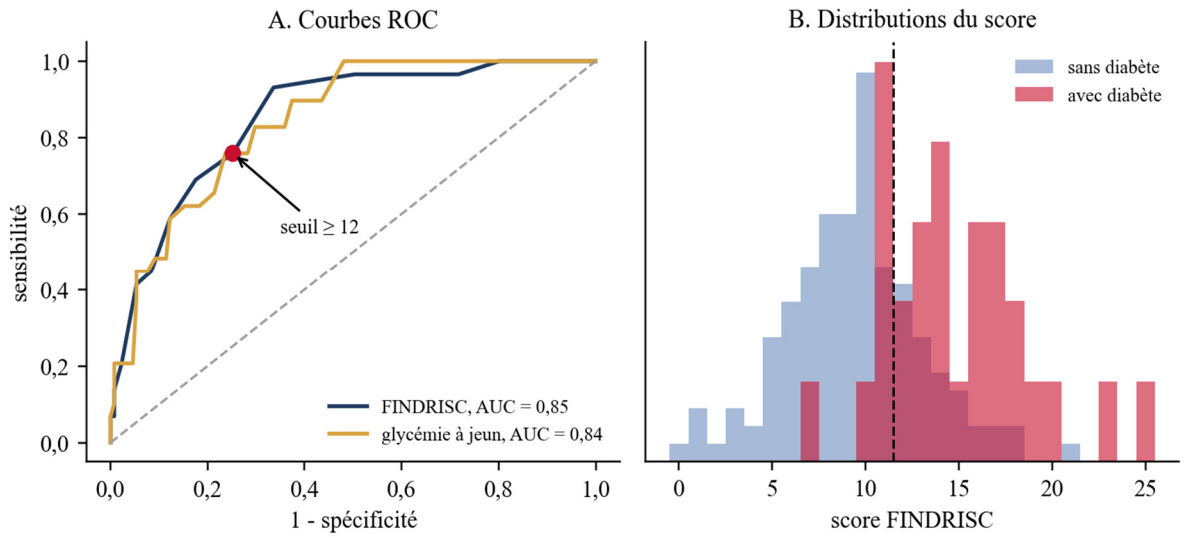


Figure 12.3. A : courbes ROC du score FINDRISC (AUC 0,85) et de la glycémie à jeun (AUC 0,84) pour le diabète confirmé par HGPO ; le point rouge marque le seuil FINDRISC ≥ 12. B : distributions du score FINDRISC chez les personnes sans diabète et avec diabète ; la ligne pointillée marque le seuil entre 11 et 12. Données synthétiques (jeu mg2026).

La différence entre les AUC est de 0,852 - 0,844 = 0,008 (IC à 95% -0,060 à 0,077), z = 0,24, p = 0,81 (méthode de DeLong, la corrélation entre les deux AUC étant d'environ 0,5). Les deux tests ont une capacité de discrimination pratiquement identique pour le diabète confirmé par HGPO. Le choix entre eux se fait selon d'autres critères : le FINDRISC ne nécessite pas de prélèvement sanguin et peut être rempli par le patient en salle d'attente, tandis que la glycémie est une mesure objective, mais exige le jeûne et un laboratoire. Observez aussi la forme des courbes : elles se croisent, si bien qu'à certains seuils un test est meilleur, à d'autres c'est l'autre. L'AUC résume la courbe entière, y compris des zones de seuils que nous n'utiliserions jamais en pratique.

Étude de cas : comparaison de plusieurs tests par l'AUROC [57,107]

Rațiu et collaborateurs (avec R. Lupușoru) ont comparé, chez 160 patients présentant une obstruction biliaire (âge moyen 69,31 ± 10,96 ans ; 53,1% d'hommes), trois méthodes par rapport au diagnostic final de malignité : l'AUROC a été de 0,88 pour la cholangio-IRM (IC à 95% 0,75-0,90), de 0,94 pour la CPRE (IC à 95% 0,85-0,99) et de 0,80 pour l'histologie (IC à 95% 0,70-0,82). Deux remarques de lecture critique. Premièrement, les intervalles de confiance sont très asymétriques autour des estimations (0,88 avec une borne supérieure de 0,90, 0,80 avec une borne supérieure de 0,82), ce qui est inhabituel avec les méthodes de calcul courantes et mérite d'être vérifié dans le texte intégral. Deuxièmement, les intervalles se chevauchent, mais le chevauchement des intervalles n'est pas un test de la différence ; pour des tests réalisés chez les mêmes patients, la comparaison correcte est la méthode de DeLong sur données appariées. Bâldea et collaborateurs (avec A. Tudor) ont comparé l'élastographie impulsionnelle (TE) à la quantification VTQ par ARFI chez 157 patients atteints d'hépatite chronique B ou C, avec la biopsie hépatique comme standard. La répartition des stades de fibrose était F0 19,7%, F1 22,3%, F2 27,4%, F3 17,8%, F4 12,7%, et la corrélation entre les deux méthodes était forte (Spearman ρ = 0,826 ; IC à 95% 0,769-0,870). Pour les seuils successifs de fibrose, les AUROC ont été de 0,876 contre 0,832 (p = 0,358), 0,826 contre 0,862 (p = 0,313), 0,907 contre 0,880 (p = 0,434) et 0,981 contre 0,974 (p = 0,423), TE contre VTQ. Aucune des différences n'a été significative. L'interprétation correcte n'est pas "les deux méthodes sont équivalentes", mais "l'étude n'a pas trouvé de différences ; les différences observées, de l'ordre de 0,03-0,04, sont compatibles avec le hasard". Pour démontrer

l'équivalence, il aurait fallu un schéma d'équivalence, avec une marge fixée à l'avance, ou les intervalles de confiance des différences.

12.7.6 Choix du seuil : indice de Youden et seuils cliniques

Le seuil ayant l'indice de Youden maximal est le point de la courbe ROC le plus éloigné verticalement de la diagonale. Pour le FINDRISC dans mg2026, le seuil de Youden est ≥ 11, avec une Se de 93,1%, une Sp de 66,4% et J = 0,595. Pour la glycémie à jeun, le seuil de Youden est ≥ 110 mg/dL, avec une Se de 82,8%, une Sp de 70,2% et J = 0,530. L'indice de Youden suppose toutefois qu'un faux négatif et un faux positif ont la même importance et que la prévalence ne compte pas, ce qui correspond rarement à la réalité clinique.

En pratique, le seuil se choisit selon l'objectif du test. Pour un test de tri, pour lequel un résultat négatif met fin aux explorations, l'important est de ne pas manquer de malades : on choisit un seuil d'exclusion (angl. rule-out) avec une sensibilité élevée et une VPN élevée. Dans mg2026, FINDRISC ≥ 10 a une Se de 96,6% et une VPN de 98,5% : un patient dont le score est inférieur à 10 a une probabilité de diabète d'environ 1,5% et peut être rassuré. Pour un test qui déclenche un traitement invasif ou coûteux, l'important est de ne pas traiter des sujets sains : on choisit un seuil de confirmation (angl. rule-in) avec une spécificité élevée et une VPP élevée. FINDRISC ≥ 16 a une Sp de 94,7% et une VPP de 63,2%. Entre les deux seuils subsiste une zone grise, dans laquelle la décision exige des informations supplémentaires.

L'exemple de la glycémie montre encore une chose importante. Le seuil clinique de diagnostic du diabète pour la glycémie à jeun, 126 mg/dL [80], a dans le jeu mg2026 une Se de seulement 37,9% et une Sp de 94,7% (VPP 61,1%, VPN 87,3%) par rapport à l'HGPO. Ce seuil est choisi pour confirmer le diagnostic (spécificité élevée), et non pour le dépistage ; utilisé seul, il manquerait près des deux tiers des diabétiques identifiés par l'HGPO. Le seuil "statistiquement optimal" (110 mg/dL) et le seuil "clinique" (126 mg/dL) répondent à des questions différentes.

Le même raisonnement se retrouve dans les études de nos collègues. Moga et collaborateurs ont rapporté pour l'élastographie time-harmonic un seuil d'exclusion de la cirrhose (< 1,61 m/s : Se 90,7%, VPN 93,9%) et un seuil de confirmation (> 1,83 m/s : Sp 96,7%, VPP 90,3%) [3]. Barbulescu et collaborateurs ont construit des règles d'exclusion et de confirmation des varices à haut risque, avec une zone grise résolue par l'albumine [123]. Popa et collaborateurs (avec A. Tudor) ont rapporté pour la GFAP, chez 85 patients admis en alerte code AVC au service d'accueil des urgences, une exactitude de 94,2% pour différencier l'AVC hémorragique de l'AVC ischémique, au seuil de 77,15 pg/mL ; la GFAP était de $142{,}91 \pm 102{,}19$ pg/mL chez les patients ayant un AVC aigu contre $37{,}76 \pm 19{,}92$ pg/mL chez ceux sans AVC ($p < 0{,}001$) [61]. L'écart-type proche de la moyenne suggère une distribution asymétrique, pour laquelle la médiane et le test de Mann-Whitney auraient été plus appropriés, et le seuil à quatre chiffres significatifs, dérivé sur 69 patients ayant eu un AVC, nécessite une validation.

Étude de cas : seuils ROC pour des maladies à plusieurs degrés [58,65,90]

La stéatose hépatique comporte quatre grades (S0-S3), et la fibrose cinq stades (F0-F4). Pour ces maladies ordinales, l'analyse ROC se fait séparément pour chaque seuil de sévérité : S0 contre S1-S3, S0-S1 contre S2-S3, et ainsi de suite. Bende et collaborateurs (avec R. Lupușoru) ont évalué le paramètre UGAP par rapport au CAP chez 176 sujets classés : S0 27,2%, S1 31,6%, S2 7,3%, S3 33,9%. Les valeurs moyennes de l'UGAP ont augmenté de façon monotone avec le grade ($198{,}3 \pm \frac{25{,}7 d_B}{m}$ en S0 et $270{,}8 \pm \frac{31{,}62 d_B}{m}$ en S3 ; $p < 0{,}001$), la corrélation avec le CAP a été de r = 0,73, et les seuils ont été de 192,5 dB/m pour S1 (AUC 0,83), de 231 dB/m pour S2 (AUC 0,90) et de 248 dB/m pour S3 (AUC 0,91). Nica et collaborateurs (avec R. Lupușoru) ont évalué la fraction de graisse échographique (USFF) chez 169 patients, avec le CAP comme référence (seuils CAP 248, 268 et 280 dB/m ; 52,6% avec

une stéatose au moins modérée). L'USFF était corrélée au CAP (r = 0,81), et les seuils > 5%, > 7,5% et > 10% ont eu des AUROC de 0,90, 0,92 et 0,91 pour la stéatose au moins légère, modérée et sévère, respectivement. Un détail intéressant pour la lecture critique : le résumé qualifie de "modérée" la corrélation r = 0,83 entre le paramètre TAI et le CAP, alors que sur l'échelle de Colton du cours elle serait "très bonne à forte". Stoichescu-Hogea et collaborateurs (avec A. Tudor) ont comparé le strain longitudinal global (GLS) dans trois groupes (62 patients atteints d'insuffisance cardiaque à fraction d'éjection préservée, 46 hypertendus, 40 témoins) : -17,6 ± 2,3% dans le premier groupe, -18,4 ± 3,0% chez les hypertendus et -20,2 ± 1,4% chez les témoins, et le seuil optimal de -19,35% a eu une AUC de 0,833. Ici, "positif" signifie une valeur moins négative, et le logiciel doit être correctement paramétré pour le sens du test (sinon l'AUC serait de 0,167). Dans les trois études, le standard de référence (le CAP ou le diagnostic clinique) est imparfait, et les seuils sont dérivés sur le même échantillon que celui sur lequel on estime la performance ; ils nécessitent donc une validation externe.

12.8 Accord entre évaluateurs

Un test diagnostique utile doit être non seulement valide (mesurer ce qu'il doit mesurer), mais aussi fiable : donner le même résultat lorsqu'il est répété ou interprété par un autre examinateur. Pour des résultats catégoriels (par exemple "hypertrophie ventriculaire gauche présente ou absente" à l'ECG), l'accord se mesure avec le coefficient kappa de Cohen. Pour des résultats numériques (rigidité hépatique en kPa, pression artérielle), on utilise le coefficient de corrélation intraclasse et la méthode de Bland-Altman [102].

12.8.1 Le coefficient kappa de Cohen

La proportion d'accord observé (Po) est insuffisante, car deux évaluateurs seraient souvent d'accord aussi par hasard, surtout lorsque l'une des catégories est très fréquente. Le kappa corrige l'accord observé par l'accord attendu sous l'effet du hasard (Pe) :

$$\kappa = \frac{\mathrm{Po} - \mathrm{Pe}}{1 - \mathrm{Pe}}$$

L'accord attendu se calcule à partir des totaux marginaux, comme les effectifs attendus du test du khi-deux : pour chaque catégorie, on fait le produit des proportions avec lesquelles les deux évaluateurs l'ont utilisée, puis la somme. Le kappa vaut 1 en cas d'accord parfait, 0 lorsque l'accord ne dépasse pas le hasard, et il est négatif lorsque l'accord est inférieur à celui dû au hasard. Pour l'interprétation, on utilise fréquemment l'échelle de Landis et Koch [103].

Tableau 12.7. Échelle de Landis et Koch pour l'interprétation du coefficient kappa.

κ	Accord
moins de 0	médiocre (inférieur au hasard)
0-0,20	léger
0,21-0,40	acceptable (modeste)
0,41-0,60	modéré
0,61-0,80	substantiel
0,81-1,00	presque parfait

L'exemple complet se trouve dans le problème résolu 12.8. Le kappa a quelques limites : il dépend de la prévalence des catégories (pour des prévalences extrêmes, le kappa peut être faible même avec un accord observé élevé) et il ne distingue pas les petits désaccords des grands dans le

cas des catégories ordonnées. Pour les échelles ordinales (par exemple les grades de stéatose), on utilise le kappa pondéré.

12.8.2 Le coefficient de corrélation intraclasse

Pour des mesures numériques, le coefficient de corrélation de Pearson n'est pas adapté à l'évaluation de l'accord : deux appareils qui diffèrent systématiquement de 5 kPa peuvent avoir r = 1 sans être en accord. Le coefficient de corrélation intraclasse (ICC) mesure la proportion de la variabilité totale due aux différences réelles entre patients, par rapport à celle due aux différences entre mesures ou entre évaluateurs. Des valeurs supérieures à 0,90 indiquent une fiabilité excellente, entre 0,75 et 0,90 bonne, entre 0,50 et 0,75 modérée.

Moga et collaborateurs ont inclus dans l'étude de l'élastographie time-harmonic une sous-étude de reproductibilité portant sur 27 patients : l'accord global entre examinateurs a été de 0,94 (IC à 95% 0,89-0,97), et les ICC pour différentes comparaisons ont été de 0,92 (IC à 95% 0,82-0,96), 0,94 (IC à 95% 0,87-0,97) et 0,97 (IC à 95% 0,95-0,99) [3]. Un test ayant une telle reproductibilité peut être utilisé par différents examinateurs sans que les différences entre eux modifient la classification des patients. L'ICC est aussi utilisé pour la fiabilité test-retest des questionnaires, par exemple des valeurs d'ICC comprises entre 0,82 et 0,91 à un intervalle de 14 jours dans une étude de validation de nos collègues [73].

12.9 Évaluation des classificateurs d'intelligence artificielle

Les algorithmes d'intelligence artificielle qui classent des images médicales (radiographies, rétinographies, images dermatoscopiques) ou des signaux (ECG) sont, du point de vue statistique, des tests diagnostiques. Ils s'évaluent avec les mêmes outils, même si la littérature informatique emploie d'autres termes.

Le tableau 2 x 2 s'appelle ici matrice de confusion (angl. confusion matrix). La sensibilité s'appelle rappel (angl. recall, ou taux de vrais positifs), la VPP s'appelle précision (angl. precision), et la spécificité reste la spécificité (plus rarement, taux de vrais négatifs). Un indicateur propre au domaine est le score F1, moyenne harmonique de la précision et de la sensibilité :

F1 = 2 x précision x sensibilité / (précision + sensibilité)

Le F1 n'est élevé que si ses deux composantes sont élevées, et il est utilisé surtout lorsque la classe positive est rare. Contrairement à l'AUC, le F1 dépend de la prévalence (par l'intermédiaire de la précision) et du seuil choisi.

Le problème classique de l'évaluation des classificateurs est le paradoxe de l'exactitude en cas de classes déséquilibrées. Si seulement 8% des rétinographies d'un programme de dépistage présentent une rétinopathie diabétique, un algorithme qui déclare toutes les images normales a une exactitude de 92%, mais une sensibilité de 0. C'est pourquoi un article qui ne rapporte que "l'exactitude du modèle" ne peut pas être évalué. D'autres pièges fréquents sont l'évaluation sur le même jeu de données que celui sur lequel le modèle a été entraîné (surapprentissage, angl. overfitting), l'absence de validation externe sur des données d'un autre centre et le choix du seuil en fonction des résultats. Le problème résolu 12.10 calcule tous ces indicateurs pour un exemple concret. Pour les modèles de prédiction fondés sur la régression ou l'apprentissage automatique, il existe une recommandation de publication dédiée, TRIPOD+AI [139].

12.10 Publication des études d'exactitude diagnostique : STARD 2015 et STARD-AI

La recommandation STARD 2015 (angl. Standards for Reporting of Diagnostic Accuracy Studies) comprend 30 items qu'un article d'exactitude diagnostique devrait rapporter [140].

Certains comptent davantage pour l'évaluation critique : l'identification de l'étude comme étude d'exactitude diagnostique dans le titre ou le résumé ; la description de la population, des critères d'éligibilité et du mode de recrutement (consécutif ou non) ; la description du test à évaluer et du standard de référence, avec suffisamment de détails pour qu'ils puissent être reproduits ; l'indication que les interprètes étaient ou non en insu du résultat de l'autre test ; l'indication que les seuils ont été fixés à l'avance ou dérivés des données ; le diagramme de flux des participants (combien étaient éligibles, combien ont reçu chaque test, combien ont été exclus et pourquoi) ; le tableau 2 x 2 complet ; les estimations d'exactitude avec leurs intervalles de confiance ; les résultats ininterprétables et la façon dont ils ont été traités.

STARD-AI étend STARD aux études dans lesquelles le test à évaluer est un algorithme d'intelligence artificielle [141]. Elle ajoute des exigences concernant la source et les caractéristiques des données d'entraînement, de validation et de test, la séparation stricte de ces jeux, la version de l'algorithme, le prétraitement des données, le mode de choix du seuil, la performance dans les sous-groupes (sexe, âge, centre, type d'appareil) et l'analyse des erreurs.

À la lecture d'un article, utilisez la liste comme un filtre : si le tableau 2 x 2 ne peut pas être reconstitué, si vous ne savez pas comment les patients ont été recrutés ou si le seuil a été choisi sur les mêmes données que celles sur lesquelles la performance est rapportée, les résultats doivent être lus avec prudence, aussi élevées que soient l'AUC et la sensibilité.

12.11 Problèmes résolus

Problème résolu 12.1. Tous les indicateurs à partir d'un tableau complet

Un test a été évalué chez 440 patients par rapport au standard de référence. Parmi les 180 malades, 150 ont eu un test positif ; parmi les 260 non-malades, 4 ont eu un test positif. Calculez tous les indicateurs de performance.

Solution. Le tableau, avec les notations du cours : a = 150, b = 180 - 150 = 30, c = 4, d = 260 - 4 = 256.

	T+	**T-**	**Total**
B+	150	30	180
B-	4	256	260
Total	154	286	440

Prévalence = 180/440 = 0,409 (40,9%).

Se $= \frac{150}{180} = 0{,}833(83{,}3\%)$. Sp $= \frac{256}{260} = 0{,}985(98{,}5\%)$.

VPP $= \frac{150}{154} = 0{,}974(97{,}4\%)$. VPN $= \frac{256}{286} = 0{,}895(89{,}5\%)$.

Ex $= \frac{150+256}{440} = \frac{406}{440} = 0{,}923(92{,}3\%)$. TE $= \frac{30+4}{440} = \frac{34}{440} = 0{,}077(7{,}7\%)$; vérification : 0,923 + 0,077 = 1.

$J = 0{,}833 + 0{,}985 - 1 = 0{,}818.$

$\mathrm{RV}^{+} = \frac{0{,}833}{1-0{,}985} = \frac{0{,}833}{0{,}0154} = 54{,}2$. $\mathrm{RV}^{-} = \frac{0{,}167}{0{,}985} = 0{,}17.$

DOR $= \frac{150\times256}{30\times4} = 38.400/120 = 320.$

Interprétation. Le test a une spécificité très élevée et un RV+ supérieur à 50 : un résultat positif confirme pratiquement la maladie (SpPin). La sensibilité de 83,3% signifie cependant que 1 malade sur 6 a un test négatif ; un résultat négatif n'exclut pas la maladie, et le RV- de 0,17 la

rend seulement modérément moins probable. Les très bonnes valeurs prédictives (97,4% et 89,5%) sont valables pour la prévalence de 40,9% ; dans une population de dépistage, de prévalence bien plus faible, la VPP diminuerait et la VPN augmenterait.

Problème résolu 12.2. Reconstitution du tableau à partir de la VPP, de la VPN et de deux cellules

Pour un test, on connaît VPP $= \frac{9}{12}$, VPN $= \frac{7}{8}$, le nombre de vrais positifs VP = 90 et le nombre de faux négatifs FN = 10. Reconstituez le tableau et calculez la Se, la Sp et la prévalence.

Solution. Étape 1 : a = VP = 90, b = FN = 10.

Étape 2 : de VPP $= \frac{a}{a+c} = \frac{9}{12}$, on déduit $\frac{90}{90+c} = \frac{9}{12}$, donc 9 x (90 + c) = 12 x 90, c'est-à-dire 810 + 9c = 1080, d'où c = 30.

Étape 3 : de VPN $= \frac{d}{b+d} = \frac{7}{8}$, on déduit $\frac{d}{10+d} = \frac{7}{8}$, donc 8d = 70 + 7d, d'où d = 70.

	T+	**T-**	**Total**
B+	90	10	100
B-	30	70	100
Total	120	80	200

Étape 4 (vérification) : VPP $= \frac{90}{120} = \frac{9}{12}$ et VPN $= \frac{70}{80} = \frac{7}{8}$.

Indicateurs : Se $= \frac{90}{100} = 0{,}90$; Sp $= \frac{70}{100} = 0{,}70$; prévalence = 100/200 = 0,50 ; Ex $= \frac{160}{200} =$ 0,80 ; TE $= \frac{10+30}{200} = 0{,}20$; $RV^+ = \frac{0{,}90}{0{,}30} = 3{,}0$; $RV^- = \frac{0{,}10}{0{,}70} = 0{,}14$; J = 0,60.

Interprétation. Le test est meilleur pour exclure que pour confirmer. La prévalence de 50% est typique d'un schéma cas-témoins à groupes égaux, ce qui signifie que la VPP de 75% n'est pas transposable à une population réelle de faible prévalence.

Problème résolu 12.3. Reconstitution du tableau à partir de la Se, de la VPP et de deux cellules

On connaît Se $= \frac{2}{3}$, VPP $= \frac{6}{8}$, FN = 30 et VN = 80. Reconstituez le tableau et calculez la Sp et la VPN.

Solution. b = 30, d = 80.

De Se $= \frac{a}{a+b} = \frac{2}{3} : \frac{a}{a+30} = \frac{2}{3}$, donc 3a = 2a + 60, d'où a = 60.

De VPP $= \frac{a}{a+c} = \frac{6}{8} : \frac{60}{60+c} = \frac{6}{8}$, donc 6 x (60 + c) = 480, c'est-à-dire 360 + 6c = 480, d'où c = 20.

	T+	**T-**	**Total**
B+	60	30	90
B-	20	80	100
Total	80	110	190

$\text{Sp} = \frac{80}{100} = 0{,}80$. $\text{VPN} = \frac{80}{110} = 0{,}727$. Prévalence = 90/190 = 0,474 ; $\text{Ex} = \frac{140}{190} = 0{,}737$; $\text{TE} = \frac{50}{190} = 0{,}263$; $\text{RV}^+ = \frac{0{,}667}{0{,}20} = 3{,}33$; $\text{RV}^- = \frac{0{,}333}{0{,}80} = 0{,}42$; $\text{DOR} = \frac{60\times80}{30\times20} = 8$.

Remarque. La sensibilité de 2/3 vaut 0,667, et non 0,66. L'arrondi prématuré de la fraction (à 0,66) conduit à a = 58,2, c'est-à-dire à une cellule contenant un nombre non entier, signe certain d'erreur. Travaillez avec les fractions jusqu'à la fin.

Problème résolu 12.4. Construction du tableau à partir de la prévalence, de la Se et de la Sp

Un test rapide a une Se = 85% et une Sp = 90%. Construisez le tableau pour 500 patients avec une prévalence de la maladie de 20% et calculez la VPP et la VPN. Refaites le calcul pour 2000 personnes avec une prévalence de 5%.

Solution. Prévalence de 20% sur 500 : a + b = 100 malades, c + d = 400 non-malades. a = 0,85 x 100 = 85, b = 15. d = 0,90 x 400 = 360, c = 40.

	T+	T-	Total
B+	85	15	100
B-	40	360	400
Total	125	375	500

$\text{VPP} = \frac{85}{125} = 68{,}0\%$; $\text{VPN} = \frac{360}{375} = 96{,}0\%$; $\text{Ex} = \frac{445}{500} = 89{,}0\%$; $\text{TE} = \frac{55}{500} = 11{,}0\%$.

Pour une prévalence de 5%, sur 2000 : 100 malades (a = 85, b = 15) et 1900 non-malades (d = 0,90 x 1900 = 1710, c = 190). $\text{VPP} = \frac{85}{85+190} = \frac{85}{275} = 30{,}9\%$; $\text{VPN} = \frac{1710}{1725} = 99{,}1\%$.

Interprétation. La sensibilité et la spécificité n'ont pas changé. Lorsque la prévalence baisse de 20% à 5%, la VPP diminue de 68,0% à 30,9%, et la VPN augmente de 96,0% à 99,1%. L'exactitude augmente légèrement (de 89,0% à (85 + 1710)/2000 = 89,8%), car la spécificité, plus élevée que la sensibilité, pèse davantage lorsque les sujets sains prédominent ; encore une raison de ne pas se fier à l'exactitude.

Problème résolu 12.5. Reconstitution à partir de la Se, de la VPN, des FN et des FP

On connaît Se = 0,80, VPN = 0,95, FN = 20 et FP = 60. Reconstituez le tableau.

Solution. b = 20, c = 60. De $\text{Se} = \frac{a}{a+20} = 0{,}80$: a = 0,80a + 16, donc 0,20a = 16, a = 80. De $\text{VPN} = \frac{d}{20+d} = 0{,}95$: d = 19 + 0,95d, donc 0,05d = 19, d = 380.

	T+	T-	Total
B+	80	20	100
B-	60	380	440
Total	140	400	540

$\text{Sp} = \frac{380}{440} = 86{,}4\%$; $\text{VPP} = \frac{80}{140} = 57{,}1\%$; prévalence = 100/540 = 18,5% ; $\text{Ex} = \frac{460}{540} = 85{,}2\%$; $\text{TE} = \frac{80}{540} = 14{,}8\%$; $\text{RV}^+ = \frac{0{,}80}{0{,}136} = 5{,}87$; $\text{RV}^- = \frac{0{,}20}{0{,}864} = 0{,}23$.

Problème résolu 12.6. Évaluation complète du FINDRISC ≥ 12 sur le jeu mg2026

En utilisant le tableau 12.3, formulez la question, calculez les indicateurs avec leurs intervalles de confiance et interprétez le résultat pour un médecin généraliste qui travaille avec une population dans laquelle la prévalence du diabète non diagnostiqué est de 5%.

Solution. (a) Question de recherche : dans quelle mesure le score FINDRISC, au seuil ≥ 12, identifie-t-il les adultes atteints d'un diabète non diagnostiqué, par rapport à l'HGPO ? (b) Schéma

: étude transversale d'exactitude diagnostique, chez 160 adultes à risque cardiométabolique, chez qui les deux tests ont été réalisés. (c) Indicateurs : Se 75,9% (IC à 95% 57,9-87,8) ; Sp 74,8% (IC à 95% 66,7-81,5) ; VPP 40,0% (IC à 95% 28,1-53,2) ; VPN 93,3% (IC à 95% 86,9-96,7) ; exactitude 75,0% ; RV+ 3,01 (IC à 95% 2,10-4,31) ; RV- 0,32 (IC à 95% 0,17-0,62) ; DOR 9,33. (d) Pour une prévalence de 5% : cote pré-test 0,0526 ; après un score positif, 0,0526 x 3,01 = 0,158, donc VPP 13,7% ; après un score négatif, 0,0526 x 0,32 = 0,017, donc probabilité post-test 1,7% (VPN 98,3%), exactement les valeurs du tableau 12.4.

Interprétation clinique. Dans une population où la prévalence est de 5%, sur 100 patients ayant un FINDRISC ≥ 12, environ 14 ont un diabète ; les 86 autres feront un test de laboratoire et recevront un résultat rassurant. Sur 100 patients ayant un score inférieur à 12, environ 98 n'ont pas de diabète. Le score est utile pour sélectionner les patients qui méritent d'être testés, et non pour poser le diagnostic. L'intervalle de confiance large de la sensibilité (57,9-87,8) montre que, sur un échantillon de 29 malades, nous ne savons pas précisément combien de diabétiques le score manque.

Problème résolu 12.7. Courbe ROC de la glycémie à jeun, construite à la main

Sur le jeu mg2026, pour la glycémie à jeun par rapport au diabète confirmé par HGPO, on a obtenu aux seuils ≥ 126, ≥ 120, ≥ 115, ≥ 110, ≥ 105 et ≥ 100 mg/dL : VP 11, 14, 19, 24, 26, 29 (sur 29 malades) et FP 7, 15, 28, 39, 54, 68 (sur 131 non-malades). Construisez les points de la courbe ROC, calculez l'aire par la méthode des trapèzes et comparez-la à l'AUC rapportée par le logiciel (0,844).

Solution. Nous calculons $\text{Se} = \frac{\text{VP}}{29}$ et $1 - \text{Sp} = \frac{\text{FP}}{131}$.

Seuil (mg/dL)	Se	1 - Sp	Sp
≥ 126	0,379	0,053	0,947
≥ 120	0,483	0,115	0,885
≥ 115	0,655	0,214	0,786
≥ 110	0,828	0,298	0,702
≥ 105	0,897	0,412	0,588
≥ 100	1,000	0,519	0,481

Nous ajoutons (0 ; 0) et (1 ; 1) et calculons les trapèzes (largeur x hauteur moyenne) : 0,053 x 0,190 = 0,0101 ; 0,061 x 0,431 = 0,0263 ; 0,099 x 0,569 = 0,0565 ; 0,084 x 0,741 = 0,0623 ; 0,115 x 0,862 = 0,0987 ; 0,107 x 0,948 = 0,1013 ; 0,481 x 1,000 = 0,4809. Somme : 0,836.

Interprétation. L'aire approchée, 0,836, est légèrement inférieure à la valeur exacte de 0,844, pour la même raison que pour le FINDRISC : des points peu nombreux "coupent les angles" de la courbe. L'indice de Youden maximal parmi les seuils du tableau se situe à ≥ 110 mg/dL (0,828 + 0,702 - 1 = 0,530). Au seuil diagnostique ≥ 126 mg/dL, la sensibilité n'est que de 37,9% : le seuil clinique est un seuil de confirmation, et non de dépistage.

Problème résolu 12.8. Le coefficient kappa pour deux médecins généralistes

Deux médecins généralistes ont interprété de façon indépendante 100 électrocardiogrammes pour les critères d'hypertrophie ventriculaire gauche. Tous deux ont considéré les critères présents sur 20 ECG et absents sur 65. Le médecin A a répondu "présent" et le médecin B "absent" pour 10 ECG ; l'inverse s'est produit pour 5. Calculez le kappa.

Solution. Le tableau d'accord :

	B : présent	B : absent	Total
A : présent	20	10	30
A : absent	5	65	70
Total	25	75	100

Accord observé : $\text{Po} = \frac{20+65}{100} = 0{,}85$.

Accord attendu sous l'effet du hasard : A répond "présent" dans 30% des cas, B dans 25% ; la probabilité que tous deux répondent "présent" par hasard est de 0,30 x 0,25 = 0,075. La probabilité que tous deux répondent "absent" par hasard est de 0,70 x 0,75 = 0,525. Pe = 0,075 + 0,525 = 0,60.

$\kappa = \frac{0{,}85-0{,}60}{1-0{,}60} = \frac{0{,}25}{0{,}40} = 0{,}625$.

Interprétation. Sur l'échelle de Landis et Koch, l'accord est substantiel (0,61-0,80). Notez que l'accord brut de 85% paraît excellent, mais que 60 points de pourcentage de cet accord seraient apparus aussi par hasard, étant donné la fréquence élevée des ECG normaux. Sur les 15 désaccords, 10 vont dans le même sens (A voit plus souvent une hypertrophie), ce qui suggère une différence systématique de critère entre les deux médecins, qui mérite d'être discutée.

Problème résolu 12.9. Reconstitution du tableau à partir d'un article : CEUS-PAT pour la cirrhose

Lupuşoru et collaborateurs ont rapporté, chez 37 patients atteints de cirrhose et 27 volontaires sains, pour le CEUS-PAT au seuil > 0,7 : Se 89,19%, Sp 100%, VPP 100%, VPN 87,1% [2]. Reconstituez le tableau et calculez le RV+, le RV-, l'exactitude et la VPP pour une prévalence de la cirrhose de 10% parmi les patients atteints d'hépatopathie chronique.

Solution. a + b = 37, c + d = 27. De $\text{Se} = \frac{a}{37} = 0{,}8919$, a = 33, donc b = 4. De $\text{Sp} = \frac{d}{27} = 1$, d = 27, c = 0. Vérification : $\text{VPP} = \frac{33}{33} = 100\%$; $\text{VPN} = \frac{27}{4+27} = \frac{27}{31} = 87{,}1\%$.

	CEUS-PAT > 0,7	CEUS-PAT ≤ 0,7	Total
Cirrhose	33	4	37
Sujets sains	0	27	27
Total	33	31	64

$\text{Ex} = \frac{60}{64} = 93{,}8\%$. $\text{RV}^- = \frac{0{,}108}{1} = 0{,}11$. Le RV+ ne peut pas être calculé (division par 0, puisque 1 - Sp = 0) ; formellement, il est infini. Toujours formellement, la VPP serait de 100% quelle que soit la prévalence.

Interprétation. Une spécificité de 100% observée sur 27 volontaires sains ne signifie pas que le test ne donne jamais de faux positifs. L'intervalle de confiance de Wilson pour 27/27 a une borne inférieure d'environ 87,5%, de sorte que la spécificité réelle pourrait être nettement plus faible, surtout chez des patients atteints d'hépatopathies non cirrhotiques, absents de l'échantillon. Ici, la VPP de "100% quelle que soit la prévalence" est une illusion produite par un petit échantillon et par un spectre favorable. La recommandation correcte face à un tel résultat est la validation chez des patients consécutifs atteints de maladie hépatique chronique.

Problème résolu 12.10. Évaluation d'un classificateur d'intelligence artificielle

Un algorithme de détection de la rétinopathie diabétique a été testé sur 1000 rétinographies, dont 80 avec rétinopathie (confirmée par deux ophtalmologistes). L'algorithme a correctement classé 72 des images avec rétinopathie et 874 des images sans rétinopathie. Calculez la matrice de confusion, la sensibilité (rappel), la spécificité, la précision (VPP), le score F1 et l'exactitude. Comparez l'exactitude à celle d'un "algorithme" qui déclare toutes les images normales.

Solution. VP = 72, FN = 8, VN = 874, FP = 920 - 874 = 46.

	Algorithme : rétinopathie	**Algorithme : normal**	**Total**
Rétinopathie	72	8	80
Sans rétinopathie	46	874	920
Total	118	882	1000

Sensibilité (rappel) = 72/80 = 90,0%. Spécificité = 874/920 = 95,0%. Précision = 72/118 = 61,0%. $F_1 = \frac{2\times0{,}610\times0{,}900}{0{,}610+0{,}900} = \frac{1{,}098}{1{,}510} = 0{,}727$. Exactitude = (72 + 874)/1000 = 94,6%.

Un "algorithme" qui déclare tout normal aurait VP = 0, VN = 920, une exactitude de 920/1000 = 92,0%, une sensibilité de 0 et un F1 de 0.

Interprétation. L'exactitude de 94,6% ne dépasse que de 2,6 points de pourcentage celle d'un classificateur inutile ; ce n'est pas un indicateur suffisant en cas de classes déséquilibrées. La sensibilité de 90% et la spécificité de 95% décrivent mieux l'algorithme. La précision de 61,0% indique que près de 4 patients sur 10 adressés à l'ophtalmologiste seront des faux positifs, un coût qu'il faut peser lors de la planification du programme de dépistage.

12.12 Exercices proposés

1. Un test a été appliqué à 200 patients : 45 malades avec un test positif, 5 malades avec un test négatif, 30 non-malades avec un test positif et 120 non-malades avec un test négatif. Construisez le tableau et calculez la prévalence, la Se, la Sp, la VPP, la VPN, l'Ex, le TE, J, le RV+, le RV- et le DOR.
2. Pour un test, on connaît VPP $= \frac{4}{5}$, VPN $= \frac{9}{10}$, VP = 80 et FN = 20. Reconstituez le tableau et calculez la Se, la Sp et N.
3. Pour un test, on connaît Sp $= \frac{3}{4}$, VPN $= \frac{15}{16}$, VN = 150 et VP = 40. Reconstituez le tableau et calculez la Se et la VPP.
4. Un test a une Se = 75% et une Sp = 95%. Construisez le tableau pour 1000 patients avec une prévalence de 10% et calculez la VPP et la VPN.
5. En utilisant les données de l'étude de Moleriu et collaborateurs (problème résolu 11.3) : 18 lésions de haut grade et 46 sans, le HPV avec 17 résultats positifs parmi les malades et 27 négatifs parmi les autres, p16/Ki-67 avec 12 et 39, construisez les deux tableaux et calculez pour chaque test l'exactitude et l'indice de Youden [131]. Quel test est le plus approprié pour exclure et lequel pour confirmer ?
6. Un marqueur présente, à cinq seuils, les points (1 - Sp ; Se) : (0,05 ; 0,40), (0,10 ; 0,60), (0,20 ; 0,80), (0,40 ; 0,90), (0,70 ; 0,98). Calculez l'aire sous la courbe ROC par la méthode des trapèzes (ajoutez les points (0 ; 0) et (1 ; 1)) et situez-la sur l'échelle du cours.
7. Deux radiologues ont lu 200 radiographies pulmonaires à la recherche d'une pneumonie. Tous deux ont répondu "pneumonie" pour 30 et "pas de pneumonie" pour 145 ; le radiologue A a répondu "pneumonie" et B "pas de pneumonie" pour 15, et l'inverse pour 10. Calculez le kappa et interprétez-le sur l'échelle de Landis et Koch.
8. Albai et collaborateurs ont rapporté pour l'ISARIC-4C, chez 159 patients avec une mortalité de 25,15%, une Se de 87,80% et une Sp de 73,72% [75]. (a) Reconstituez approximativement le tableau 2 x 2 (nombre de décès, VP, FN, FP, VN, arrondis à l'entier). (b) Calculez la VPP à partir du tableau reconstitué. (c) Calculez l'intervalle de

confiance de Wilson approximatif de la sensibilité et comparez-le à l'intervalle rapporté (87,66-87,93).

9. Un algorithme de classification des lésions cutanées a été testé sur 500 images, dont 50 mélanomes. L'algorithme a correctement identifié 40 mélanomes et 360 lésions bénignes. Calculez la sensibilité, la spécificité, la précision, le F1 et l'exactitude, et commentez.
10. Danila et collaborateurs ont rapporté une AUC de 0,624 pour le dNLR et de 0,752 pour l'ALRI chez les patients atteints de COVID-19 et de diabète [105]. Situez ces deux valeurs sur l'échelle du cours et expliquez pourquoi une AUC significativement différente de 0,5 ($p < 0,001$) ne garantit pas l'utilité clinique.
11. Sipos et collaborateurs ont rapporté pour l'IL-6 dans le sepsis néonatal une Se de 80,1% et une Sp de 85,7% [126]. Calculez la VPP et la VPN pour une prévalence du sepsis de 10% parmi les nouveau-nés testés.
12. Dans le jeu mg2026, la glycémie à jeun ≥ 126 mg/dL a une Se de 37,9% et une Sp de 94,7% par rapport à l'HGPO. Calculez le RV+ et le RV- et expliquez pourquoi ce seuil convient à la confirmation, mais pas au dépistage.

Pièges fréquents

- Calculer le taux d'erreur avec la formule (b + d)/N. Correct : $\text{TE} = \frac{b+c}{N} = \frac{\text{FN+FP}}{N} = 1 - \text{Ex}$.
- Inverser les lignes et les colonnes : dans certains logiciels, la maladie est en colonnes, et b et c échangent leur place. Vérifiez ce que contient chaque cellule.
- Rapporter la VPP et la VPN d'une étude cas-témoins comme des valeurs cliniques ; la prévalence de l'échantillon est fixée par le schéma d'étude.
- Arrondir les fractions (2/3 = 0,66) avant de reconstituer le tableau.
- Considérer l'exactitude comme un indicateur suffisant, surtout lorsque la maladie est rare.
- Interpréter le chevauchement des intervalles de confiance des AUC comme une preuve d'équivalence, ou l'absence de chevauchement comme une preuve de différence ; la comparaison se fait avec la méthode de DeLong.
- Choisir le seuil sur le même échantillon que celui sur lequel la performance est rapportée, sans validation.
- Rapporter une spécificité de 100% sans intervalle de confiance, sur un petit échantillon de témoins sains.

À retenir

- Notations du cours : a = VP, b = FN, c = FP, d = VN ; $\text{Se} = \frac{a}{a+b}$, $\text{Sp} = \frac{d}{c+d}$, $\text{VPP} = \frac{a}{a+c}$, $\text{VPN} = \frac{d}{b+d}$, $\text{Ex} = \frac{a+d}{N}$, $\text{TE} = \frac{b+c}{N}$.
- La Se et la Sp ne dépendent pas de la prévalence (en principe), mais dépendent du spectre des patients ; la VPP et la VPN dépendent de la prévalence et se recalculent avec le théorème de Bayes.
- SnNout : un test très sensible, s'il est négatif, exclut la maladie ; SpPin : un test très spécifique, s'il est positif, la confirme.
- J = Se + Sp - 1 ; $\text{RV}^+ = \frac{\text{Se}}{1-\text{Sp}}$; $\text{RV}^- = \frac{1-\text{Se}}{\text{Sp}}$; $\text{DOR} = \frac{a \times d}{b \times c}$.
- Courbe ROC : Se en fonction de 1 - Sp pour tous les seuils ; AUC = probabilité qu'un malade choisi au hasard ait une valeur plus élevée qu'un sujet sain choisi au hasard.
- Échelle du cours : 0,91-1 excellent ; 0,81-0,90 très bon ; 0,71-0,80 bon ; 0,61-0,70 satisfaisant ; moins de 0,6 faible. L'AUC est rapportée avec son IC à 95%.

- Sur mg2026 : FINDRISC AUC 0,852 (IC à 95% 0,762-0,943), glycémie à jeun 0,844, différence non significative (p = 0,81) ; seuil de Youden ≥ 11 (Se 93,1%, Sp 66,4%).
- Le seuil se choisit selon l'objectif : exclusion (Se et VPN élevées), confirmation (Sp et VPP élevées), avec une zone grise entre les deux.
- Le kappa mesure l'accord au-delà du hasard pour des catégories (échelle de Landis et Koch) ; l'ICC pour des mesures numériques.
- Les classificateurs d'IA s'évaluent avec les mêmes indicateurs (rappel = Se, précision = VPP, F1) ; la publication suit STARD 2015 et STARD-AI.

Questions d'autoévaluation

1. (QRU) Le taux d'erreur de classification d'un test, avec les notations du cours (a = VP, b = FN, c = FP, d = VN), est :
   a) $\frac{a+d}{N}$
   b) $\frac{b+d}{N}$
   c) $\frac{b+c}{N}$
   d) $\frac{b}{a+b}$
   e) $\frac{c}{c+d}$
2. (QRU) Un test ayant une sensibilité de 99% est négatif chez un patient dont la probabilité pré-test est modérée. La conclusion la plus appropriée est :
   a) la maladie est confirmée
   b) la maladie devient très peu probable (SnNout)
   c) le résultat n'apporte aucune information
   d) il faut calculer la spécificité pour interpréter un résultat négatif, la sensibilité n'étant pas pertinente
   e) la valeur prédictive positive est de 99%
3. (QRU) Lequel des indicateurs suivants dépend le plus directement de la prévalence de la maladie dans la population testée ?
   a) la sensibilité
   b) la spécificité
   c) l'indice de Youden
   d) la valeur prédictive positive
   e) le rapport de vraisemblance positif
4. (QRU) L'aire sous la courbe ROC du score FINDRISC pour le diabète est de 0,852. L'interprétation correcte est :
   a) 85,2% des patients sont correctement classés
   b) la sensibilité du score est de 85,2%
   c) un patient diabétique choisi au hasard a un score plus élevé qu'un patient non diabétique choisi au hasard dans environ 85,2% des paires
   d) la valeur prédictive positive est de 85,2%
   e) 85,2% des patients ayant un score positif ont un diabète
5. (QRU) Deux évaluateurs obtiennent un coefficient kappa de 0,55. Sur l'échelle de Landis et Koch, l'accord est :
   a) léger

   b) acceptable
   c) modéré
   d) substantiel
   e) presque parfait

6. (QRM) À propos du biais de spectre :
   a) il apparaît fréquemment dans les études cas-témoins qui comparent des malades typiques à des volontaires sains
   b) il apparaît lorsque les patients de l'étude diffèrent de ceux chez qui le test sera utilisé en pratique
   c) il apparaît lorsque le standard de référence n'est appliqué qu'aux patients dont le test est positif
   d) il peut modifier à la fois la sensibilité et la spécificité
   e) il est éliminé en augmentant le nombre de patients de l'étude
7. (QRM) À propos de la courbe ROC :
   a) elle porte la sensibilité sur l'axe vertical et 1 - spécificité sur l'axe horizontal
   b) la diagonale correspond à un test sans valeur discriminante (AUC = 0,5)
   c) chaque point de la courbe correspond à un seuil de décision
   d) l'aire sous la courbe dépend directement de la prévalence de la maladie dans l'échantillon
   e) le point ayant l'indice de Youden maximal est le plus éloigné verticalement de la diagonale
8. (QRM) Pour le tableau VP = 22, FN = 7, FP = 33, VN = 98 :
   a) la sensibilité est de 75,9%
   b) la valeur prédictive positive est de 40,0%
   c) le taux d'erreur de classification est (7 + 98)/160
   d) le rapport de cotes diagnostique est d'environ 9,33
   e) la spécificité est 98/105
9. (QRM) À propos de l'évaluation des classificateurs d'intelligence artificielle :
   a) la précision (angl. precision) équivaut à la valeur prédictive positive
   b) le rappel (recall) équivaut à la sensibilité
   c) l'exactitude est un indicateur suffisant lorsque la classe positive est rare
   d) le score F1 est la moyenne harmonique de la précision et de la sensibilité
   e) la recommandation STARD-AI exige la description de la séparation des données d'entraînement, de validation et de test
10. (QRM) À propos de la comparaison de deux courbes ROC :

a) pour deux tests réalisés chez les mêmes patients, la méthode de DeLong tient compte de la corrélation entre les AUC
b) le chevauchement des intervalles de confiance des AUC prouve que les tests sont équivalents
c) la différence FINDRISC - glycémie (0,008 ; p = 0,81) dans le jeu mg2026 montre qu'aucune différence significative de discrimination n'a été trouvée
d) une différence non significative entre les AUC démontre l'équivalence des tests
e) deux courbes ROC peuvent se croiser, de sorte que chaque test soit meilleur à des seuils différents

Chapitre 13. Les essais cliniques : classification, conception, protocole, outils de recueil des données et analyse intégrée des données biomédicales

Objectifs

À la fin de ce chapitre, l'étudiant sera capable de :

- distinguer une étude clinique interventionnelle d'une étude observationnelle et situer une étude dans une phase de développement (0-IV) ;
- reconnaître le schéma d'une étude (groupes parallèles, croisé, factoriel, en grappes, pragmatique ou explicatif, pilote à un seul bras) et ses conséquences sur l'analyse statistique ;
- expliquer la randomisation simple, par blocs et stratifiée, la dissimulation de l'allocation et l'insu, avec l'argument statistique qui sous-tend chacune ;
- définir le critère de jugement principal et les critères secondaires, distinguer un critère de substitution d'un critère clinique "dur" et calculer la taille de l'échantillon pour une moyenne ou une proportion ;
- comparer l'analyse en intention de traiter à l'analyse per protocole et expliquer pourquoi les données manquantes ne s'"effacent" pas tout simplement ;
- décrire la structure d'un protocole, les obligations d'enregistrement et le cadre éthique et juridique (Déclaration d'Helsinki, BPC, Règlement UE 536/2014, RGPD) ;
- concevoir un formulaire de recueil des données et un questionnaire, choisir le type de question adapté à chaque variable et interpréter les indicateurs de fidélité et de validité (alpha de Cronbach, ICC, KMO, analyse factorielle) ;
- construire un formulaire électronique dans Google Forms, avec des règles de validation, des sections et des embranchements, et nettoyer un jeu de données brut ;
- rédiger le plan d'analyse d'un essai randomisé et l'appliquer au jeu mg2026, du "Tableau 1" jusqu'à la conclusion et aux limites, en rapportant les résultats conformément à CONSORT 2025.

13.1 Qu'est-ce qu'un essai clinique et qu'exige-t-il de la statistique

Jusqu'ici, nous avons travaillé sur des données déjà recueillies. Nous avons décrit des variables, estimé des intervalles de confiance, testé des hypothèses, mesuré des associations et jugé la performance d'un test diagnostique. Dans ce chapitre, nous faisons un pas en arrière, là où naissent les données : dans le projet d'une étude. Un résultat statistique impeccable calculé sur des données mal recueillies reste un mauvais résultat. Réciproquement, une étude bien pensée rend l'analyse presque mécanique.

Un essai clinique est une recherche menée chez l'être humain, dans laquelle l'investigateur attribue, selon un plan établi à l'avance, une intervention (un médicament, une procédure, un programme éducatif, une application mobile) et suit son effet sur certains critères d'évaluation de la santé. Le mot qui compte est "attribue". Dans une étude observationnelle, le chercheur se contente d'observer : les patients fument ou non, sont obèses ou non, reçoivent un traitement parce que leur médecin traitant en a décidé ainsi. Dans une étude interventionnelle, l'exposition est décidée par le protocole, idéalement par tirage au sort. Cette différence change tout dans l'interprétation : des études observationnelles, nous tirons des conclusions d'association ; des essais randomisés bien conduits, nous nous autorisons des conclusions causales [63,64].

Pourquoi un essai clinique a-t-il besoin de la statistique dès le début ? Parce que nous devons savoir, avant d'inclure le premier patient, combien de patients il nous faut (chapitre 4), quelle hypothèse nous testons et avec quel seuil de signification (chapitre 5), quel test nous utiliserons pour le critère de jugement principal (chapitres 6-8), comment nous exprimerons l'effet (différence de moyennes, risque relatif, NST, chapitre 10) et comment nous traiterons les patients qui abandonnent. Toutes ces décisions sont écrites dans le protocole et dans le plan d'analyse statistique. Si nous les prenons après avoir vu les données, nous risquons de choisir, consciemment ou non, la variante qui "sort" significative. Les guides méthodologiques de la discipline insistent précisément sur ce point : le plan précède les données [18,111].

Dans la hiérarchie des preuves de la médecine fondée sur les preuves, l'essai contrôlé randomisé (ECR ; angl. randomised controlled trial, RCT) se situe au-dessus des études de cohorte, cas-témoins et transversales, et en dessous de la synthèse systématique de plusieurs ECR. Cette hiérarchie n'est cependant pas une échelle rigide. Un petit ECR, avec beaucoup d'abandons et sans insu, peut fournir des preuves plus faibles qu'une grande cohorte soigneusement ajustée. C'est pourquoi le médecin qui lit un article doit juger la qualité concrète de l'étude, et pas seulement l'étiquette sur la couverture. Les systèmes modernes de gradation de la certitude des preuves partent du schéma d'étude, puis abaissent ou relèvent le niveau en fonction du risque d'erreur systématique, de l'imprécision et de l'incohérence [128].

Tout au long du chapitre, nous nous appuierons sur le jeu de données pédagogique mg2026. Il simule un essai randomisé pragmatique mené dans 12 cabinets de médecine générale : 160 adultes à risque cardiométabolique, répartis en trois bras (A, conseil standard ; B, conseil et application mobile d'automesure ; C, conseil, application et téléconsultation mensuelle), suivis à 0, 3 et 6 mois. Nous le regarderons successivement avec les yeux de celui qui conçoit l'étude, de celui qui recueille les données et de celui qui les analyse. À la fin, nous rédigerons ensemble le plan d'analyse intégré.

13.2 Classification des études cliniques

13.2.1 Interventionnelle ou observationnelle

La première question à se poser en ouvrant un article est : qui a décidé de l'exposition ? Si c'est l'investigateur, l'étude est interventionnelle (expérimentale). Si l'exposition s'est produite indépendamment de l'étude, celle-ci est observationnelle : cohorte, cas-témoins, transversale (chapitre 10, figure 10.1). Les études observationnelles peuvent être prospectives ou rétrospectives, analytiques ou descriptives, mais aucune n'a le pouvoir de la randomisation d'équilibrer les facteurs de confusion inconnus.

Il existe aussi des formes intermédiaires. Une étude non randomisée à deux groupes, dans laquelle les patients ont choisi eux-mêmes le type de suivi ou ont été répartis selon la disponibilité des appareils, est interventionnelle en ce sens que l'intervention est appliquée selon un protocole, mais elle ressemble à une cohorte du point de vue de la comparabilité des groupes. Une étude avant-après sans groupe contrôle est elle aussi interventionnelle, mais elle ne peut pas séparer l'effet du traitement de l'évolution naturelle de la maladie, de la régression vers la moyenne ou de l'effet placebo. Nous les rencontrerons toutes dans les études de cas de ce chapitre.

13.2.2 Les phases du développement d'un médicament

Pour les médicaments, les essais cliniques suivent un parcours standardisé, présenté dans la figure 13.1. Avant l'être humain, il existe une étape préclinique (laboratoire, modèles animaux). On ajoute parfois une phase 0, avec des microdoses administrées à un très petit nombre de volontaires, afin de vérifier si la substance se comporte dans l'organisme humain comme le

suggéraient les études précliniques ; elle n'a pas de but thérapeutique et ne répond pas aux questions d'efficacité.

La phase I inclut habituellement 20-80 volontaires (sains ou, en oncologie, des patients) et étudie la sécurité, la tolérance, la pharmacocinétique et la dose maximale tolérée. La statistique y est essentiellement descriptive : fréquences des effets indésirables, moyennes et intervalles pour les concentrations. La phase II, avec 100-300 patients, recherche des signaux d'efficacité et la dose optimale ; c'est là qu'apparaissent les premières comparaisons avec un groupe contrôle et les premières estimations de la taille de l'effet, utilisées ensuite pour calculer la taille de l'échantillon de la phase III. La phase III est le grand essai randomisé, avec des centaines ou des milliers de patients, qui compare le nouveau traitement au standard ou au placebo sur un critère clinique pertinent ; c'est sur ses résultats que l'autorisation de mise sur le marché est accordée. La phase IV commence après l'autorisation et étudie les effets indésirables rares, l'utilisation dans des populations non incluses dans les études antérieures et l'efficience en pratique réelle (pharmacovigilance).

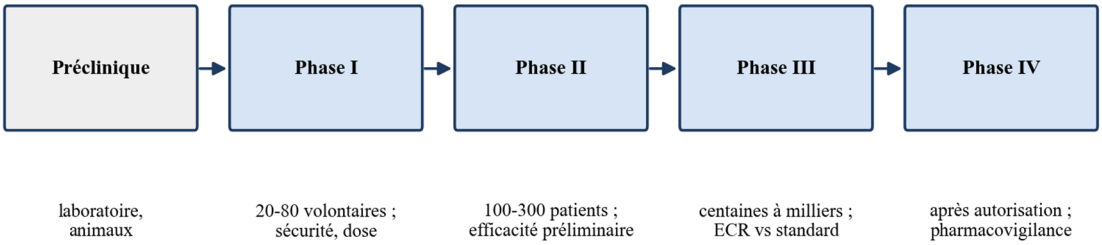


Figure 13.1. Les étapes du développement d'un médicament : préclinique (laboratoire, animaux), phase I (20-80 volontaires, sécurité et dose), phase II (100-300 patients, efficacité initiale), phase III (centaines-milliers de patients, ECR versus standard), phase IV (après autorisation, pharmacovigilance).

La logique statistique des phases est celle d'un entonnoir. Un effet indésirable d'une fréquence de 1 sur 1000 a de fortes chances de ne pas apparaître du tout dans un essai de phase III portant sur 600 patients. La "règle de 3", approximative, dit que, si nous n'avons observé aucun événement chez n patients, la borne supérieure de l'intervalle de confiance à 95% de sa fréquence est d'environ 3/n ; pour n = 600, cela signifie 0,5%. L'absence d'un effet indésirable dans les essais d'autorisation ne prouve pas qu'il n'existe pas, mais seulement qu'il est probablement plus rare que cela. D'où la nécessité de la phase IV.

Les études qui ne testent pas de médicaments (programmes éducatifs, applications mobiles, interventions chirurgicales, dispositifs) ne respectent pas toujours cette numérotation, mais elles suivent la même progression : étude de faisabilité ou pilote, puis étude comparative de grande taille. Le scénario mg2026 correspondrait à une étude de type phase III pour une intervention non pharmacologique : il compare des stratégies déjà considérées comme sûres, sur un critère d'efficacité, dans des conditions proches de la pratique réelle.

13.2.3 Selon l'objectif : supériorité, non-infériorité, équivalence

La plupart des études que vous lirez sont des études de supériorité : elles veulent montrer que le nouveau traitement est meilleur que le comparateur. Les hypothèses sont celles du chapitre 5 : H0,

il n'existe pas de différence ; H1, il existe une différence. Le rejet de H0 soutient la supériorité (si la différence va dans le sens attendu).

Parfois, nous ne cherchons pas un traitement plus efficace : nous voulons un traitement aussi efficace, mais moins coûteux, plus commode ou avec moins d'effets indésirables. Nous formulons alors une étude de non-infériorité. Nous fixons à l'avance une marge de non-infériorité Δ, la plus grande perte d'efficacité que nous jugeons cliniquement acceptable. Le nouveau traitement est déclaré non inférieur si la borne de l'intervalle de confiance de la différence (nouveau moins standard) ne dépasse pas -Δ. Un exemple : si la téléconsultation remplaçait les visites en présentiel, nous pourrions accepter une baisse de la glycémie inférieure d'au plus 3 mg/dL à celle obtenue avec les visites en présentiel ; si l'IC à 95% de la différence est, par exemple, de -2,1 à 4,5 mg/dL, la borne inférieure -2,1 est au-dessus de -3 et nous concluons à la non-infériorité. Dans une étude d'équivalence, nous exigeons que l'intervalle entier se situe entre -Δ et +Δ.

Remarquez qu'ici c'est l'intervalle de confiance, et non la valeur p, qui décide. Une étude de supériorité au résultat non significatif ne démontre pas l'équivalence : "l'absence de preuve n'est pas la preuve de l'absence" [79]. Si l'étude était petite, l'intervalle est large et inclut à la fois des effets importants en faveur du traitement et des effets importants en sa défaveur. Seule une étude conçue dès le départ pour la non-infériorité, avec une marge cliniquement justifiée et un échantillon suffisant, peut étayer une telle conclusion.

13.2.4 Selon le schéma d'étude

La figure 13.2 résume quatre schémas que vous rencontrerez fréquemment. Chacun a une conséquence directe sur le test statistique choisi.

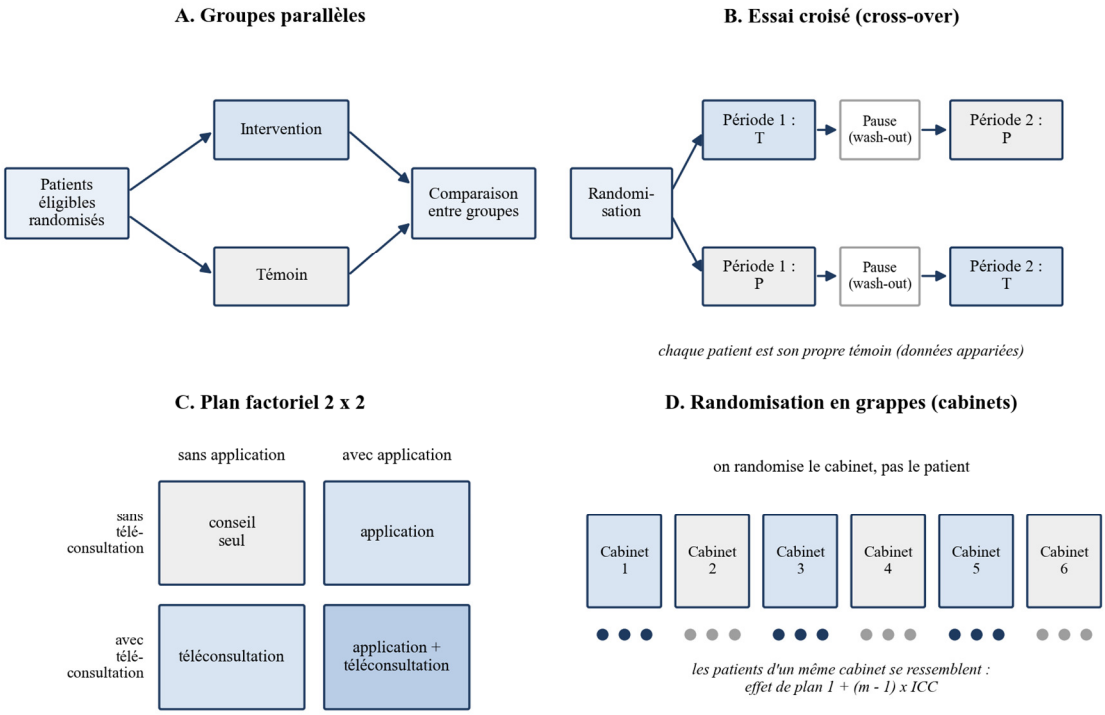


Figure 13.2. Quatre schémas d'essai clinique. A : groupes parallèles, chaque patient reçoit une seule intervention et l'on compare les groupes entre eux. B : schéma croisé (cross-over), chaque patient reçoit les deux traitements dans un ordre aléatoire, séparés par une période d'élimination, les données sont donc appariées. C : plan factoriel 2 x 2, deux interventions testées simultanément. D : randomisation en grappes, on randomise le cabinet, et les patients d'un même cabinet se ressemblent entre eux.

Groupes parallèles. Chaque patient est alloué à un seul bras et y reste jusqu'à la fin. C'est le schéma de l'étude mg2026 et de la majorité des ECR. Les groupes sont indépendants ; nous utilisons donc le test t pour échantillons indépendants ou le test de Mann-Whitney pour deux bras, l'ANOVA ou le test de Kruskal-Wallis pour trois bras ou plus, le khi-deux pour les proportions.

Schéma croisé (cross-over). Chaque patient reçoit successivement les deux traitements, l'ordre étant tiré au sort ; entre les périodes, on ménage une période d'élimination (angl. washout) afin que l'effet du premier traitement disparaisse. Chaque patient est son propre témoin, la variabilité entre patients disparaît de la comparaison et l'échantillon nécessaire diminue fortement. Le fichier BPOC (données de BPCO) utilisé en travaux pratiques de la discipline provient d'une étude de ce type, avec des mesures au début de l'étude et après le traitement. L'analyse utilise des tests pour données appariées : t apparié ou Wilcoxon. Ce schéma ne fonctionne que pour des maladies chroniques stables et des traitements à effet réversible ; il ne peut pas être utilisé pour une intervention qui guérit la maladie ni pour un critère comme le décès. Le risque spécifique est l'effet résiduel (angl. carry-over) : si la période d'élimination a été trop courte, le traitement de la première période contamine la seconde.

Plan factoriel. Il teste simultanément deux interventions (ou plus). Dans un plan 2 x 2, les patients sont randomisés en quatre groupes : rien, A seul, B seul, A et B. Avec le même nombre de patients, nous répondons à deux questions. L'étude mg2026 n'est pas factorielle : le bras C ajoute la téléconsultation à l'application, mais il n'existe pas de bras "téléconsultation seule". C'est pourquoi nous ne pouvons pas estimer l'effet de la téléconsultation sans l'application. Un plan factoriel l'aurait permis et aurait aussi permis de tester une interaction (si l'application et la téléconsultation se potentialisent ou non).

Randomisation en grappes. Parfois, nous ne pouvons pas randomiser des patients individuels. Si le médecin généraliste apprend à faire des téléconsultations, il aura tendance à les appliquer à tous ses patients, y compris à ceux alloués au conseil standard ; il se produit alors une contamination entre les bras. La solution est la randomisation des cabinets (des grappes, angl. clusters) : tous les patients d'un cabinet reçoivent la même intervention. L'étude mg2026 aurait pu être conçue ainsi, les 12 cabinets étant randomisés à raison de 4 par bras. Le prix à payer est statistique. Les patients d'un même cabinet se ressemblent (même médecin, même zone, même style de conseil) et n'apportent donc pas une information indépendante. La corrélation intraclasse (ICC) mesure cette ressemblance, et l'effet de plan indique de combien il faut augmenter l'échantillon :

effet de plan = 1 + (m - 1) x ICC

où m est le nombre moyen de patients par grappe. Avec 13 patients par cabinet et ICC = 0,05, l'effet de plan est de 1 + 12 x 0,05 = 1,6 : il nous faut 60% de patients de plus que dans un essai randomisé individuellement. L'analyse doit tenir compte des grappes ; si nous les ignorons, les erreurs standard sont trop petites et les valeurs p trop optimistes.

Pragmatique ou explicatif. Un essai explicatif demande "l'intervention peut-elle fonctionner dans des conditions idéales ?" : critères d'inclusion stricts, patients motivés, observance surveillée attentivement, souvent un seul centre spécialisé. Un essai pragmatique demande "l'intervention fonctionne-t-elle dans la pratique courante ?" : critères larges, cabinets ordinaires, observance aussi proche que possible de la réalité, critères de jugement pertinents pour le patient. L'étude mg2026 est pragmatique ; c'est pourquoi ses effets sont probablement plus faibles qu'ils ne l'auraient été dans un essai explicatif, mais plus proches de ce que verra le médecin généraliste. La plupart des essais réels se situent quelque part entre ces deux extrêmes.

L'étude pilote à un seul bras. Avant un ECR coûteux, les chercheurs vérifient la faisabilité : peut-on recruter les patients, acceptent-ils l'intervention, les instruments de mesure sont-ils

applicables, quelle semble être l'ampleur de l'effet ? Une étude pilote à un seul bras applique l'intervention à tous et compare les valeurs dans le temps. L'analyse utilise des tests pour données appariées ou pour mesures répétées (t apparié, Wilcoxon, ANOVA à mesures répétées, Friedman). Sa limite est celle de toute étude avant-après : sans groupe contrôle, nous ne pouvons pas attribuer avec certitude l'amélioration à l'intervention.

Étude de cas : l'application mobile chez les patients asthmatiques et l'étude pilote à un seul bras dans la dermatite atopique

Deux études de collègues de la discipline illustrent les extrêmes évoqués ci-dessus. Munteanu et collaborateurs ont testé une application mobile destinée à l'observance thérapeutique chez 93 patients asthmatiques, dans une étude pilote contrôlée à deux groupes (traitement standard versus traitement plus application) avec des mesures répétées sur une année [94]. Le schéma ressemble de façon frappante aux bras A et B de mg2026. L'analyse respecte la structure des données : les groupes indépendants ont été comparés avec le test de Mann-Whitney ($p < 0,001$), l'évolution dans le temps a été analysée avec le test de Friedman pour échantillons dépendants, avec correction de Bonferroni pour les comparaisons multiples ($p < 0,001$), et la proportion d'exacerbations a été comparée par le khi-deux : 10,30% contre 46,30%, $\chi^2(1) = 13,707$, $p < 0,001$. La question que nous nous posons à la lecture est de savoir comment les patients ont été alloués aux deux groupes ; s'ils n'ont pas été randomisés, la différence peut aussi refléter la motivation de ceux qui ont accepté l'application.

À l'autre extrémité, Miulescu et collaborateurs ont suivi 21 enfants atteints de dermatite atopique et d'une dysbiose confirmée au laboratoire, dans une étude pilote prospective à un seul bras, avec des évaluations à l'inclusion, à 30 et à 90 jours [91]. Le score POEM est passé de 14,67 ± 5,22 à 7,10 ± 3,82, et le SCORAD de 41,66 ± 15,28 à 17,93 ± 11,68 (les deux $p < 0,001$), analysés par une ANOVA à mesures répétées, le test adapté à trois temps de mesure chez les mêmes enfants. Environ 76% des enfants ont obtenu une baisse d'au moins 50% du POEM. Le résultat est prometteur, mais le schéma ne permet pas de conclusion causale : la dermatite atopique a une évolution fluctuante, les enfants sont souvent inclus lors d'un pic de la maladie (régression vers la moyenne), et les parents qui savent qu'une nouvelle thérapie a été appliquée peuvent rapporter les symptômes de manière plus favorable. La conclusion correcte est celle que formule le titre lui-même : une étude pilote, qui justifie un ECR.

13.3 Les principes méthodologiques d'un essai randomisé

13.3.1 La randomisation

La randomisation désigne l'allocation des patients aux bras par un mécanisme aléatoire, que ni le patient ni le médecin ne peuvent influencer. Son effet est statistique : à long terme, toutes les caractéristiques des patients, mesurées ou non, connues ou non, se répartissent en moyenne de la même façon entre les bras. Les différences qui subsistent entre les groupes à l'inclusion sont dues au hasard, et le hasard est précisément ce que les tests statistiques savent quantifier. C'est pourquoi un effet significatif dans un ECR bien conduit peut être attribué à l'intervention.

La randomisation simple alloue chaque patient de façon indépendante, comme à pile ou face (ou avec un dé, pour trois bras). Sur de grands échantillons, elle fonctionne bien. Sur de petits échantillons, elle peut produire des déséquilibres d'effectifs : avec 30 patients et deux bras, la probabilité qu'un bras reçoive 10 patients ou moins n'est pas négligeable. La randomisation par blocs corrige ce problème. Nous choisissons des blocs de, par exemple, 6 patients pour trois bras et, dans chaque bloc, nous permutons au hasard la séquence AABBCC. Un bloc pourrait être B A C C A B, le suivant C B A B A C. À la fin de chaque bloc, les bras ont exactement le même nombre de patients. L'inconvénient : si le médecin connaît la taille du bloc, il peut deviner la

dernière allocation du bloc. C'est pourquoi on utilise des blocs de taille variable (3, 6 ou 9, eux aussi choisis au hasard).

La randomisation stratifiée s'utilise lorsqu'un facteur pronostique important doit être équilibré de façon certaine, et pas seulement en moyenne. On crée une liste de randomisation (habituellement par blocs) pour chaque strate : par exemple, pour chaque cabinet ou pour diabète confirmé oui/non. Dans mg2026, une stratification par cabinet garantirait que chaque médecin a des patients dans les trois bras, de sorte que l'"effet médecin" ne se confond pas avec l'effet de l'intervention. Nous ne stratifions pas selon trop de facteurs : avec 12 cabinets, deux sexes et trois catégories de tabagisme, nous obtenons 72 strates, dont beaucoup ne comptent qu'un seul patient, et la stratification perd son sens.

13.3.2 La dissimulation de l'allocation

La séquence de randomisation peut être parfaite et l'étude pourtant compromise. Si le médecin qui inclut les patients voit la liste et sait que le prochain patient ira dans le bras C, il peut retarder l'inclusion d'un patient "difficile" jusqu'à une allocation qui lui convient. La dissimulation de l'allocation (angl. allocation concealment) signifie que l'allocation d'un patient n'est connue qu'après l'inclusion définitive de ce patient. Les méthodes sûres sont la randomisation centralisée, par téléphone ou par une plateforme web, et les enveloppes opaques, scellées, numérotées séquentiellement (plus vulnérables, mais acceptables si elles sont gérées correctement). À ne pas confondre avec l'insu : la dissimulation de l'allocation protège le moment de l'inclusion et est toujours possible ; l'insu protège la période qui suit l'allocation et n'est pas toujours possible.

13.3.3 L'insu

L'insu (angl. blinding) signifie qu'un ou plusieurs participants à l'étude ne savent pas quelle intervention a reçue chaque patient : le patient, le médecin traitant, la personne qui évalue le résultat, l'analyste statisticien. Les termes "simple insu" et "double insu" sont utilisés de façon incohérente dans la littérature ; c'est pourquoi la recommandation actuelle est de préciser explicitement qui était en insu, et pas seulement l'étiquette [142]. Le patient qui sait qu'il reçoit le traitement actif peut rapporter de meilleurs symptômes ; le médecin qui le sait peut interpréter plus favorablement un résultat clinique ambigu ; l'évaluateur peut mesurer avec une attention différente.

Dans l'étude mg2026, les patients ne peuvent pas être en insu : ils savent s'ils ont reçu l'application et les téléconsultations. Les médecins généralistes non plus. Il est possible, en revanche, de faire doser la glycémie par un laboratoire qui ne connaît pas le bras et de transmettre à l'analyste des données dont les bras sont codés (X, Y, Z). Les critères objectifs (glycémie de laboratoire, poids mesuré sur une balance étalonnée) sont moins vulnérables à l'absence d'insu que les critères subjectifs (satisfaction, douleur, qualité de vie). La satisfaction mesurée dans mg2026 doit donc être interprétée avec prudence : les patients du bras C ont reçu plus d'attention et sont, naturellement, plus satisfaits.

13.3.4 Le groupe contrôle et le placebo

Sans groupe contrôle, nous ne pouvons pas séparer l'effet de l'intervention de l'évolution naturelle, de la régression vers la moyenne, de l'effet placebo et de l'effet de la simple participation à une étude (les patients observés modifient leur comportement, phénomène appelé effet Hawthorne). Le contrôle peut être un placebo (une substance inactive d'aspect identique), le traitement standard, l'absence d'intervention ou une liste d'attente. L'éthique limite l'utilisation du placebo : lorsqu'il existe un traitement dont l'efficacité est prouvée, les patients du groupe contrôle doivent le recevoir, et le nouveau traitement s'y ajoute ou lui est comparé. Dans mg2026, le bras A reçoit

le conseil standard ; il serait contraire à l'éthique de laisser des patients prédiabétiques sans aucune recommandation sur le mode de vie.

13.3.5 Les critères de jugement

Le critère de jugement principal (angl. primary endpoint) est la variable sur laquelle repose la conclusion de l'étude et pour laquelle la taille de l'échantillon a été calculée. Il doit être unique, défini précisément (ce que l'on mesure, comment, quand) et choisi avant le début de l'étude. Dans mg2026, il s'agit de la baisse de la glycémie à jeun entre l'inclusion et 6 mois (glicemie0 - glicemie6). Les critères secondaires apportent une information supplémentaire : perte de poids d'au moins 5%, baisse de la pression artérielle, satisfaction. Leurs résultats sont considérés comme des éléments de soutien ou comme générateurs d'hypothèses, car, lorsqu'on teste de nombreux critères, la probabilité que l'un d'eux sorte significatif par simple hasard augmente. Avec 10 critères indépendants, testés chacun à α = 0,05, la probabilité d'obtenir au moins un résultat faux positif est de 1 - 0,95^10 = 0,40.

Un critère de substitution est un marqueur biologique qui tient lieu d'un résultat clinique important : la glycémie ou l'HbA1c à la place des complications du diabète, le cholestérol LDL à la place de l'infarctus, la pression artérielle à la place de l'accident vasculaire cérébral. Les critères de substitution permettent des études plus petites et plus courtes, mais l'histoire de la médecine compte suffisamment d'exemples de médicaments qui ont amélioré le critère de substitution et aggravé le résultat clinique. Les critères "durs" (décès, infarctus, accident vasculaire cérébral, hospitalisation) sont ceux qui comptent pour le patient, mais ils exigent de grands échantillons et un suivi long. Un critère composite réunit plusieurs événements (par exemple décès cardiovasculaire, infarctus ou AVC) afin d'augmenter le nombre d'événements ; il doit être lu avec attention, car l'effet peut être dû presque entièrement à la composante la moins grave.

13.3.6 La taille de l'échantillon et la puissance

Nous avons appris au chapitre 4 que la taille de l'échantillon dépend de quatre éléments [42,53] : le seuil α (habituellement 0,05), la puissance souhaitée 1 - β (habituellement 80% ou 90%), la taille de l'effet que nous voulons détecter et la variabilité du critère. Pour la comparaison de deux moyennes, la formule approchée par groupe est

$$n = \frac{2 \times \left(z\left(1 - \frac{\alpha}{2}\right) + z(1-\beta) \right)^2 \times \sigma^2}{\Delta^2}$$

où Δ est la plus petite différence cliniquement pertinente et σ l'écart-type. Exprimée à l'aide de la taille d'effet standardisée $d = \frac{\Delta}{\sigma}$, elle devient $n = \frac{2 \times (1{,}96 + 0{,}84)^2}{d^2}$. Pour d = 0,5, la formule donne environ 63, et le calcul exact avec la distribution t, réalisé dans des logiciels comme G*Power [77], donne 64 patients par groupe ; pour d = 0,3, 176 par groupe. Pour deux proportions p1 et p2 :

$$n = \frac{\left[z\left(1 - \frac{\alpha}{2}\right) \times \sqrt{2 \times \bar{p} \times (1 - \bar{p})} + z(1-\beta) \times \sqrt{p_1 \times (1-p_1) + p_2 \times (1-p_2)} \right]^2}{(p_1 - p_2)^2}$$

avec $\bar{p} = \frac{p_1 + p_2}{2}$. Au résultat obtenu, nous ajoutons une marge pour les abandons : si nous estimons que 10% des patients seront perdus, nous divisons n par 0,9. Le problème résolu 13.1 présente le calcul complet pour mg2026.

Une étude trop petite est inutile et, plus grave, contraire à l'éthique : elle expose des patients aux risques de la recherche sans chance raisonnable de répondre à la question. Un résultat non significatif dans une étude de faible puissance ne nous dit presque rien sur l'absence d'effet.

13.3.7 L'intention de traiter et l'analyse per protocole

Dans toute étude réelle, certains patients ne respectent pas le protocole : ils n'installent pas l'application, ne se présentent pas aux téléconsultations, changent de bras, abandonnent. Nous disposons de deux grandes options d'analyse. L'analyse en intention de traiter (angl. intention-to-treat, ITT) analyse chaque patient dans le bras dans lequel il a été randomisé, quel que soit ce qu'il a effectivement reçu. L'analyse per protocole (PP) ne conserve que les patients qui ont respecté le protocole.

L'analyse ITT préserve le bénéfice de la randomisation. Les patients qui n'adhèrent pas au traitement ne forment pas un sous-ensemble aléatoire : ils sont habituellement plus malades, moins motivés ou présentent des effets indésirables. Si nous les éliminons du bras actif, mais pas les patients comparables du bras contrôle, les groupes restants ne sont plus comparables. L'analyse ITT répond à la question pratique "que se passe-t-il si je recommande cette stratégie ?" et tend à sous-estimer l'effet biologique, car elle dilue l'effet avec les patients non observants. Dans les études de supériorité, cette sous-estimation est conservatrice, et l'ITT est donc l'analyse principale. Dans les études de non-infériorité, la dilution favoriserait la conclusion de non-infériorité ; c'est pourquoi on exige les deux analyses en espérant qu'elles concordent. Dans mg2026, il n'y a pas de perdus de vue (figure 13.3), de sorte que l'ITT coïncide avec l'analyse des 160 patients ; sur des données réelles, ce n'est jamais le cas.

13.3.8 Les données manquantes et les perdus de vue

Un patient qui ne vient pas à la visite de 6 mois laisse une cellule vide dans le tableau. La solution la plus commode, l'analyse des cas complets (suppression du patient), n'est correcte que si les données sont manquantes de façon complètement aléatoire, c'est-à-dire si leur absence n'a de lien ni avec les caractéristiques du patient ni avec le résultat. C'est rarement le cas. Le patient qui n'a pas maigri et se sent coupable ne revient plus en consultation ; si nous le supprimons, son bras paraît plus efficace qu'il ne l'est.

On décrit trois mécanismes. Données manquantes de façon complètement aléatoire : les données manquent pour des raisons sans rapport avec l'étude (un tube cassé). Données manquantes de façon aléatoire conditionnellement aux variables observées : la probabilité d'absence dépend d'éléments que nous avons mesurés (par exemple, les personnes âgées sont plus souvent absentes), et nous pouvons corriger en les utilisant. Données manquantes de façon non aléatoire : la probabilité d'absence dépend de la valeur manquante elle-même (ceux dont la glycémie est élevée ne reviennent plus) ; ici, aucune méthode ne corrige complètement. Les méthodes modernes (imputation multiple, modèles mixtes) dépassent le niveau de ce cours. Ce qu'il faut retenir : rapportez combien de données manquent dans chaque bras et pourquoi ; ne remplacez pas la valeur manquante par la moyenne du groupe (cela réduit artificiellement la variabilité) ; le remplacement par la dernière valeur observée est une solution ancienne, qui suppose que le patient est resté "figé" et peut fausser le résultat ; faites une analyse de sensibilité, avec différentes hypothèses sur les perdus de vue, et vérifiez si la conclusion reste la même.

13.3.9 Les analyses intermédiaires

Dans les études longues, un comité indépendant de surveillance des données et de la sécurité analyse les données à des moments préétablis. Il peut arrêter l'étude plus tôt en raison d'un bénéfice évident, d'effets indésirables ou de futilité (lorsqu'il devient improbable que l'étude montre un jour un effet). Chaque analyse intermédiaire est un test statistique de plus et, si nous utilisions chaque fois le seuil de 0,05, le risque global d'erreur de type I augmenterait. C'est pourquoi on utilise des seuils beaucoup plus stricts lors des analyses intermédiaires (de l'ordre de $p < 0{,}001$), en conservant pour l'analyse finale un seuil proche de 0,05. Les études arrêtées

précocement pour bénéfice tendent à surestimer l'effet, car elles se sont arrêtées précisément au moment où le hasard avait produit une grande différence.

Étude de cas : un essai contrôlé contre placebo rapporté selon CONSORT et une cohorte non randomisée de faible puissance

Gherbon et collaborateurs ont comparé, pendant 3 mois, 20 personnes obèses ayant reçu un régime et de l'Aloe ferox (2 gélules par jour, 460 mg) à 20 personnes appariées (angl. matched) ayant reçu un régime et un placebo [85]. L'étude est rapportée explicitement selon CONSORT, ce qui signifie que le lecteur trouve dans l'article le flux des patients, le mode d'allocation et la définition des critères. Le résumé rapporte deux séries de valeurs p : d'une part pour l'IMC (p = 0,03), le cholestérol total (p = 0,032), le LDL (p = 0,01) et la glycémie à jeun (p = 0,018), et d'autre part pour l'IMC (p = 0,002), le LDL (p = 0,039), la glycémie à jeun (p < 0,001) et la pression diastolique (p = 0,006). À la lecture, la première chose à clarifier est de savoir quelle série correspond à la comparaison entre groupes à 3 mois et laquelle à la comparaison avant-après dans le groupe traité. Seule la comparaison entre groupes teste l'effet de l'Aloe ferox ; la comparaison avant-après dans le groupe actif inclut aussi l'effet du régime, que les deux groupes ont reçu. Deuxièmement : avec 20 patients par groupe, l'étude n'a de puissance que pour des effets importants ; de nombreux critères testés simultanément augmentent le risque d'un résultat faux positif, si bien que le critère de jugement principal doit être identifié dans le protocole.

Popovici et collaborateurs ont comparé, dans une cohorte prospective de 100 patients ayant subi une CPRE, la sédation guidée par le monitorage de l'entropie au monitorage standard [78]. Les groupes n'ont pas été randomisés. Les épisodes hypertensifs ont été plus rares dans le groupe entropie (6% contre 26%, p = 0,007), et les scores d'Aldrete ont été meilleurs à 5 minutes (p = 0,003) et à 15 minutes (p < 0,001). Les auteurs rapportent honnêtement que la puissance statistique a été limitée (29,3%). La leçon est double : une puissance de 29,3% signifie que, si l'effet recherché existait, l'étude n'aurait qu'environ une chance sur trois de le détecter, de sorte qu'un résultat non significatif pour le critère de jugement principal n'exclut pas l'effet ; et l'absence de randomisation laisse place aux facteurs de confusion (qui a décidé du monitorage reçu par le patient ?). Les résultats sont, pour le formuler correctement, générateurs d'hypothèses pour un ECR.

13.4 Le protocole, l'enregistrement et le cadre éthique

13.4.1 Le protocole

Le protocole est le document qui décrit l'étude avant qu'elle ne commence. Il fonctionne comme un contrat entre les chercheurs, les patients, le comité d'éthique et les futurs lecteurs : tout ce qui est fait dans l'étude doit y avoir été annoncé, et tout écart doit être justifié. Pour les protocoles d'essais cliniques, il existe une recommandation internationale de contenu, SPIRIT, construite en miroir de la recommandation de publication CONSORT : ce qui est promis dans le protocole est ensuite rapporté dans l'article. Des collègues de la discipline ont eux aussi publié des protocoles d'étude, par exemple le protocole de traduction en roumain de trois questionnaires de littératie en santé de l'étude ALSATION [26], ce qui montre que le protocole n'est pas réservé aux essais médicamenteux.

Un protocole complet contient, pour l'essentiel, les éléments suivants : le titre et l'identifiant d'enregistrement ; la justification (ce que l'on sait, ce que l'on ne sait pas, pourquoi l'étude est nécessaire) ; les objectifs et les hypothèses, formulés selon le format PICO (population, intervention, comparateur, critère de jugement) [24] ; le schéma d'étude (parallèle, croisé, en grappes ; supériorité ou non-infériorité) ; le lieu de réalisation ; les critères d'inclusion et

d'exclusion ; les interventions, décrites avec une précision suffisante pour qu'un autre centre puisse les reproduire ; le critère de jugement principal et les critères secondaires, avec le moment de la mesure ; le calendrier des visites ; le calcul de la taille de l'échantillon, avec toutes ses hypothèses ; la méthode de randomisation, la dissimulation de l'allocation et l'insu ; les méthodes de recueil et de gestion des données ; le plan d'analyse statistique (populations d'analyse ITT et PP, tests, traitement des données manquantes, analyses de sous-groupes préétablies, analyses intermédiaires) ; la surveillance de la sécurité ; les aspects éthiques, le consentement et la confidentialité ; la politique de publication et de partage des données.

Pour un étudiant de première année, la partie la plus utile à exercer est la traduction de la question clinique en éléments statistiques. Dans mg2026 : P, adultes de 35-79 ans à risque cardiométabolique suivis dans des cabinets de médecine générale ; I, conseil plus application (B) ou conseil plus application plus téléconsultation (C) ; C, conseil standard (A) ; O principal, baisse de la glycémie à jeun à 6 mois. L'hypothèse statistique principale devient H0 : $\mu A = \mu B = \mu C$ pour la baisse de la glycémie, et le test prévu est l'ANOVA à un facteur avec comparaisons de Tukey, ou le test de Kruskal-Wallis si les conditions de l'ANOVA ne sont pas remplies.

13.4.2 L'enregistrement de l'étude

Un essai clinique est enregistré dans un registre public avant l'inclusion du premier patient. Les plus connus sont ClinicalTrials.gov (le registre américain, utilisé aussi par de nombreux chercheurs européens) et, pour les essais médicamenteux de l'Union européenne, le système CTIS (Clinical Trials Information System), introduit par le Règlement UE 536/2014 [143]. Les revues qui suivent les recommandations de l'ICMJE ne publient pas les résultats d'un essai clinique qui n'a pas été enregistré de façon prospective [36]. La raison est statistique, et non bureaucratique : l'enregistrement fixe publiquement le critère de jugement principal. Si, à la fin, les auteurs présentent comme "principal" un critère qui était initialement secondaire, parce que celui-ci est sorti significatif, la comparaison avec le registre révèle le changement. Les études aux résultats négatifs, qui resteraient autrement non publiées, deviennent elles aussi visibles, et les synthèses systématiques ne sont plus faussées par le biais de publication.

13.4.3 L'éthique de la recherche sur les êtres humains

Le fondement éthique de la recherche médicale est la Déclaration d'Helsinki de l'Association médicale mondiale, révisée pour la dernière fois en 2024 [35]. Ses principes peuvent se résumer ainsi : le bien-être et les droits du participant priment sur l'intérêt de la science et de la société ; les risques doivent être proportionnés aux bénéfices attendus ; la recherche doit être scientifiquement justifiée et conduite par des personnes qualifiées ; le protocole est approuvé par un comité d'éthique indépendant avant le début de l'étude ; la participation est volontaire et repose sur le consentement éclairé ; les groupes vulnérables bénéficient d'une protection supplémentaire ; l'utilisation du placebo n'est autorisée que dans des conditions strictes ; les résultats, y compris négatifs, doivent être rendus publics.

Le consentement éclairé est un processus, et pas seulement une signature. Le patient reçoit des informations sur le but de l'étude, les procédures, les risques et les bénéfices possibles, les alternatives, le droit de se retirer à tout moment sans conséquence sur sa prise en charge médicale et la manière dont ses données seront protégées. Les informations sont données dans un langage que le patient comprend, avec un délai de réflexion et la possibilité de poser des questions. Pour les mineurs, on exige l'accord des parents et, selon l'âge, l'assentiment de l'enfant.

Le comité d'éthique de la recherche (dans les hôpitaux et les universités de Roumanie, le comité d'éthique de l'établissement) évalue le protocole, les documents destinés aux patients, la compétence de l'équipe et la justification scientifique. Du point de vue statistique, le comité

vérifie aussi la taille de l'échantillon : une étude sans puissance suffisante expose des patients à des risques sans perspective de résultat utile.

Les bonnes pratiques cliniques (BPC ; angl. Good Clinical Practice, GCP) constituent la norme internationale pour la conception, la conduite, l'enregistrement et la publication des essais cliniques ; la version actuelle de la ligne directrice ICH E6, révision 3, met l'accent sur la qualité pensée dès la conception et sur l'approche fondée sur le risque [144]. Dans l'Union européenne, les essais cliniques de médicaments sont régis par le Règlement 536/2014, qui a harmonisé la procédure d'autorisation et introduit le portail unique CTIS [143]. Le traitement des données personnelles de santé dans la recherche relève du Règlement général sur la protection des données (RGPD), qui considère les données concernant la santé comme une catégorie particulière, ne pouvant être traitée à des fins de recherche qu'avec des garanties appropriées, telles que la pseudonymisation et la minimisation des données [32], complété en Roumanie par la loi 190/2018 [33].

Pour l'analyste, le RGPD a des conséquences pratiques. Le fichier d'analyse ne contient ni nom, ni CNP (numéro personnel d'identification roumain), ni adresse, ni numéro de téléphone, mais un code d'étude (dans mg2026, P001-P160). La table de correspondance code-identité est conservée séparément, chez l'investigateur principal, avec un accès restreint. On ne recueille que les variables nécessaires à la question de recherche : si l'âge suffit, nous ne demandons pas la date de naissance. Les fichiers ne sont pas envoyés par courriel non chiffré et ne sont pas conservés sur des clés USB oubliées au fond des poches. Un formulaire Google Forms utilisé dans une étude réelle avec des patients doit être vérifié du point de vue du lieu de stockage des données et de l'accord de l'établissement ; pour les exercices des travaux pratiques, nous utilisons des données fictives.

13.5 Les outils de recueil des données

13.5.1 Le cahier d'observation

Le cahier d'observation (angl. case report form, CRF) est le document, papier ou électronique, dans lequel on consigne les données de chaque patient à chaque visite. Un bon CRF est conçu à partir du plan d'analyse : chaque champ correspond à une variable du dictionnaire des données et chaque variable du plan d'analyse a un champ dans le CRF. Rien de plus, rien de moins.

Quelques règles de conception préviennent la plupart des erreurs que nous verrons à la section 13.6. Chaque champ numérique porte l'unité de mesure imprimée à côté ("taille : ___ cm"), afin de ne pas recevoir 1,68 au lieu de 168. Les champs catégoriels ont des modalités prédéfinies, à cocher, et non du texte libre ; sinon, nous aurons "f", "F", "Feminin" et "femeie" pour la même valeur. Les dates ont un format unique (jj.mm.aaaa). Pour chaque champ, on prévoit un moyen d'indiquer pourquoi la valeur manque (non mesurée, refus du patient, sans objet), de sorte qu'une cellule vide ne soit pas ambiguë. Les variables dérivées (IMC, perte de poids en pourcentage, score FINDRISC) ne sont pas calculées à la main dans le CRF, mais dans le fichier d'analyse, à partir des variables primaires. Le premier champ de chaque page est le code du patient.

13.5.2 Le questionnaire : de la question à la variable

De nombreuses études de médecine générale utilisent des questionnaires : symptômes, qualité de vie, observance, satisfaction, connaissances, comportements alimentaires. La conception d'un questionnaire est un métier, et sa première leçon est que le type de question détermine le type de variable et, à travers lui, le test statistique (chapitres 2 et 8). Le tableau 13.1 établit ce lien.

Tableau 13.1. Types de questions et variables qu'elles produisent.

Type de question	Exemple	Variable obtenue	Description et tests usuels
Réponse numérique libre	"Quel âge avez-vous ? ___ ans"	numérique (continue ou discrète)	moyenne et ÉT ou médiane et EIQ ; t, ANOVA, corrélation
Choix unique, deux modalités	"Fumez-vous actuellement ? Oui / Non"	nominale dichotomique	effectifs, % ; khi-deux, Fisher, RR, OR
Choix unique, modalités non ordonnées	"Milieu de résidence : urbain / rural"	nominale	effectifs ; khi-deux
Choix unique, modalités ordonnées	"Niveau d'études : collège / lycée / universitaire / postuniversitaire"	ordinale	médiane, EIQ ; Mann-Whitney, Kruskal-Wallis, Spearman
Item de Likert	"Je suis satisfait du programme : 1 pas du tout... 5 tout à fait"	ordinale	médiane, effectifs ; tests non paramétriques
Score composé de plusieurs items	somme de 10 items de Likert	quasi numérique	moyenne ou médiane, selon la distribution
Choix multiple	"Quelles maladies chroniques avez-vous ? (cochez tout ce qui s'applique)"	une variable dichotomique par modalité	fréquence de chaque modalité
Texte libre	"Autres observations"	qualitative	analyse thématique ; n'entre pas dans les tests statistiques

Deux erreurs reviennent chaque année dans les projets des étudiants. La première : l'âge ou le poids recueillis en classes ("18-25 ans, 26-35 ans..."). Nous avons transformé une variable numérique en variable ordinale et perdu de l'information ; nous ne pouvons plus calculer la moyenne ni la corrélation. La règle : recueillez la valeur exacte et regroupez, si nécessaire, au moment de l'analyse. La seconde : les questions à deux idées ("Faites-vous du sport et mangez-vous sainement ?") ; le patient qui fait du sport mais mange mal ne sait pas quoi répondre.

La formulation des questions demande elle aussi de l'attention. Les questions doivent être courtes, neutres (et non "N'est-ce pas que l'application vous a aidé ?"), sans double négation, avec une période de référence claire ("au cours des 7 derniers jours"), avec des modalités de réponse exhaustives et qui ne se chevauchent pas ("0-1 heure, 1-2 heures" se chevauchent à la valeur 1). Les questions sensibles (revenus, consommation d'alcool, vie sexuelle) sont posées vers la fin, une fois la confiance établie.

13.5.3 Les échelles de Likert

Une échelle de Likert propose une affirmation et un nombre impair (habituellement 5 ou 7) de modalités ordonnées d'accord ou de désaccord, avec un point neutre au milieu. La variable satisfactie (satisfaction) de mg2026 est un item de Likert à 5 niveaux. Un seul item de Likert est une variable ordinale : la distance entre "1, pas du tout satisfait" et "2" n'est pas nécessairement égale à la distance entre "4" et "5, très satisfait". C'est pourquoi nous le décrivons par la médiane et les effectifs et le comparons avec des tests non paramétriques, comme nous l'avons fait au chapitre 7 (Mann-Whitney C versus A : médianes 4 contre 3, $U = 2265{,}5$, $p < 0{,}001$). Un score obtenu en additionnant plusieurs items de Likert qui mesurent le même construit possède davantage de valeurs possibles et une distribution souvent proche de la loi normale ; en pratique, il est traité comme une variable numérique, après vérification de la distribution.

Les items formulés à l'envers ("Je trouve difficile d'utiliser l'application", dans une échelle qui mesure la facilité d'utilisation) sont recodés avant la sommation : 1 devient 5, 2 devient 4, et ainsi de suite. L'oubli du recodage est l'une des erreurs les plus fréquentes et se reconnaît facilement, car l'item concerné présente une corrélation négative avec le score total.

13.5.4 La fidélité d'un instrument

Un instrument de mesure est fidèle (angl. reliable) s'il donne des résultats cohérents. Pour les questionnaires, on utilise deux types de fidélité.

La cohérence interne indique si les items d'une échelle mesurent la même chose. L'indicateur classique est le coefficient alpha de Cronbach :

$$\alpha = \frac{k}{k-1} \times \left(1 - \frac{\sum s^2(\text{item})}{s^2(\text{ total })}\right)$$

où k est le nombre d'items, s²(item) la variance de chaque item et s²(total) la variance du score total. Si les items sont fortement corrélés entre eux, la variance du total est bien supérieure à la somme des variances des items, et α se rapproche de 1. À titre indicatif, α ≥ 0,70 est considéré comme acceptable pour la recherche, ≥ 0,80 bon, ≥ 0,90 très bon pour des décisions individuelles ; des valeurs supérieures à 0,95 peuvent signaler des items redondants, la même question posée plusieurs fois en d'autres termes. Le coefficient oméga de McDonald est une alternative qui repose sur des hypothèses moins restrictives que l'alpha et qui est de plus en plus souvent rapporté à ses côtés. La corrélation item-total corrigée (corrélation de chaque item avec la somme des autres) identifie les items "faibles" ; un seuil usuel est 0,30.

La fidélité test-retest indique si le même patient, dans le même état, donne les mêmes réponses lors de deux passations séparées par un intervalle (habituellement 1-3 semaines). Pour les scores numériques, on utilise le coefficient de corrélation intraclasse (ICC), et non le coefficient de Pearson. La raison : le r de Pearson mesure seulement si les valeurs augmentent ensemble ; si, lors de la seconde passation, tous les patients ont 5 points de plus, r peut valoir 1 alors que l'accord est faible. L'ICC pénalise aussi de telles différences systématiques. À titre de repère, un ICC inférieur à 0,50 indique une fidélité faible, 0,50-0,75 modérée, 0,75-0,90 bonne et au-delà de 0,90 excellente. Pour les variables catégorielles, l'accord entre deux passations ou entre deux évaluateurs se mesure avec le coefficient kappa [103].

13.5.5 La validité d'un instrument

Un instrument fidèle n'est pas automatiquement valide. Une balance qui affiche constamment 5 kg de trop est fidèle, mais pas valide. La validité indique si l'instrument mesure ce qu'il prétend mesurer.

La validité de contenu est jugée par des experts et par des patients : les items couvrent-ils tous les aspects importants du construit, sont-ils formulés de façon intelligible ? La validité de construit vérifie, par plusieurs voies, si les scores se comportent comme le prédit la théorie. Validité convergente : le score est fortement corrélé à d'autres mesures du même construit. Validité discriminante : il est faiblement corrélé à des construits différents. Validité par groupes connus (angl. known-groups) : des groupes dont nous savons qu'ils diffèrent (par exemple des diététiciens par rapport à la population générale, pour un questionnaire de connaissances en nutrition) obtiennent des scores différents. La validité de critère compare le score à un standard de référence, comme au chapitre 12.

La structure interne d'un questionnaire s'étudie par l'analyse factorielle. L'analyse factorielle exploratoire (AFE, angl. EFA) recherche des groupes d'items corrélés entre eux et suggère combien de facteurs (dimensions) sous-tendent les réponses. Avant de l'appliquer, on vérifie si les données s'y prêtent : l'indice de Kaiser-Meyer-Olkin (KMO) doit être de préférence supérieur à 0,80 (en dessous de 0,50, l'analyse n'a pas de sens), et le test de sphéricité de Bartlett doit être significatif (nous rejetons l'hypothèse que les items ne sont pas corrélés). Le nombre de facteurs se choisit avec l'analyse parallèle, qui compare les valeurs propres observées à celles obtenues à partir de données aléatoires, ou selon d'autres critères. L'analyse factorielle confirmatoire (AFC,

angl. CFA) teste une structure établie à l'avance et rapporte des indices d'ajustement : CFI et TLI (souhaitables au-dessus de 0,90, idéalement au-dessus de 0,95), RMSEA (souhaitable en dessous de 0,08, idéalement en dessous de 0,06) et SRMR (en dessous de 0,08). La modélisation par équations structurelles (SEM) va plus loin et estime les relations entre facteurs. Ces méthodes dépassent le calcul à la main d'un étudiant de première année, mais vous devez les reconnaître et être capables de les lire dans un article.

Lorsqu'un questionnaire validé dans une autre langue est utilisé en Roumanie, il doit être traduit et adapté culturellement, et pas seulement traduit. La procédure standard comprend deux traductions indépendantes en roumain, leur synthèse, la rétrotraduction dans la langue d'origine par des traducteurs qui ne connaissent pas l'original, la comparaison avec l'original, le test sur un petit groupe de patients (entretiens cognitifs : "qu'avez-vous compris par cette question ?"), puis la validation psychométrique dans la population roumaine. Le protocole ALSATION décrit un tel parcours pour trois questionnaires de littératie en santé [26].

Étude de cas : la validation de questionnaires dans les travaux de nos collègues de la discipline

Marin et collaborateurs ont développé un questionnaire fondé sur le modèle d'acceptation de la technologie (TAM) pour la télésurveillance à domicile des patients atteints de cirrhose hépatique et l'ont validé chez 130 patients [10]. La cohérence interne a été très bonne (α de Cronbach = 0,934 ; ω de McDonald = 0,950). L'AFE a identifié une structure à deux facteurs expliquant 77,9% de la variance, et l'AFC l'a confirmée avec des indices acceptables (CFI = 0,927 ; TLI = 0,908, tous deux au-dessus de 0,90, mais en dessous de 0,95). Le modèle structurel a montré que l'utilité perçue prédit fortement l'intention d'utilisation ($\beta = 0,823$, $p < 0,001$), et que la facilité perçue influence l'utilité ($\beta = 0,343$, $p < 0,001$) ; le modèle a expliqué 69,5% de la variance de l'intention. La question critique pour le lecteur est de savoir si 130 patients suffisent pour une AFE et une AFC sur le même échantillon ; idéalement, la structure trouvée de façon exploratoire est confirmée sur un échantillon indépendant.

Dragomir et collaborateurs ont construit le questionnaire MMBQ-44 sur le fardeau de la ménopause, auprès de 308 femmes de 45-65 ans [60]. Les étapes suivent exactement la section 13.5.5 : les corrélations item-total corrigées ont toutes dépassé 0,30 ; les données se prêtaient à l'analyse factorielle (KMO = 0,931 ; χ^2 de Bartlett = 9006,2, $p < 0,001$) ; l'analyse parallèle a identifié quatre facteurs expliquant 48,6% de la variance totale ; la cohérence interne a été de α 0,813-0,941 et ω 0,864-0,953. Les auteurs signalent aussi des effets plancher (beaucoup de réponses à la valeur minimale) pour deux sous-échelles, 37,0% et 23,4%. Cette information compte : une sous-échelle présentant un effet plancher ne peut pas détecter une amélioration chez les patientes qui partent déjà du minimum.

Dicu et collaborateurs ont validé un instrument multidimensionnel d'évaluation des produits de boulangerie à base d'ingrédients alternatifs auprès de 504 consommateurs, avec un sous-échantillon de 30 personnes réévaluées après 14 jours [73]. Outre l'AFE (KMO = 0,933) et l'AFC (CFI = 0,965 ; TLI = 0,959 ; RMSEA = 0,061 ; SRMR = 0,033, tous dans la zone satisfaisante), l'étude rapporte la fidélité test-retest par un ICC de 0,82-0,91, c'est-à-dire une fidélité bonne à excellente, et un α de Cronbach de 0,898-0,938. Le retest n'a cependant porté que sur 30 personnes, et les intervalles de confiance de l'ICC sur un sous-échantillon aussi petit sont larges ; les valeurs ponctuelles doivent être lues avec prudence.

Quatre autres études complètent le tableau. Frandeș et collaborateurs ont validé un instrument d'acceptabilité de la technologie mobile dans l'autogestion du diabète chez 103 patients, avec un α de Cronbach de 0,9 pour l'attitude envers l'utilisation de la technologie mobile et de 0,97 pour une autre dimension, et ont utilisé des corrélations de rang (Spearman, Kendall), adaptées aux données ordinales [54]. Putnoky et collaborateurs ont adapté un

> questionnaire de connaissances générales en nutrition en quatre étapes : fidélité interne sur 412 personnes (0,878), fidélité test-retest sur 46 personnes (supérieure à 0,880), validité par groupes connus sur 96 et validité convergente sur 508 [145]. Şerban et collaborateurs ont validé le questionnaire d'alimentation consciente chez des adultes roumains, avec une fidélité interne de 0,72 et une fidélité externe de 0,83 [146] ; la valeur de 0,72 est à la limite de l'acceptable et nous rappelle que l'alpha dépend aussi du nombre d'items, une sous-échelle courte ayant en règle générale un alpha plus faible. Enfin, Cîrstea et collaborateurs ont traduit et validé la version roumaine du questionnaire PSPSQ 2.0 chez 220 patients atteints d'affections dermatologiques chroniques [11] : $\alpha = 0{,}978$ (sous-échelles 0,961-0,969), une structure expliquant 84,0% de la variance, CFI = TLI = 0,999, mais RMSEA = 0,109, valeur qui indique un certain défaut d'ajustement du modèle. Lorsque les indices d'ajustement se contredisent, les auteurs doivent discuter pourquoi ; un alpha de 0,978 suggère en outre une possible redondance de certains items.

13.5.6 Le recueil électronique des données

La saisie informatique des données recueillies sur papier est une source importante d'erreurs. Le recueil électronique (angl. electronic data capture, EDC) supprime cette étape et permet de vérifier les données au moment même de leur saisie. En travaux pratiques de la discipline, nous utilisons Google Forms, relié à une feuille Google Sheets dans laquelle les réponses sont rassemblées [46,147]. Pour les essais cliniques réels, on utilise des plateformes spécialisées, comme REDCap (répandue en milieu universitaire, avec journal d'audit et gestion des droits d'accès) ou Epi Info, le logiciel gratuit des CDC qui comporte à la fois un module de construction de formulaires et des modules d'analyse [51].

Un formulaire Google Forms bien construit utilise quatre fonctionnalités. La première est le type de question adéquat : "Réponse courte" pour les valeurs numériques, "Choix multiples" pour un choix unique, "Cases à cocher" pour un choix multiple, "Échelle linéaire" pour les items de Likert, "Liste déroulante" pour les listes longues (par exemple, le département), "Date" pour les dates. La deuxième est constituée par les règles de validation de la réponse [45] : pour l'âge, "Nombre", "Entre 18 et 99" ; pour la taille, "Entre 120 et 220" (cela aurait arrêté la valeur 1,68 de mg2026_brut) ; pour la glycémie, "Nombre", sans texte (cela aurait arrêté "112 mg/dl") ; chacune avec un message d'erreur explicite ("saisissez la taille en centimètres, par exemple 168"). La troisième est constituée par les questions obligatoires pour les variables indispensables (code du patient, bras, critère de jugement principal), mais pas pour toutes : un formulaire dans lequel tout est obligatoire contraint le répondant à inventer des réponses. La quatrième est constituée par les sections et les embranchements [148]. La première section est la page d'information et de consentement, avec la question "J'accepte de participer", la réponse "Je n'accepte pas" renvoyant directement à la fin du formulaire ; la question "Fumez-vous ?" avec la réponse "Oui" ouvre une section sur le nombre de cigarettes par jour, tandis que la réponse "Non" la fait sauter.

Une fois les réponses rassemblées, la feuille de calcul est téléchargée et vérifiée : une ligne par patient, une colonne par variable, des noms de colonnes courts, sans espaces ni signes diacritiques (glicemie0, et non "Glycémie à l'inclusion (mg/dL)"), des codes cohérents pour les catégories. Le tableau croisé dynamique de Google Sheets [44] est utile pour des vérifications rapides des effectifs.

13.5.7 Le dictionnaire des données et le test pilote

Le dictionnaire des données (angl. data dictionary ou codebook) est le tableau qui décrit chaque variable : le nom court, le libellé complet, le type, l'unité, les valeurs autorisées, les codes des catégories, le mode de codage des valeurs manquantes. Le tableau 13.2 reproduit un extrait du dictionnaire du jeu mg2026.

Tableau 13.2. Extrait du dictionnaire des données du jeu mg2026.

Variable	Description	Type	Valeurs autorisées / unité
id	code du patient	identifiant	P001-P160, unique
brat	bras de randomisation	nominale	A, B, C
sex	sexe	nominale dichotomique	F, M
varsta	âge à l'inclusion	numérique	ans, 18-99
educatie	niveau d'études	ordinale	1 collège, 2 lycée/enseignement professionnel, 3 universitaire, 4 postuniversitaire
inaltime	taille	numérique	cm, 120-220
glicemie0	glycémie à jeun à l'inclusion	numérique	mg/dL, 40-400
hba1c0	hémoglobine glyquée à l'inclusion	numérique	%, 3,5-15, avec une décimale
diabet_ogtt	diabète confirmé par HGPO	nominale dichotomique	0 non, 1 oui
satisfactie	satisfaction à 6 mois	ordinale	1-5 (Likert)

Avant le lancement de l'étude, le formulaire est testé auprès de quelques collègues et de quelques patients (test pilote). On chronomètre le remplissage, on vérifie si les questions sont comprises de la même façon par tous, si les embranchements fonctionnent, si les données arrivent correctement dans la feuille de calcul et si le fichier téléchargé peut être importé directement dans JASP ou MedCalc. Cinq formulaires de test remplis avec des valeurs extrêmes (âge 18 et 99, champs laissés vides, réponse "Je n'accepte pas") révèlent plus de problèmes qu'une heure de relecture.

13.6 La gestion et le nettoyage des données

Aussi bon que soit le formulaire, les données brutes contiennent des erreurs. Le fichier mg2026_brut.csv, déposé sur Moodle à côté de la version nettoyée, a été construit précisément pour cet exercice : il comporte 161 lignes au lieu de 160 et une série d'erreurs typiques. Le tableau 13.3 les présente toutes, avec la méthode qui permet de les détecter.

Tableau 13.3. Les erreurs du fichier mg2026_brut.csv et la méthode de détection.

Patient	Variable	Valeur dans le fichier brut	Type d'erreur	Comment la détecter
P005	sex	"f"	codage incohérent	tableau de fréquences : plus de 2 catégories apparaissent
P021	sex	"Feminin"	codage incohérent	idem
P034	tas0	1540	valeur impossible (un zéro de trop)	minimum/maximum, boîte à moustaches
P042	varsta	manquante	valeur manquante	nombre de valeurs valides inférieur à 160
P057	inaltime	1,72	unité erronée (m au lieu de cm)	minimum, histogramme, IMC recalculé aberrant
P064	glicemie0	"87 mg/dl"	texte dans une colonne numérique	la colonne n'est pas reconnue comme numérique
P084	fumat	"Fumator "	majuscule et espace	tableau de fréquences : une

Patient	Variable	Valeur dans le fichier brut	Type d'erreur	Comment la détecter
			superflu	catégorie de trop
P089	hba1c0	"5,2" stocké comme texte	séparateur décimal différent	type de la colonne, valeur mal alignée
P102	greutate0	7,26	virgule décimale déplacée	minimum, IMC recalculé aberrant
P121	mediu	"Urban"	majuscule	tableau de fréquences
P134	glicemie6	manquante	valeur manquante au suivi	nombre de valeurs valides
P091	ligne entière	doublon (lignes 91 et 161)	enregistrement en double	recherche des codes en double

La procédure de nettoyage suit un ordre logique. On conserve d'abord une copie intacte du fichier brut ; on ne travaille jamais sur l'original. On recherche ensuite les doublons d'après le code du patient. On vérifie le type de chaque colonne (numérique ou texte). On établit des tableaux de fréquences pour toutes les variables catégorielles et, pour les variables numériques, le minimum, le maximum, le nombre de valeurs valides et une boîte à moustaches. On vérifie les concordances logiques : IMC recalculé à partir du poids et de la taille, perte de poids compatible avec la variable scadere5, glycémie à 6 mois plausible par rapport à la glycémie initiale.

La correction est la partie délicate. Une erreur de forme ("Urban", "Fumator " avec une espace, "5,2" stocké comme texte) peut être corrigée sans hésitation si le sens est évident. Une valeur impossible ne se "devine" pas. La tentation est de croire que 1540 mmHg était 154, que 7,26 kg était 72,6 et que 1,72 m signifie 172 cm. Dans le fichier mg2026, ces suppositions se confirment, mais uniquement parce que nous avons vérifié chaque valeur dans le document source (le dossier du patient) et, pour la taille et le poids, recalculé l'IMC conservé dans le tableau : $\sqrt{\frac{77,7}{26,3}} = 1,72m$ et 27,0 x 1,64² = 72,6 kg. Dans des données réelles, un même chiffre déformé peut provenir de plusieurs erreurs différentes ; 1540 pouvait tout aussi bien être 150 ou 154, et une pression de 154 mmHg change le classement du patient comme hypertendu. La correction ne se fait qu'à partir de la source ; si la source n'est plus disponible, la valeur est déclarée manquante. Les vraies valeurs manquantes (l'âge de P042, la glycémie à 6 mois de P134) sont récupérées dans les documents sources si possible ; à défaut, elles restent manquantes et sont signalées.

Chaque modification est consignée dans un journal des modifications : la date, l'auteur, la ligne, la variable, l'ancienne valeur, la nouvelle valeur, le motif et le document source. Dans les plateformes professionnelles, cette traçabilité est tenue automatiquement et s'appelle piste d'audit (angl. audit trail) ; la ligne directrice des BPC exige que toute modification des données soit traçable et n'efface pas la valeur initiale [144]. Pour un projet étudiant, une feuille séparée dans le fichier Google Sheets, nommée "modificari", suffit. L'analyse statistique ne commence qu'une fois la base de données déclarée définitive ("gelée"), et toute modification ultérieure est justifiée par écrit.

13.7 La publication des résultats : CONSORT 2025

Les résultats d'un ECR sont rapportés selon la recommandation CONSORT, mise à jour en 2025 [56]. CONSORT est une liste de contrôle des éléments qui doivent figurer dans l'article (un titre précisant que l'étude est randomisée, l'enregistrement, l'accès au protocole, le schéma d'étude, les participants, les interventions, les critères de jugement, la taille de l'échantillon, la randomisation et la dissimulation de l'allocation, l'insu, les méthodes statistiques, le flux des participants, les caractéristiques initiales, les résultats avec la taille de l'effet et sa précision, les effets indésirables, les limites, la généralisabilité), accompagnée d'un diagramme de flux. Pour les études

observationnelles, la recommandation correspondante est STROBE [25], et pour les études d'exactitude diagnostique, STARD [140].

Le diagramme de flux CONSORT montre ce qu'il est advenu de chaque patient, de l'évaluation de l'éligibilité jusqu'à l'analyse. La figure 13.3 présente le diagramme de mg2026 : 212 personnes évaluées, 52 exclues (31 ne remplissaient pas les critères, 21 ont refusé), 160 randomisées (53 dans le bras A, 53 dans le bras B, 54 dans le bras C), toutes suivies à 3 et 6 mois et analysées en intention de traiter.

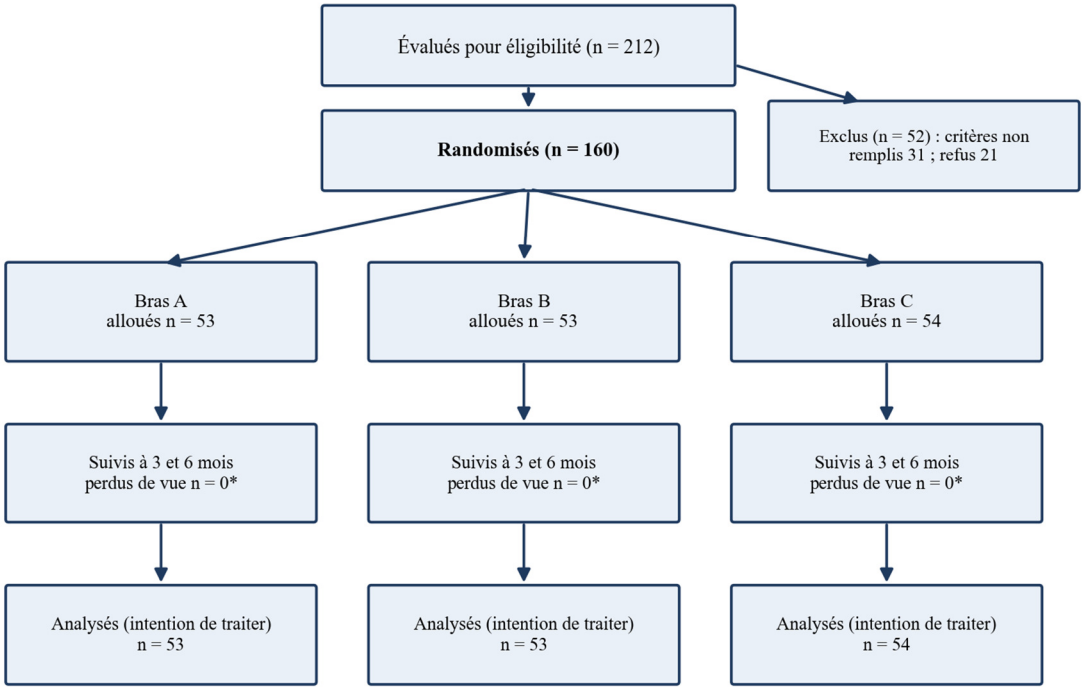


Figure 13.3. Diagramme de flux CONSORT de l'étude simulée mg2026 : 212 personnes évaluées pour l'éligibilité, 52 exclues (31 ne remplissent pas les critères, 21 refusent), 160 randomisées en trois bras (A 53, B 53, C 54), sans perdus de vue dans le jeu pédagogique, toutes analysées en intention de traiter.

Le diagramme répond à des questions que le lecteur doit se poser. Combien de patients éligibles ont refusé ? Si le refus a été massif, l'échantillon pourrait ne pas être représentatif : ceux qui acceptent une étude avec une application mobile sont peut-être plus jeunes et plus familiers de la technologie. Combien ont été perdus dans chaque bras et pour quels motifs ? Des pertes inégales entre les bras sont un signal d'alarme. Combien ont été analysés, par rapport au nombre de randomisés ? Si les nombres diffèrent, l'analyse n'est plus une analyse ITT. Dans une étude réelle, la case "perdus de vue" ne contient presque jamais zéro ; dans le jeu pédagogique, nous avons choisi cette simplification pour ne pas compliquer les exercices, et la note de la figure le dit explicitement.

13.8 Analyse intégrée de l'étude mg2026 : un exemple complet

Nous arrivons à la partie où tous les chapitres précédents travaillent ensemble. Nous parcourrons l'analyse de l'étude mg2026 dans l'ordre où elle apparaîtrait dans un article : le plan, la population, les caractéristiques initiales, le critère de jugement principal, les critères secondaires, les tailles d'effet, la conclusion et les limites, puis les analyses observationnelles que l'on peut mener sur les mêmes données. Tous les chiffres proviennent du jeu mg2026_curat.csv et peuvent être reproduits dans JASP 0.98.1 [47] ou MedCalc 24.0 [50], comme vous le ferez lors des travaux pratiques 13,

en suivant les étapes décrites également dans les recueils d'applications pratiques des disciplines de biostatistique [14,149].

13.8.1 Le plan d'analyse statistique, rédigé avant les données

Le plan d'analyse de l'étude, rédigé avant la levée des codes des bras, prévoit ce qui suit. La population d'analyse principale est la population ITT (les 160 patients randomisés, dans leur bras d'allocation). Le critère de jugement principal est la baisse de la glycémie à jeun à 6 mois (glicemie0 - glicemie6, en mg/dL), comparée entre les trois bras par une ANOVA à un facteur, après vérification de la normalité dans chaque bras (Shapiro-Wilk) et de l'homogénéité des variances (Levene) ; les comparaisons deux à deux se font avec le test de Tukey ; si les conditions ne sont pas remplies, on utilise le test de Kruskal-Wallis avec des comparaisons de Mann-Whitney corrigées. Le seuil de signification est $\alpha = 0{,}05$, bilatéral. Les critères secondaires sont : la proportion de patients ayant une perte de poids d'au moins 5% (khi-deux, risque relatif, différence de risques, NST, avec IC à 95%) ; la perte de poids en kg (ANOVA) ; la baisse de la PAS à 6 mois (ANOVA) ; la satisfaction à 6 mois (Kruskal-Wallis) ; l'évolution de la glycémie à 0, 3 et 6 mois (Friedman et ANOVA à mesures répétées). Les résultats secondaires sont interprétés comme des éléments de soutien, sans correction formelle, mais avec prudence face aux comparaisons multiples. Analyses observationnelles complémentaires, à l'inclusion : l'association obésité-diabète et la performance du score FINDRISC par rapport à l'HGPO.

Un tel paragraphe n'a rien de spectaculaire, mais c'est ce paragraphe qui donne leur crédibilité aux résultats. Remarquez que le test alternatif (Kruskal-Wallis) est également annoncé à l'avance, avec la règle selon laquelle il est choisi. Choisir le test après avoir vu lequel donne la plus petite valeur p est l'une des formes les plus fréquentes de "pêche" à la significativité.

13.8.2 Les caractéristiques initiales : le Tableau 1

Le premier tableau de tout article rapportant un ECR décrit les patients à l'inclusion, par bras. Les variables numériques de distribution approximativement normale sont décrites par la moyenne et l'écart-type, les variables asymétriques par la médiane et l'écart interquartile, les variables catégorielles par l'effectif absolu et le pourcentage (chapitre 2). Le tableau 13.4 est le "Tableau 1" de l'étude mg2026.

Tableau 13.4. Caractéristiques des patients à l'inclusion, par bras (étude mg2026, n = 160). Moyenne (ÉT), médiane [EIQ] ou n (%).

Caractéristique	Bras A (n = 53)	Bras B (n = 53)	Bras C (n = 54)	Total (n = 160)
Âge, ans	57,5 (9,2)	56,9 (9,4)	54,5 (7,5)	56,3 (8,8)
Sexe féminin	32 (60,4)	26 (49,1)	28 (51,9)	86 (53,8)
Milieu urbain	25 (47,2)	36 (67,9)	34 (63,0)	95 (59,4)
Études universitaires ou postuniversitaires	26 (49,1)	29 (54,7)	28 (51,9)	83 (51,9)
Fumeur actuel	18 (34,0)	18 (34,0)	17 (31,5)	53 (33,1)
Ancien fumeur	5 (9,4)	11 (20,8)	17 (31,5)	33 (20,6)
Non-fumeur	30 (56,6)	24 (45,3)	20 (37,0)	74 (46,3)
Activité physique (1-5)	3 [2-4]	3 [2-3]	3 [2-3]	3 [2-3]
IMC, kg/m²	29,2 [26,6-31,1]	27,9 [26,1-31,4]	27,7 [25,9-31,4]	28,25 [26,2-31,4]
Obésité (IMC ≥ 30)	18 (34,0)	20 (37,7)	16 (29,6)	54 (33,8)
Tour de taille, cm	89,4 (12,3)	87,8 (12,1)	88,4 (12,1)	88,5 (12,1)
PAS, mmHg	131,5 (13,8)	132,8 (13,8)	129,1 (14,0)	131,1 (13,9)
PAD, mmHg	80,5 (8,7)	80,1 (8,3)	77,1 (8,6)	79,2 (8,6)
Cholestérol total, mg/dL	209,3 (39,0)	202,8 (37,5)	215,9 (33,6)	209,4 (36,9)

Caractéristique	Bras A (n = 53)	Bras B (n = 53)	Bras C (n = 54)	Total (n = 160)
HDL-cholestérol, mg/dL	50,7 (11,5)	51,1 (12,4)	46,9 (11,2)	49,5 (11,8)
Triglycérides, mg/dL	143 [95-204]	144 [112-185]	166,5 [107-213]	151,5 [107-203,25]
Glycémie à jeun, mg/dL	107,8 (15,6)	106,3 (14,7)	102,3 (16,6)	105,4 (15,8)
HbA1c, %	5,61 (0,45)	5,57 (0,43)	5,55 (0,46)	5,58 (0,45)
Score FINDRISC	10,7 (4,4)	10,1 (4,2)	10,3 (4,2)	10,4 (4,3)
Diabète confirmé par HGPO	10 (18,9)	9 (17,0)	10 (18,5)	29 (18,1)

Le tableau ne comporte pas de colonne de valeurs p, et c'est voulu. CONSORT recommande de ne pas tester statistiquement les différences à l'inclusion dans un ECR [56]. Le raisonnement est simple : si la randomisation a été correcte, toute différence est, par définition, due au hasard, et un test statistique testerait une hypothèse dont nous savons déjà qu'elle est vraie. Avec 20 lignes testées à $\alpha = 0,05$, nous nous attendons, en moyenne, à un résultat "significatif" par simple hasard. Ce qui compte, c'est l'ampleur de la différence et l'importance pronostique de la variable, et non la valeur p.

Que lisons-nous néanmoins dans le tableau ? Les bras sont bien équilibrés pour l'âge, l'IMC, la pression artérielle, l'HbA1c, le FINDRISC et la prévalence du diabète (18,9%, 17,0% et 18,5%). Quelques différences sautent aux yeux : le bras C compte davantage d'anciens fumeurs (31,5% contre 9,4% dans le bras A), le bras A compte davantage de patients ruraux (52,8%), et la glycémie de départ est inférieure d'environ 5 mg/dL dans le bras C par rapport au bras A. Cette dernière différence compte pour l'interprétation : un patient qui part d'une glycémie plus élevée a, en principe, une plus grande marge de baisse. Le fait que le bras C, qui part plus bas, baisse malgré tout le plus rend le résultat plus convaincant, et non moins. Si nous considérions qu'une différence initiale peut influencer le résultat, une analyse de sensibilité avec ajustement (régression linéaire avec la glycémie initiale et le bras comme prédicteurs) devrait être prévue dans le plan, et non décidée après avoir vu le Tableau 1.

13.8.3 Le critère de jugement principal : la baisse de la glycémie à 6 mois

La description par bras vient toujours avant le test. La baisse moyenne de la glycémie a été de 3,0 mg/dL (ÉT 8,8) dans le bras A, de 11,1 mg/dL (ÉT 11,2) dans le bras B et de 16,3 mg/dL (ÉT 10,6) dans le bras C. Les intervalles de confiance à 95% de la moyenne de chaque bras, calculés avec la distribution t (chapitre 4), sont de 0,6-5,4 mg/dL pour A, 8,0-14,2 pour B et 13,4-19,2 pour C. Nous voyons déjà ici que l'intervalle du bras A ne chevauche aucun des deux autres.

Nous vérifions les conditions de l'ANOVA. Le test de Shapiro-Wilk appliqué à la baisse de la glycémie dans chaque bras donne $p = 0,575$ (A), $p = 0,052$ (B) et $p = 0,903$ (C) ; aucun ne rejette la normalité, et avec plus de 50 patients par bras l'ANOVA est de toute façon robuste à des écarts modérés. Le test de Levene donne $p = 0,163$; nous ne rejetons donc pas l'homogénéité des variances. L'ANOVA est par conséquent le test approprié, comme le prévoyait le plan.

Le résultat : $F(2 ; 157) = 22,93$, $p < 0,001$. Nous rejetons H0 selon laquelle la baisse moyenne de la glycémie est la même dans les trois bras. La taille de l'effet est $\eta^2 = 0,226$: environ 23% de la variabilité de la baisse de la glycémie est due au bras de traitement. L'oméga², une variante moins optimiste, calculée à partir des mêmes sommes de carrés, vaut 0,215. Les comparaisons de Tukey montrent que les trois paires diffèrent toutes significativement : B versus A, différence $8,1$ mg/dL$(\text{IC } 95\% \; 3,4 - 12,8;\ p = 0,0002)$; C versus A, 13,3 mg/dL (IC à 95% 8,6-18,0 ; $p < 0,001$) ; C versus B, $5,2$ mg/dL$(\text{IC } 95\% \; 0,6 - 9,9;\ p = 0,024)$. Le test de Kruskal-Wallis, prévu comme alternative, confirme le résultat ($H = 36,63$, $p < 0,001$). La partie gauche de la figure 13.4 présente graphiquement les différences de Tukey.

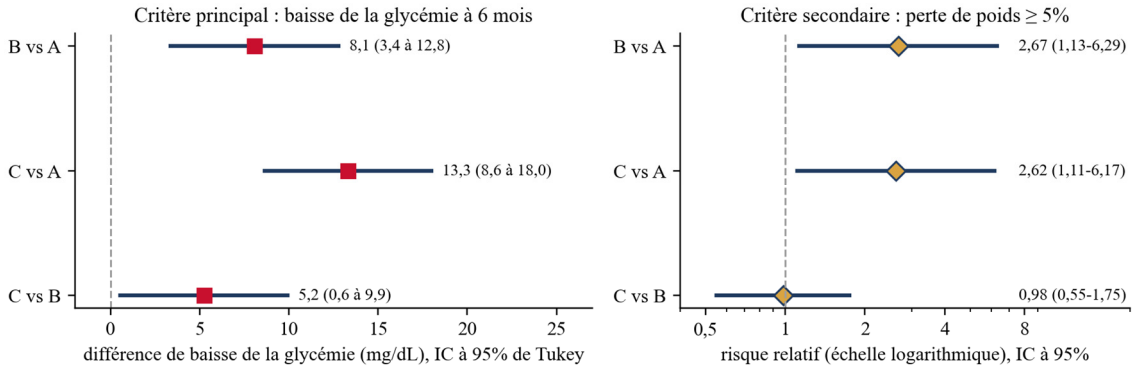


Figure 13.4. Effets des interventions dans l'étude mg2026. À gauche : différences entre les bras dans la baisse de la glycémie à 6 mois (mg/dL), avec les intervalles de confiance à 95% de Tukey ; aucun intervalle n'inclut 0. À droite : risque relatif d'obtenir une perte de poids d'au moins 5%, avec IC à 95%, sur échelle logarithmique ; les intervalles pour B et C versus A n'incluent pas 1, celui pour C versus B inclut 1.

Comment lire ces chiffres sur le plan clinique ? Une baisse supplémentaire de 13,3 mg/dL de la glycémie à jeun pour le bras C par rapport au conseil standard est considérable dans une population dont les valeurs moyennes de départ sont de 105 mg/dL, c'est-à-dire exactement dans la zone du prédiabète. La taille d'effet standardisée (d de Cohen, chapitre 6) est de d = 1,37 pour C versus A, de d = 0,80 pour B versus A et de d = 0,48 pour C versus B ; selon les repères usuels, les deux premiers effets sont grands et le troisième modéré [76]. La différence C versus B a toutefois un IC à 95% qui descend jusqu'à 0,6 mg/dL, c'est-à-dire très près de zéro : les données sont compatibles aussi bien avec un bénéfice supplémentaire substantiel de la téléconsultation qu'avec un bénéfice presque négligeable. C'est l'intervalle, et non la valeur p = 0,024, qui nous dit quelle est la précision de l'estimation.

13.8.4 Les critères secondaires

Perte de poids d'au moins 5%. Ont atteint ce seuil 6 patients sur 53 dans le bras A (11,3%), 16 sur 53 dans le bras B (30,2%) et 16 sur 54 dans le bras C (29,6%). Le test du khi-deux pour le tableau 3 x 2 donne $\chi^2 = 6{,}77$, ddl = 2, p = 0,034 ; le plus petit effectif attendu est de 12,6, la condition d'application (tous les effectifs attendus au moins égaux à 5) est donc remplie. Pour la comparaison C versus A, le risque relatif est de 2,62 (IC à 95% 1,11-6,17) : la probabilité de perdre au moins 5% du poids a été 2,6 fois plus élevée dans le bras C que dans le bras A. La différence absolue de risques est de 18,3 points de pourcentage (IC à 95% 3,4-33,2), et le nombre de sujets à traiter est $\text{NNT} = \frac{1}{0{,}183} = 5{,}5$, arrondi par excès à 6 (IC à 95% 3,0-29,1) : il faut inclure 6 patients dans le programme C, au lieu du conseil standard, pour qu'un patient de plus atteigne une perte de poids de 5%. Pour B versus A, RR = 2,67 (IC à 95% 1,13-6,29), avec une différence de risques de 18,9 points de pourcentage. Pour C versus B, RR = 0,98 (IC à 95% 0,55-1,75) : la téléconsultation n'a apporté aucun gain supplémentaire sur ce critère. La partie droite de la figure 13.4 montre ces trois rapports.

Observez la largeur des intervalles. Le NST peut se situer n'importe où entre 3 et 29, c'est-à-dire entre "très efficace" et "modeste". Avec seulement 22 événements dans les deux bras comparés, la précision est limitée, et un rapport honnête le dit.

Perte de poids en kilogrammes. A : $0{,}71 \pm 2{,}48$ kg ; B : $2{,}84 \pm 2{,}36$ kg ; C : $3{,}37 \pm 2{,}39$ kg ; ANOVA F = 18,25, p < 0,001. Le résultat continu et le résultat dichotomisé (≥ 5%) racontent la même histoire, et la version continue conserve davantage d'information. La dichotomisation d'une variable numérique fait perdre de la puissance statistique, raison pour laquelle le critère de jugement principal d'une étude est, de préférence, la variable continue.

Baisse de la pression artérielle systolique. Nous avons ici un exemple instructif. La baisse moyenne de la PAS à 6 mois a été de 2,8 mmHg (ÉT 9,0) dans le bras A, de 4,5 mmHg (ÉT 9,6) dans le bras B et de 7,0 mmHg (ÉT 8,2) dans le bras C. Le test prévu dans le plan, l'ANOVA, donne F = 2,94, p = 0,056. Le test de Kruskal-Wallis, appliqué aux mêmes données, donnerait p = 0,046. Que rapportons-nous ? Le test du plan : la différence n'atteint pas le seuil de signification statistique (p = 0,056). Changer de test après avoir vu le résultat, pour "passer" sous 0,05, serait exactement la pratique décrite plus haut comme pêche à la significativité. La formulation correcte est qu'il existe une tendance à une baisse plus importante de la PAS dans les bras avec application, mais que l'étude n'avait pas une puissance suffisante pour ce critère secondaire et que le résultat doit être confirmé. Sur l'ensemble des 160 patients, la PAS a baissé en moyenne de 4,8 mmHg (ÉT 9,0 ; t apparié = 6,68, p < 0,001), mais cette baisse globale ne peut pas être attribuée à l'intervention, car elle n'a pas de comparateur : elle inclut la régression vers la moyenne et l'effet de la participation à l'étude.

Satisfaction. Les médianes ont été de 3 (EIQ 2-3) dans le bras A, de 4 (EIQ 3-4) dans le bras B et de 4 (EIQ 3-5) dans le bras C ; Kruskal-Wallis H = 32,16, p < 0,001 ; comparaison C versus A par le test de Mann-Whitney : U = 2265,5, p < 0,001, avec une corrélation bisériale des rangs de 0,58 (effet de grande taille). Il s'agit d'un critère subjectif, rapporté par des patients qui savent quelle intervention ils ont reçue. La différence peut refléter aussi bien la valeur réelle du programme que le simple fait d'avoir reçu davantage d'attention.

Évolution de la glycémie dans le temps. Sur l'ensemble de l'échantillon, les moyennes de la glycémie ont été de 105,4, 99,4 et 95,2 mg/dL à 0, 3 et 6 mois (médianes 104,5, 99,0 et 92,5) ; le test de Friedman donne χ^2 = 74,67, p < 0,001, et l'ANOVA à mesures répétées F = 64,97, p < 0,001. Par bras, le test de Friedman est à la limite dans le bras A (χ^2 = 6,04, p = 0,049 ; moyennes 107,8, 105,7, 104,8 mg/dL) et fortement significatif dans le bras B (χ^2 = 21,07, p < 0,001 ; moyennes 106,3, 101,3, 95,2) et dans le bras C (χ^2 = 64,63, p < 0,001 ; moyennes 102,3, 91,2, 86,0). Attention à un piège : le fait que le test soit significatif dans un bras et "plus significatif" dans un autre ne prouve pas que les bras diffèrent entre eux. La question "la glycémie évolue-t-elle différemment selon les bras ?" trouve une réponse directe dans le critère de jugement principal (comparaison des baisses entre les bras) ou, de façon plus élégante, dans le terme d'interaction temps x bras d'une ANOVA mixte, méthode qui dépasse le niveau de ce cours.

13.8.5 Les comparaisons multiples et la hiérarchie des résultats

Nous avons réalisé un seul test principal et au moins cinq tests secondaires, plus les comparaisons post hoc. Si toutes les H0 étaient vraies, la probabilité qu'au moins un test sur dix sorte significatif par simple hasard serait d'environ 40%. L'étude mg2026 se protège par une hiérarchie : la conclusion repose sur le critère de jugement principal, choisi à l'avance et testé une seule fois, les comparaisons post hoc étant contrôlées par le test de Tukey ; les critères secondaires sont cohérents avec le critère principal et sont présentés comme des éléments de soutien. Si nous avions voulu qu'un critère secondaire soit concluant, nous aurions pu utiliser la correction de Bonferroni (le seuil de 0,05 divisé par le nombre de tests) ou une procédure de tests hiérarchiques préétablie, dans laquelle le deuxième critère n'est testé que si le premier a été significatif.

13.8.6 La conclusion et les limites

Un paragraphe de conclusion, rédigé dans le style d'un article, pourrait se lire ainsi. "Dans un essai randomisé pragmatique mené dans 12 cabinets de médecine générale, l'ajout d'une application d'automesure au conseil standard a augmenté la baisse de la glycémie à jeun à 6 mois de 8,1 mg/dL (IC à 95% 3,4-12,8), et l'ajout supplémentaire d'une téléconsultation mensuelle l'a augmentée de 13,3 mg/dL (IC à 95% 8,6-18,0) par rapport au conseil standard. La proportion de patients ayant perdu au moins 5% de leur poids a doublé dans les bras avec application (RR 2,62

pour le bras C versus A, IC à 95% 1,11-6,17 ; NST 6). Le bénéfice supplémentaire de la téléconsultation par rapport à l'application seule a été plus faible et estimé de façon imprécise."

Les limites doivent être énoncées tout aussi clairement. Les patients et les médecins n'étaient pas en insu, ce qui affecte surtout les critères subjectifs (la satisfaction). Le critère de jugement principal est un critère de substitution (la glycémie), et non un résultat clinique dur (incidence du diabète, événements cardiovasculaires). Le suivi de 6 mois ne dit rien du maintien de l'effet. La randomisation a été individuelle dans des cabinets où le médecin appliquait les trois stratégies, si bien qu'une contamination entre les bras est possible et aurait eu tendance à réduire les différences. Les données sont complètes, situation irréaliste pour une étude réelle. Enfin, les patients qui ont accepté de participer (160 sur 181 éligibles) peuvent être plus motivés que la population générale des patients à risque cardiométabolique.

13.8.7 Analyses observationnelles sur les mêmes données

Un jeu de données issu d'un ECR peut aussi servir à répondre à des questions observationnelles, à une condition que beaucoup oublient : la randomisation ne protège que la comparaison entre les bras. Toute autre association dans le même jeu (obésité et diabète, âge et pression artérielle) est observationnelle, transversale, exposée aux facteurs de confusion comme dans toute étude de ce type (chapitre 10).

À l'inclusion, la prévalence du diabète confirmé par HGPO a été de 33,3% chez les patients obèses (18 sur 54) et de 10,4% chez les non-obèses (11 sur 106). Le rapport de prévalences est de 3,21 (IC à 95% 1,64-6,31), et le rapport de cotes OR = 4,32 (IC à 95% 1,86-10,03) ; $\chi^2 = 12,70$, $p < 0,001$. S'agissant d'un schéma transversal, la mesure naturelle est le rapport de prévalences ; l'OR le surestime, comme on pouvait s'y attendre lorsque la maladie n'est pas rare. Dans la régression logistique ajustée sur l'âge et le sexe, l'OR de l'obésité devient 6,01 (IC à 95% 2,35-15,39 ; p = 0,0002), et chaque année d'âge augmente la cote de diabète de 11% (OR 1,11 ; IC à 95% 1,05-1,18). Nous ne pouvons pas dire, sur la base de ces données, que l'obésité "cause" le diabète : nous ne savons pas ce qui est apparu en premier, et d'autres facteurs non contrôlés peuvent expliquer l'association.

Deuxième question observationnelle : dans quelle mesure le score FINDRISC identifie-t-il les patients atteints d'un diabète confirmé par HGPO [125] ? L'aire sous la courbe ROC (chapitre 12) est de 0,852 (IC à 95% 0,762-0,943), proche de celle de la glycémie à jeun (0,844). Le seuil qui maximise l'indice de Youden est FINDRISC ≥ 11 (Se 93,1%, Sp 66,4%, J = 0,595). Au seuil ≥ 12, le tableau 2 x 2 comporte 22 vrais positifs, 33 faux positifs, 7 faux négatifs et 98 vrais négatifs : Se 75,9%, Sp 74,8%, VPP 40,0%, VPN 93,3%. Avec une prévalence de 18,1% dans l'échantillon, un score positif n'augmente la probabilité de diabète que jusqu'à 40%, mais un score négatif l'abaisse en dessous de 7%. Le FINDRISC fonctionne donc mieux pour exclure que pour confirmer, et l'HGPO reste nécessaire chez les sujets positifs. La figure 13.5 reprend la courbe ROC du chapitre 12.

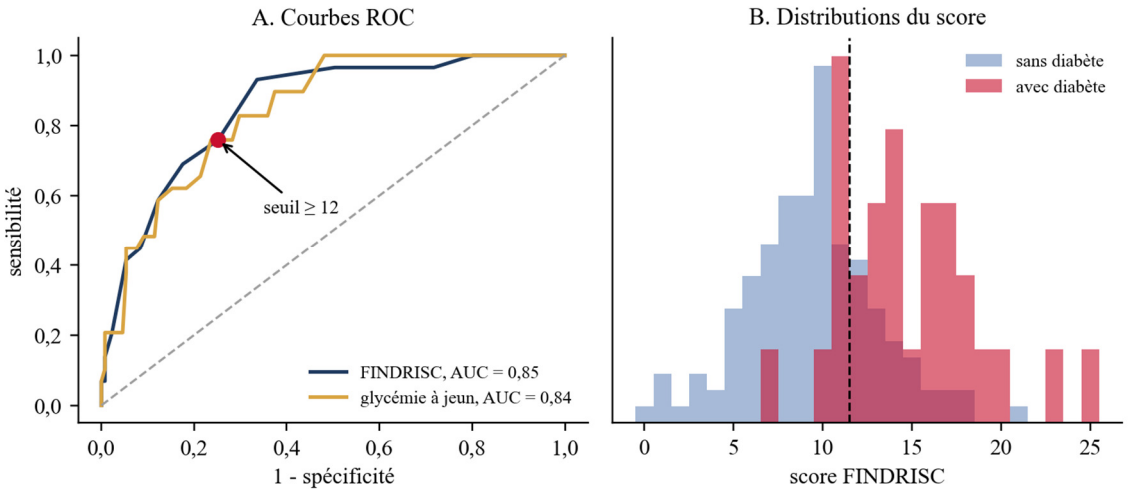


Figure 13.5. Courbe ROC du score FINDRISC pour le diabète confirmé par HGPO à l'inclusion dans l'étude mg2026 (aire sous la courbe 0,852, IC à 95% 0,762-0,943), analyse observationnelle transversale sur les données d'un essai randomisé.

Étude de cas : la confusion, les groupes parallèles non randomisés et le schéma avant-après

Christodorescu et collaborateurs ont suivi pendant 12 mois 278 patients hospitalisés pour une insuffisance cardiaque décompensée, afin de voir si le rapport hémoglobine/indice de distribution des globules rouges (HRR) prédit les événements [23]. La médiane du HRR était de 0,89, et le critère de jugement principal est survenu chez 167 patients (60,1%). Dans l'analyse de Kaplan-Meier, les patients ayant un HRR bas ont eu une survie sans événement plus faible (log-rank p = 0,027), mais dans le modèle de Cox ajusté sur l'âge, le sexe, le NT-proBNP, la fraction d'éjection et le débit de filtration glomérulaire, l'association n'était plus significative (p = 0,240). C'est l'exemple classique de confusion : un HRR bas accompagne les patients plus âgés, dont la fonction rénale est moins bonne et le NT-proBNP plus élevé, et ces patients ont de toute façon un pronostic plus réservé. Dans une cohorte, l'ajustement est la seule défense ; dans un ECR, la randomisation aurait équilibré ces facteurs. Les auteurs rapportent correctement les deux résultats.

Uivarosan et collaborateurs ont comparé, chez 70 patients ayant eu un accident vasculaire cérébral, deux groupes recevant une thérapie combinée (techniques modernes de rééducation et médication neurotrophique, n = 22 et n = 18) à un groupe contrôle recevant le traitement standard (n = 30), suivis pendant 12 mois [89]. Le résumé rapporte des résultats "significativement meilleurs" ($p < 0,05$) dans les groupes sous thérapie combinée et des améliorations à 12 mois par rapport à l'état initial ($p < 0,05$), sans préciser les tests. Le schéma à groupes parallèles avec contrôle est correct dans son principe, mais le lecteur doit découvrir dans le texte intégral comment les groupes ont été constitués (randomisation ou allocation selon un autre critère ?), pourquoi ils ont des effectifs différents et quel test a été appliqué. Pour trois groupes indépendants, l'ANOVA ou le test de Kruskal-Wallis seraient appropriés, et pour l'évolution dans le temps, un test pour mesures répétées.

Buda et collaborateurs ont mesuré la thrombospondine-1 sérique chez des patients hypertendus présentant une dysfonction endothéliale, avant et après un an de traitement par périndopril [86]. Les valeurs ont augmenté de 8061,9 ± 3684,80 à 15380 ± 5887 ng/mL($p < 0,001$), contre $9221,03 \pm 6510,21$ ng/mL chez les témoins non hypertendus, et un autre marqueur a diminué de 0,91 ± 0,51 à $0,50 \pm 0,24$ ng/mL($p < 0,001$). L'analyse appropriée est celle des données appariées (t apparié ou Wilcoxon, en notant que les écarts-types élevés par rapport aux moyennes suggèrent une asymétrie). Le groupe témoin de personnes saines sert de

> repère, et non de groupe contrôle du traitement : il n'a reçu ni périndopril ni placebo. C'est pourquoi le changement avant-après ne peut pas être attribué avec certitude au médicament, et le nombre de patients, non mentionné dans le résumé, est la première information que le lecteur doit rechercher.

13.9 Problèmes résolus

Problème résolu 13.1. Combien de patients faut-il inclure dans l'étude mg2026 ?

Énoncé. L'équipe qui conçoit l'étude mg2026 considère comme cliniquement pertinente une différence de 6 mg/dL dans la baisse de la glycémie entre un bras d'intervention et le bras contrôle, et, d'après des études antérieures, estime l'écart-type de la baisse à 10 mg/dL. Pour le critère secondaire, on s'attend à ce que 10% des patients du bras A et 30% de ceux du bras C perdent au moins 5% de leur poids. On travaille avec α = 0,05 (bilatéral) et une puissance de 80%. On estime à 15% la proportion de perdus de vue. Combien de patients faut-il inclure par bras ?

Solution. Pour le critère de jugement principal, nous utilisons la formule pour deux moyennes, avec $z\left(1-\frac{\alpha}{2}\right) = 1{,}96$ et z(1-β) = 0,84 (pour une puissance de 80%) :

$$n = \frac{2 \times (1{,}96 + 0{,}84)^2 \times \sigma^2}{\Delta^2} = \frac{2 \times 7{,}84 \times 100}{36} = 43{,}6$$

soit 44 patients par bras. La taille d'effet standardisée est $d = \frac{6}{10} = 0{,}6$. Si l'équipe avait été plus prudente et avait supposé d = 0,5, elle aurait obtenu 64 patients par bras (calcul exact avec la distribution t).

Pour le critère secondaire, p1 = 0,10, p2 = 0,30, $\bar{p} = 0{,}20$:

$\sqrt{2 \times 0{,}20 \times 0{,}80} = \sqrt{0{,}32} = 0{,}566;\ \ 1{,}96 \times 0{,}566 = 1{,}109\sqrt{0{,}10 \times 0{,}90 + 0{,}30 \times 0{,}70} = \sqrt{0{,}30} = 0{,}548;\ 0{,}84 \times 0{,}548 = 0{,}461n = \frac{(1{,}109+0{,}461)^2}{(0{,}30-0{,}10)^2} = \frac{1{,}570^2}{0{,}04} = \frac{2{,}464}{0{,}04} = 61{,}6$

soit 62 patients par bras. C'est plus que pour le critère principal : les critères binaires exigent, en règle générale, des échantillons plus grands.

Correction pour les pertes : si nous voulons que 62 patients restent évaluables dans chaque bras et que nous nous attendons à 15% de pertes, nous incluons 62/0,85 = 72,9, soit 73 patients par bras, environ 219 au total.

Interprétation. Avec 53-54 patients par bras, l'étude mg2026 a une puissance suffisante pour le critère principal (44 nécessaires), mais elle est légèrement en dessous du nécessaire pour le critère secondaire binaire. C'est exactement ce que l'on observe dans les résultats : le critère principal est très net, alors que les intervalles du RR sont larges (1,11-6,17). La formule ne tient pas compte du fait que nous avons trois bras et trois comparaisons ; une correction pour les comparaisons multiples augmenterait légèrement l'échantillon.

Problème résolu 13.2. Construction d'une liste de randomisation par blocs

Énoncé. Construisez les 12 premières allocations d'une liste de randomisation par blocs de 6 pour trois bras (A, B, C), en utilisant les nombres aléatoires suivants (générés, par exemple, avec la fonction =RAND() de Google Sheets) : bloc 1 : 0,71 ; 0,15 ; 0,88 ; 0,42 ; 0,07 ; 0,56 ; bloc 2 : 0,33 ; 0,91 ; 0,26 ; 0,64 ; 0,48 ; 0,12. Combien de séquences distinctes sont possibles dans un bloc ? Pourquoi une taille de bloc fixe est-elle dangereuse ?

Solution. Chaque bloc contient les lettres A, A, B, B, C, C. Nous associons à chaque lettre un nombre aléatoire, dans l'ordre, et nous trions par ordre croissant des nombres.

Bloc 1 : A 0,71 ; A 0,15 ; B 0,88 ; B 0,42 ; C 0,07 ; C 0,56. Par ordre croissant : C (0,07), A (0,15), B (0,42), C (0,56), A (0,71), B (0,88). La séquence : C A B C A B.

Bloc 2 : A 0,33 ; A 0,91 ; B 0,26 ; B 0,64 ; C 0,48 ; C 0,12. Ordonné : C (0,12), B (0,26), A (0,33), C (0,48), B (0,64), A (0,91). La séquence : C B A C B A.

Les 12 premiers patients : C, A, B, C, A, B, C, B, A, C, B, A. Après chaque groupe de 6 patients, chaque bras en compte exactement 2.

Le nombre de séquences distinctes dans un bloc est le nombre de permutations du multiensemble AABBCC : $\frac{6!}{2!\times 2!\times 2!} = \frac{720}{8} = 90$.

Le danger d'une taille fixe : si le médecin qui inclut les patients sait que les blocs comptent 6 places et a déjà vu 5 allocations du bloc, la sixième est entièrement prévisible ; après 4 allocations, il connaît déjà dans de nombreux cas les deux lettres restantes. Il peut alors influencer l'inclusion du patient suivant. Les solutions sont la dissimulation de l'allocation (le médecin ne voit pas la liste) et les blocs de taille variable (3, 6 ou 9, dans un ordre aléatoire). Pour une stratification par cabinet, on génère une telle liste pour chacun des 12 cabinets.

Problème résolu 13.3. Pourquoi ne testons-nous pas les différences du Tableau 1 ?

Énoncé. Un étudiant propose d'ajouter au tableau 13.4 une colonne de valeurs p pour chacune des 20 lignes. Il observe que, pour le tabagisme, le khi-deux donne p = 0,072, et pour le milieu de résidence p = 0,075, et conclut que "la randomisation a échoué de justesse". Que lui répondez-vous ?

Solution. Si la randomisation a été correcte, H0 "les bras proviennent de la même population" est vraie par construction pour toutes les caractéristiques initiales. Toute différence observée est une erreur d'échantillonnage. Un test statistique dans cette situation ne nous apprend rien de nouveau : il donnera un résultat "significatif" avec la probabilité α, c'est-à-dire dans environ 5% des cas. Avec 15 tests indépendants à α = 0,05, la probabilité qu'au moins l'un d'eux sorte significatif par simple hasard est

$$1 - 0{,}95^{15} = 1 - 0{,}463 = 0{,}537$$

soit plus de la moitié. Un p = 0,072 n'indique pas un "échec de justesse" de la randomisation ; il indique une fluctuation ordinaire. Ce qu'il faut juger, c'est si la différence (par exemple 31,5% d'anciens fumeurs dans le bras C contre 9,4% dans le bras A) est suffisamment grande et si la variable est suffisamment liée au résultat pour pouvoir influencer la conclusion. Si c'est le cas, l'analyse de sensibilité avec ajustement est prévue dans le plan. La conclusion correcte : le Tableau 1 se lit cliniquement, et non à l'aide de tests.

Problème résolu 13.4. ANOVA calculée à la main à partir des statistiques de groupe

Énoncé. En partant uniquement des statistiques descriptives de la baisse de la glycémie (A : n = 53, m = 3,02, s = 8,76 ; B : n = 53, m = 11,09, s = 11,21 ; C : n = 54, m = 16,33, s = 10,59), calculez F, l'êta² et le d de Cohen pour C versus A.

Solution. La moyenne générale, pondérée par la taille des groupes :

$$m = \frac{53 \times 3{,}02 + 53 \times 11{,}09 + 54 \times 16{,}33}{160} = \frac{160{,}06 + 587{,}77 + 881{,}82}{160} = \frac{1629{,}65}{160} = 10{,}19$$

La somme des carrés intergroupes :

$$\begin{aligned} \text{SSB} &= 53 \times (3{,}02 - 10{,}19)^2 + 53 \times (11{,}09 - 10{,}19)^2 + 54 \times (16{,}33 - 10{,}19)^2 \\ &= 53 \times 51{,}34 + 53 \times 0{,}82 + 54 \times 37{,}76 = 2721{,}1 + 43{,}4 + 2038{,}9 = 4803{,}4 \end{aligned}$$

La somme des carrés intragroupes, à partir des écarts-types :

$$\text{SSW} = 52 \times 8{,}76^2 + 52 \times 11{,}21^2 + 53 \times 10{,}59^2 = 3990{,}4 + 6534{,}5 + 5943{,}8 = 16468{,}7$$

Les degrés de liberté : intergroupes k - 1 = 2 ; intragroupes N - k = 157. Les carrés moyens : $\text{MSB} = \frac{4803{,}4}{2} = 2401{,}7$; $\text{MSW} = \frac{16468{,}7}{157} = 104{,}9$. Le rapport $F = \frac{2401{,}7}{104{,}9} = 22{,}90$. Le logiciel, qui travaille avec les valeurs non arrondies, donne 22,93 ; la différence provient uniquement des arrondis.

$\eta^2 = \frac{\text{SSB}}{\text{SSB+SSW}} = \frac{4803{,}4}{21272{,}1} = 0{,}226.$

Pour le d de Cohen C versus A, l'écart-type commun est

$$\text{sp} = \sqrt{\left[\frac{53 \times 10{,}59^2 + 52 \times 8{,}76^2}{54 + 53 - 2}\right]} = \sqrt{\frac{9934{,}2}{105}} = \sqrt{94{,}6} = 9{,}73$$

$$d = \frac{16{,}33 - 3{,}02}{9{,}73} = 1{,}37$$

Interprétation. La valeur critique de F pour 2 et 157 degrés de liberté à α = 0,05 est d'environ 3,05 ; F = 22,9 est bien au-dessus, donc p < 0,001. Le bras explique 22,6% de la variabilité de la baisse de la glycémie, et la différence C-A est de près de 1,4 écart-type, soit un effet de grande taille.

Problème résolu 13.5. Risque relatif, différence de risques et NST avec intervalles de confiance

Énoncé. Calculez le RR, la différence de risques et le NST, avec leurs IC à 95%, pour la perte de poids d'au moins 5% dans le bras C (16 sur 54) par rapport au bras A (6 sur 53).

Solution. Les risques : $p_C = \frac{16}{54} = 0{,}296$; $p_A = \frac{6}{53} = 0{,}113$. $\text{RR} = \frac{0{,}296}{0{,}113} = 2{,}62$.

L'intervalle du RR se calcule sur échelle logarithmique (chapitre 10) :

$\text{ES}(\ln \text{RR}) = \sqrt{\frac{1}{16} - \frac{1}{54} + \frac{1}{6} - \frac{1}{53}} = \sqrt{0{,}0625 - 0{,}0185 + 0{,}1667 - 0{,}0189} = \sqrt{0{,}1918} = 0{,}438 \ln 2{,}62 = 0{,}963$; $0{,}963 \pm 1{,}96 \times 0{,}438 = 0{,}963 \pm 0{,}858$, soit 0,105 à $1{,}821$; $\text{IC}\,95\%\,\text{RR} = e^{0{,}105}$ à $e^{1{,}821} = 1{,}11$ à 6,18 (le logiciel, avec les valeurs non arrondies, donne 1,11-6,17)

Différence de risques : 0,296 - 0,113 = 0,183 (18,3 points de pourcentage).

$\text{ES(DR)} = \sqrt{\frac{0{,}296 \times 0{,}704}{54} + \frac{0{,}113 \times 0{,}887}{53}} = \sqrt{0{,}00386 + 0{,}00189} = 0{,}0759;\ \text{IC}\,95\% = 0{,}183 \pm 1{,}96 \times 0{,}0759 = 0{,}183 \pm 0{,}149$, soit 0,034 à 0,332

$\text{NNT} = \frac{1}{0{,}183} = 5{,}5$, arrondi par excès à 6. L'intervalle du NST s'obtient en inversant les bornes de l'intervalle de la différence de risques : 1/0,332 = 3,0 et 1/0,034 = 29 (le logiciel, avec les valeurs non arrondies, donne 29,1).

Interprétation statistique. L'IC à 95% du RR n'inclut pas 1, et celui de la différence de risques n'inclut pas 0 ; la différence est donc significative à α = 0,05 (en accord avec le test exact de Fisher, p = 0,030). Interprétation clinique. Pour 6 patients traités dans le programme C au lieu du conseil standard, un patient de plus atteint une perte de poids de 5% ; l'estimation est toutefois imprécise, compatible avec un NST compris entre 3 et 29.

Problème résolu 13.6. L'alpha de Cronbach calculé à la main

Énoncé. Une échelle de 4 items de Likert (1-5) mesurant la confiance dans l'utilisation de l'application mobile a été appliquée à 40 patients des bras B et C. Les variances des items sont de

1,2 ; 1,0 ; 1,4 et 1,1, et la variance du score total (somme des 4 items) est de 14,2. Calculez l'α de Cronbach et interprétez-le.

Solution. Somme des variances des items : 1,2 + 1,0 + 1,4 + 1,1 = 4,7.

$$\alpha = \frac{4}{3} \times \left(1 - \frac{4{,}7}{14{,}2}\right) = 1{,}333 \times (1 - 0{,}331) = 1{,}333 \times 0{,}669 = 0{,}892$$

Interprétation. α = 0,89 indique une bonne cohérence interne : les items semblent mesurer le même construit. Pensez à ce qui se serait passé si les items n'avaient pas été corrélés : la variance du total aurait été approximativement égale à la somme des variances (4,7), et α aurait été proche de 0. Plus les items covarient, plus la variance du total dépasse la somme des variances et plus α se rapproche de 1. Un α élevé ne démontre cependant ni que l'échelle mesure la confiance et non autre chose (la validité), ni que les patients répondraient de la même façon deux semaines plus tard (la fidélité test-retest, mesurée par l'ICC).

Problème résolu 13.7. Intention de traiter versus analyse per protocole

Énoncé. Supposons, pour l'exercice, que dans le bras C de l'étude, 10 des 54 patients n'aient jamais utilisé l'application et ne se soient pas présentés aux téléconsultations. Parmi les 44 patients observants, 15 auraient atteint la perte de poids de 5%, et parmi les 10 non observants, 1. Dans le bras A, tous les patients auraient reçu le conseil (6 sur 53 avec une perte de poids de 5%). Calculez le RR dans l'analyse ITT et dans l'analyse per protocole, et commentez.

Solution. ITT : dans le bras C, les 54 patients randomisés restent tous, avec 15 + 1 = 16 événements ; $p_C = \frac{16}{54} = 29{,}6\%$; RR $= \frac{0{,}296}{0{,}113} = 2{,}62$. Per protocole : nous ne gardons que les 44 patients observants ; $p_C = \frac{15}{44} = 34{,}1\%$; RR $= \frac{0{,}341}{0{,}113} = 3{,}01$.

Commentaire. L'analyse per protocole donne un effet plus important, mais perd la comparabilité des bras. Les 10 non observants ne sont pas choisis au hasard : ils sont probablement moins motivés, moins familiers du smartphone, peut-être plus âgés. Dans le bras A, il existe certainement des patients semblables, mais nous ne les avons pas éliminés, car il n'y avait là aucune intervention à laquelle ne pas adhérer. Nous comparons donc "les patients motivés du bras C" à "tous les patients du bras A", et une partie de la différence tient à la motivation, et non à l'intervention. L'ITT répond à la question du médecin généraliste : "si je recommande ce programme à mes patients, avec leur observance réelle, qu'est-ce que je gagne ?". L'analyse per protocole approche l'effet chez le patient observant, mais avec le risque d'une erreur systématique. Lorsque les deux analyses vont dans le même sens, la conclusion est robuste.

Problème résolu 13.8. Ce que produit une seule valeur erronée

Énoncé. Dans le fichier brut, le patient P034 a la valeur tas0 = 1540 mmHg (la valeur réelle, d'après le dossier source, est 154). Calculez la moyenne, l'écart-type et la médiane de tas0 dans les deux versions et commentez.

Solution. Sur les données correctes : moyenne 131,13 mmHg, ÉT 13,88, médiane 131. Le remplacement de la valeur 154 par 1540 ajoute à la somme des valeurs 1540 - 154 = 1386, de sorte que la moyenne augmente de 1386/160 = 8,66 mmHg, pour atteindre 139,79 mmHg. L'écart-type explose à 112,24 mmHg, car l'écart quadratique de cette seule valeur par rapport à la moyenne domine toute la somme. La médiane reste 131 : la valeur extrême a fait passer un seul élément d'en dessous de la médiane au-dessus d'elle, ce qui peut déplacer la médiane d'une position au plus, et, autour du centre de ce jeu, plusieurs valeurs sont égales à 131.

Interprétation. Une seule erreur de frappe déplace la moyenne de plus de 8 mmHg et rend l'intervalle de confiance absurdement large ; tout test t appliqué à ces données serait compromis.

La médiane est robuste, exactement comme nous l'avons appris au chapitre 2. La leçon pour la gestion des données : le minimum, le maximum et une boîte à moustaches pour chaque variable numérique, avant toute analyse. Et encore une fois : la correction se fait à partir du dossier source, et non en devinant, même si ici la source confirme la valeur de 154 mmHg.

13.10 Exercices proposés

1. Une étude compare deux programmes d'exercice physique pour abaisser la PAS. La plus petite différence pertinente est de 8 mmHg, et l'écart-type estimé de la baisse est de 12 mmHg. Calculez le nombre de patients par groupe pour $\alpha = 0,05$ et une puissance de 90% ($z(1-\beta) = 1,28$). Que devient-il si l'on estime à 10% la proportion de pertes ?
2. Dans une étude de vaccination antigrippale chez les personnes âgées, la proportion attendue de patients présentant une infection respiratoire sévère est de 35% sans intervention et de 20% avec un programme de vaccination à domicile. Calculez l'échantillon par groupe pour $\alpha = 0,05$ et une puissance de 80%.
3. Un essai randomisé en grappes inclut des cabinets de 20 patients chacun. La corrélation intraclasse estimée est de 0,02. Calculez l'effet de plan et le nombre de patients par bras, sachant qu'un essai randomisé individuellement aurait nécessité 64 patients par bras.
4. Une échelle de 5 items a une somme des variances des items de 5,5 et une variance du score total de 18,0. Calculez l'α de Cronbach et interprétez-le.
5. Dans un ECR comptant 150 patients par bras, l'événement défavorable (hospitalisation dans les 12 mois) survient chez 30 patients traités par le nouveau programme et chez 45 patients du groupe contrôle. Calculez le risque relatif, la réduction absolue du risque et le NST.
6. Une étude comporte, outre le critère de jugement principal, 8 critères secondaires, testés chacun à $\alpha = 0,05$. Quelle est la probabilité d'obtenir au moins un résultat faux positif si toutes les H0 sont vraies et que les tests sont indépendants ? Quel seuil imposerait la correction de Bonferroni ?
7. Une étude de non-infériorité compare la téléconsultation aux visites en présentiel, avec une marge de non-infériorité $\Delta = 3$ mg/dL pour la baisse de la glycémie. Quelle conclusion tirez-vous si l'IC à 95% de la différence (téléconsultation moins visites en présentiel) est : a) de -2,4 à 1,8 mg/dL ; b) de -4,1 à 0,9 mg/dL ; c) de 0,5 à 4,2 mg/dL ?
8. Identifiez le schéma de chaque étude : a) 60 patients atteints de BPCO reçoivent, dans un ordre aléatoire, un nouveau bronchodilatateur et un bronchodilatateur standard, pendant 4 semaines chacun, avec une période d'élimination de 2 semaines entre les deux ; b) 24 écoles sont randomisées pour un programme d'éducation nutritionnelle ; c) 21 enfants reçoivent une nouvelle thérapie et sont évalués à 0, 30 et 90 jours ; d) 400 patients sont randomisés en quatre groupes : placebo, vitamine D, oméga-3, vitamine D et oméga-3 ; e) des patients ayant eu un infarctus sont suivis pendant 5 ans selon qu'ils ont reçu ou non des statines à la sortie de l'hôpital, sur décision de leur médecin traitant.
9. Dans la colonne "inaltime" (taille) d'un fichier brut apparaissent les valeurs : 168, 172, 1,75, 159, 1680, 165, "1 m 70", 181. Identifiez les valeurs suspectes, le type d'erreur et ce que vous faites de chacune.
10. Dans un essai de phase III portant sur 300 patients, aucun cas d'hépatotoxicité n'est survenu. En utilisant la règle de 3, estimez la borne supérieure de l'IC à 95% de la fréquence de cet effet indésirable. Qu'est-ce que cela signifie pour la phase IV ?

11. Dans le bras B de l'étude mg2026, 16 patients sur 53 ont perdu au moins 5% de leur poids. Calculez la proportion et son IC à 95% de Wald. Comparez-le à l'intervalle de Wilson.
12. Pour la comparaison B versus A dans l'étude mg2026 (16/53 contre 6/53), calculez le RR, la différence de risques et le NST. Comparez avec les résultats pour C versus A.

Pièges fréquents

- Confondre la dissimulation de l'allocation et l'insu : la première protège le moment de l'inclusion et est toujours possible, le second protège l'évaluation et n'est pas toujours possible.
- Tester statistiquement les différences initiales d'un ECR et déclarer une randomisation "ratée" à cause d'un $p < 0,05$ apparu par hasard.
- Changer de test ou de critère de jugement principal après avoir vu les données, pour obtenir $p < 0,05$ (par exemple, Kruskal-Wallis au lieu de l'ANOVA pour la PAS, $p = 0,046$ au lieu de 0,056).
- Interpréter un résultat non significatif d'une étude de supériorité comme une preuve d'équivalence.
- Éliminer les patients non observants et appeler l'analyse obtenue "intention de traiter".
- Remplacer les valeurs manquantes par la moyenne du groupe ou supprimer les patients ayant des données manquantes sans indiquer combien et pourquoi.
- Recueillir l'âge ou le poids en classes, au lieu de la valeur exacte.
- "Corriger" une valeur impossible en la devinant (1540 devient 154) sans vérification dans le document source.
- Traiter un seul item de Likert comme une variable numérique et en calculer la moyenne sans discussion.
- Considérer qu'un α de Cronbach élevé prouve la validité d'un questionnaire.
- Analyser un essai randomisé en grappes comme si les patients étaient indépendants.
- Lire les associations observationnelles d'un ECR (obésité et diabète) comme si elles étaient protégées par la randomisation.

À retenir

- L'étude interventionnelle attribue l'exposition ; la randomisation est la seule méthode qui équilibre aussi les facteurs de confusion inconnus.
- Le schéma d'étude détermine le test : groupes parallèles, tests pour échantillons indépendants ; schéma croisé ou avant-après, tests pour données appariées ; trois temps de mesure, Friedman ou ANOVA à mesures répétées ; grappes, méthodes qui tiennent compte de la corrélation intraclasse.
- Le critère de jugement principal, le test, le seuil et la règle de choix du test alternatif sont fixés dans le protocole et dans le registre public, avant les données.
- La taille de l'échantillon dépend de α, de la puissance, de l'effet minimal pertinent et de la variabilité ; on y ajoute une marge pour les pertes.
- L'analyse principale d'une étude de supériorité est l'analyse ITT ; l'analyse per protocole est une analyse de sensibilité.
- L'éthique (Helsinki 2024, consentement éclairé, comité d'éthique, BPC, Règlement 536/2014, RGPD) n'est pas une formalité, mais une condition de l'étude.
- Le type de question du questionnaire détermine le type de variable ; la fidélité (alpha, ICC) et la validité (contenu, construit, AFE/AFC) sont rapportées pour tout instrument nouveau ou traduit.
- Le nettoyage des données se fait sur une copie, avec un journal des modifications, et les corrections ne se font qu'à partir du document source.
- Un ECR est rapporté selon CONSORT 2025, avec le diagramme de flux, un Tableau 1 sans valeurs p et des effets exprimés par leur ampleur et leur intervalle de confiance.

Questions d'autoévaluation

1. (QRU) Laquelle des caractéristiques suivantes définit une étude interventionnelle ?
    a) le suivi des patients dans le temps
    b) l'existence d'un groupe contrôle
    c) l'attribution de l'exposition par l'investigateur, conformément au protocole
    d) le recueil prospectif des données
    e) l'utilisation d'un questionnaire validé
2. (QRU) Dans un essai croisé (cross-over) à deux traitements, la comparaison des valeurs obtenues par un même patient sous les deux traitements se fait, pour une variable numérique dont les différences suivent une loi normale, avec :
    a) le test t pour échantillons indépendants
    b) le test t pour échantillons appariés
    c) le test du khi-deux
    d) l'ANOVA à un facteur
    e) le test de Mann-Whitney
3. (QRU) La dissimulation de l'allocation (allocation concealment) a pour rôle :
    a) d'empêcher le patient de savoir quel traitement il reçoit
    b) d'empêcher l'évaluateur de connaître le bras du patient lors de la visite finale
    c) d'empêcher l'anticipation de l'allocation au moment de l'inclusion du patient
    d) d'équilibrer le nombre de patients entre les bras
    e) de réduire le nombre de patients perdus de vue
4. (QRU) Pour la perte de poids d'au moins 5% dans l'étude mg2026, la comparaison C versus A donne une différence absolue de risques de 18,3 points de pourcentage. Le NST est :
    a) 2
    b) 3
    c) 6
    d) 18
    e) 29
5. (QRU) Pour un questionnaire, le coefficient ICC calculé entre deux passations à 14 jours d'intervalle chez les mêmes patients évalue :
    a) la cohérence interne
    b) la fidélité test-retest
    c) la validité de contenu
    d) la validité convergente
    e) la structure factorielle
6. (QRM) À propos de l'analyse en intention de traiter (ITT) dans une étude de supériorité, sont vraies les affirmations suivantes :
    a) elle analyse chaque patient dans le bras dans lequel il a été randomisé
    b) elle préserve le bénéfice de la randomisation
    c) elle tend à surestimer l'effet biologique du traitement
    d) c'est l'analyse principale recommandée
    e) elle exclut les patients qui n'ont pas respecté le protocole
7. (QRM) Lesquels des éléments suivants sont des critères de substitution ?

a) l'HbA1c dans les essais d'antidiabétiques
b) le décès de cause cardiovasculaire
c) le cholestérol LDL dans les essais d'hypolipémiants
d) la pression artérielle dans les essais d'antihypertenseurs
e) l'accident vasculaire cérébral

8. (QRM) Dans le "Tableau 1" d'un essai randomisé :
   a) les variables numériques de distribution normale sont décrites par la moyenne et l'écart-type
   b) les variables asymétriques sont décrites par la médiane et l'écart interquartile
   c) il est recommandé de tester chaque différence entre les bras avec un test statistique
   d) les variables catégorielles sont décrites par l'effectif absolu et le pourcentage
   e) une différence importante sur une variable pronostique majeure peut justifier une analyse de sensibilité prévue dans le plan
9. (QRM) Dans le fichier mg2026_brut.csv, lesquelles des erreurs peuvent être détectées par un tableau de fréquences des variables catégorielles ?
   a) le sexe codé "f" et "Feminin"
   b) "Fumator " avec une espace finale
   c) "Urban" avec une majuscule
   d) $\mathrm{PAS}_0 = 1540$ mmHg
   e) la taille 1,72 saisie en mètres chez un patient
10. (QRM) À propos de la validation d'un questionnaire, sont vraies les affirmations suivantes :

a) un indice KMO de 0,93 indique des données très adaptées à l'analyse factorielle
b) un test de Bartlett non significatif est une condition nécessaire à l'analyse factorielle
c) l'analyse parallèle aide à choisir le nombre de facteurs
d) un α de Cronbach supérieur à 0,95 peut suggérer des items redondants
e) un α de Cronbach élevé démontre la validité de construit

Chapitre 14. Exemples récapitulatifs : interprétation intégrée des résultats statistiques

Objectifs

À la fin de ce chapitre, l'étudiant sera capable de :

- situer toute méthode apprise sur la carte de la discipline et expliquer le lien entre la question clinique, le schéma d'étude, le type de données, la description, l'estimation, le test, l'association et la décision ;
- lire les sections méthodes et résultats d'un article médical selon une méthode pas à pas ;
- reconnaître le test approprié à une question donnée et repérer les incohérences entre schéma d'étude, données et analyse ;
- interpréter ensemble la valeur p, la taille de l'effet et l'intervalle de confiance, en distinguant signification statistique et signification clinique ;
- recalculer à partir des chiffres publiés des indicateurs simples (intervalle de confiance d'une proportion, tableau 2 x 2, VPP à une autre prévalence, R^2 à partir de r) ;
- rédiger correctement un paragraphe de résultats et reconnaître les formulations erronées ;
- utiliser le tableau de synthèse des formules du cours et résoudre des QCM de type concours de résidanat qui combinent plusieurs chapitres.

14.1 La carte de la discipline

Quatorze chapitres semblent former quatorze sujets distincts. Ce n'est pas le cas. Tous répondent, sous des angles différents, à la même succession de questions que se pose tout chercheur et tout médecin qui lit un article : qu'est-ce que je veux savoir, comment ai-je obtenu les données, de quel type de données est-ce que je dispose, à quoi ressemblent-elles, avec quelle précision est-ce que j'estime ce qui m'intéresse, la différence ou l'association est-elle réelle, quelle est son ampleur, quelle décision clinique dois-je prendre. La figure 14.1 représente cette succession.

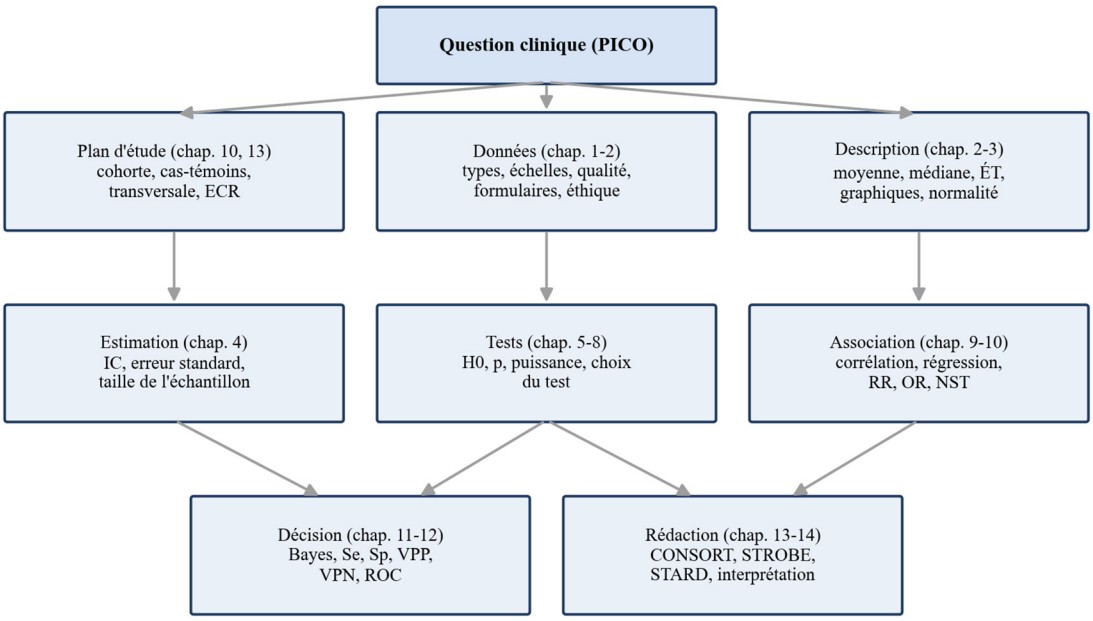


Figure 14.1. La carte de la discipline : de la question clinique (PICO) au schéma d'étude (chapitres 10 et 13), aux données (1-2), à la description (2-3), à l'estimation (4), aux tests (5-8), à l'association (9-10), à la décision diagnostique (11-12) et à la publication des résultats (13-14).

Tout part de la question clinique, formulée aussi précisément que possible, de préférence au format PICO [24]. La question détermine le schéma d'étude. Si je veux savoir si une intervention fonctionne, le schéma idéal est l'essai randomisé (chapitre 13). Si je veux savoir si une exposition est un facteur de risque, j'utilise une cohorte ou, pour les maladies rares, une étude cas-témoins (chapitre 10). Si je veux savoir quelle est la fréquence d'une maladie, une étude transversale. Si je veux juger un test diagnostique, une étude d'exactitude par rapport à un standard de référence (chapitre 12).

Le schéma d'étude et les outils de recueil produisent des données d'un certain type : numériques, ordinales, nominales (chapitre 2). Le type de données détermine la description : moyenne et écart-type pour les variables numériques symétriques, médiane et écart interquartile pour les variables asymétriques ou ordinales, effectifs et pourcentages pour les variables catégorielles. La forme de la distribution (chapitre 3) détermine si nous pouvons utiliser les méthodes fondées sur la loi normale ou si nous transformons les données ou passons aux méthodes non paramétriques.

À partir de l'échantillon, nous voulons dire quelque chose de la population. L'estimation (chapitre 4) fournit la valeur la plus probable et l'intervalle de confiance, c'est-à-dire la précision. Les tests d'hypothèses (chapitre 5) nous disent si une différence ou une association est compatible avec le hasard. Le choix du test (chapitres 6-8) dépend de trois éléments : le type de la variable résultat, le nombre de groupes et le caractère indépendant ou apparié des groupes. Les associations entre variables numériques se mesurent par la corrélation et la régression (chapitre 9), celles entre une exposition et une maladie par le risque relatif et le rapport de cotes (chapitre 10). Pour la décision au lit du malade, le théorème de Bayes (chapitre 11) et les indicateurs des tests diagnostiques (chapitre 12) traduisent le résultat d'un test en probabilité que le patient ait la maladie. Enfin, les résultats sont rapportés selon la recommandation adaptée au schéma d'étude : CONSORT pour les essais randomisés [56], STROBE pour les études observationnelles [25], STARD pour les études d'exactitude diagnostique [140].

Le tableau 14.1 condense cette carte sous une forme que vous pouvez utiliser à l'examen.

Tableau 14.1. De la question à la méthode : synthèse de la discipline.

Question	Schéma d'étude typique	Variable résultat	Mesure de l'effet	Test	Chapitre
Quelle est la fréquence de la maladie ?	transversale	dichotomique	prévalence avec IC à 95%	IC d'une proportion	4, 10
La moyenne diffère-t-elle entre deux groupes ?	transversale, cohorte, ECR	numérique	différence des moyennes, d de Cohen	t indépendant / Mann-Whitney	6, 7
La valeur a-t-elle changé après le traitement ?	avant-après, croisé	numérique	différence moyenne appariée	t apparié / Wilcoxon	6, 7
La moyenne diffère-t-elle entre 3 groupes ou plus ?	ECR à plusieurs bras, transversale	numérique	êta², différences de Tukey	ANOVA / Kruskal-Wallis	6, 7
La valeur change-t-elle entre 3 temps de mesure ?	mesures répétées	numérique ou ordinale	différences entre temps de mesure	ANOVA à mesures répétées / Friedman	6, 7
Deux variables catégorielles sont-elles associées ?	tout schéma	nominale	différence des proportions, RR, OR	khi-deux / Fisher	7, 10
Deux variables numériques varient-elles ensemble ?	transversale	numérique	r, rho, R^2, pente	Pearson / Spearman, régression	9
L'exposition est-elle un facteur de risque ?	cohorte	dichotomique	RR, DR, NST/NNH	khi-deux, IC du RR	10
L'exposition est-elle associée à la maladie (maladie rare) ?	cas-témoins	dichotomique	OR	khi-deux, IC de l'OR	10
Quelle est la qualité du test diagnostique ?	exactitude diagnostique	dichotomique	Se, Sp, VPP, VPN, RV, AUC	ROC, IC des proportions	11, 12
L'intervention fonctionne-t-elle ?	ECR	tout type	différence, RR, NST, avec IC	selon le type de variable	13

Deux remarques sur ce tableau. Premièrement, aucune colonne ne peut être lue indépendamment des autres ; le test dépend aussi du schéma d'étude (apparié ou indépendant), et pas seulement du type de variable. Deuxièmement, la colonne "mesure de l'effet" est au moins aussi importante que celle du test. Un article qui ne rapporte que des valeurs p, sans taille de l'effet ni intervalle de confiance, ne raconte que la moitié de l'histoire [70,74].

14.2 Comment lire la statistique d'un article médical

Lors des séminaires, les étudiants me disent souvent qu'ils lisent le résumé, sautent les méthodes et regardent directement la conclusion. C'est exactement l'ordre inverse de l'ordre correct. La conclusion est la partie la moins vérifiable d'un article ; ce sont les méthodes et les tableaux qui vous permettent de juger si elle est justifiée. Les manuels de base de la discipline traitent en détail de la lecture critique de la littérature médicale [18,19,64,150,151] ; nous la réduisons ici à une

méthode courte, à utiliser pour chaque article. Je propose une méthode en douze questions, que nous appliquerons ensuite dans les études de cas.

1. Le titre et la question. Quelle question l'étude pose-t-elle ? S'agit-il d'une question de fréquence, de différence, d'association, de pronostic, de diagnostic ou d'efficacité ? Peut-elle être formulée au format PICO ?
2. Le schéma d'étude. Qui a décidé de l'exposition ? S'agit-il d'une étude transversale, cas-témoins, de cohorte (prospective ou rétrospective), d'un ECR, d'une étude avant-après, d'une étude d'exactitude diagnostique ? L'étiquette du titre correspond-elle à ce qui a été effectivement fait ?
3. Population et échantillon. Qui a été inclus, d'où, comment les sujets ont-ils été sélectionnés, combien ont refusé ou ont été perdus de vue ? L'échantillon est-il représentatif des patients auxquels je veux appliquer le résultat ?
4. Variables. Quel est le critère de jugement principal et de quel type de variable s'agit-il ? Quelles sont les expositions ou les groupes ? Des variables numériques ont-elles été dichotomisées ?
5. Description. Les variables sont-elles correctement décrites (moyenne ± ÉT pour les distributions symétriques, médiane et EIQ pour les distributions asymétriques) ? Un écart-type presque égal ou supérieur à la moyenne d'une variable positive (durée d'hospitalisation, marqueur biologique) signale une asymétrie.
6. Tests. Les tests sont-ils adaptés au schéma d'étude et au type de données (indépendantes ou appariées, deux groupes ou plus, paramétrique ou non paramétrique) ? Les tests sont-ils nommés ?
7. Taille de l'effet. Quelle est l'ampleur de la différence ou de l'association : différence des moyennes, RR, OR, HR, r, AUC ? Est-elle cliniquement pertinente ?
8. Intervalle de confiance. Quelle est la précision de l'estimation ? Contient-il la valeur "nulle" (0 pour les différences, 1 pour les rapports) ? Est-il suffisamment étroit pour prendre une décision ?
9. Valeur p. Quelle hypothèse teste-t-elle ? Combien de tests ont été réalisés ? A-t-on tenu compte des comparaisons multiples ?
10. Confusion et biais. Un troisième facteur pourrait-il expliquer l'association ? L'analyse a-t-elle été ajustée ? Pour les études diagnostiques, y a-t-il eu un biais de spectre (malades sévères comparés à des volontaires sains) ?
11. Limites. Les auteurs les reconnaissent-ils ? Manque-t-il des limites évidentes (puissance, absence de groupe témoin, absence d'insu, données manquantes) ?
12. Recommandations de rédaction. CONSORT, STROBE ou STARD ont-ils été suivis ? Les éléments obligatoires (diagramme de flux, Tableau 1, IC de l'effet principal) sont-ils présents ?

La méthode ne demande pas de calculs sophistiqués. Elle demande de la patience et trois savoir-faire que vous avez exercés tout au long du semestre : reconnaître le type de variable, associer le schéma d'étude au test et lire un intervalle de confiance. Dans bien des cas, vous pouvez vérifier vous-mêmes un chiffre de l'article : à partir du pourcentage et de n, vous calculez l'intervalle de confiance d'une proportion ; à partir de Se, Sp et de la prévalence, vous reconstituez le tableau 2 x 2 ; à partir de r, vous calculez R^2. Les études de cas ci-dessous font exactement cet exercice.

Les études retenues sont issues des recherches de collègues de notre discipline et de leurs collaborations cliniques. Nous n'utilisons que les chiffres rapportés dans les résumés publiés. Lorsque nous signalons un problème méthodologique, nous le faisons à titre d'exercice de lecture critique : tout article a des limites, et les reconnaître est une forme de respect pour le travail des auteurs, non une critique à leur égard.

> **Étude de cas : la méthode des douze questions appliquée à une étude en hémodialyse**
>
> Marc et ses collaborateurs ont comparé deux cohortes de patients hémodialysés provenant de 10 centres de Roumanie, évaluées en 2015 (n = 1 401) et en 2019 (n = 1 698), avant et après le traitement national par antiviraux à action directe [39]. Parcourons rapidement la méthode. La question : la prévalence des infections par les virus des hépatites B et C a-t-elle changé ? Le

> schéma d'étude : transversal répété, deux échantillons indépendants, et non les mêmes patients suivis dans le temps. La description : âge moyen de 59,7 ± 12,92 ans en 2015 et de 61 ± 12,93 ans en 2019. Les tests : comparaison des moyennes (test t) et des proportions (khi-deux) entre les cohortes. Les résultats : l'âge diffère significativement (p = 0,005), bien que la différence ne soit que de 1,3 an, sans pertinence clinique ; avec plus de 1 400 patients dans chaque cohorte, même de petites différences deviennent significatives. La mortalité à un an (9,9% contre 10,7%, p = 0,46) et la prévalence du VHB (4,7% contre 4,8%, p = 0,604) ne diffèrent pas, tandis que la prévalence du VHC baisse de 16,9% à 10,5% (p < 0,0001), une différence également pertinente sur le plan clinique. La confusion : les cohortes diffèrent aussi par la durée de dialyse (p < 0,0001) et par la fréquence des calcifications vasculaires (p = 0,015), les populations ne sont donc pas identiques ; la baisse du VHC est compatible avec l'effet du traitement antiviral, mais le schéma d'étude ne permet pas de la lui attribuer avec certitude. Ce qui serait encore utile : les différences de prévalence avec leur IC à 95%. L'exercice montre que la méthode s'applique en quelques minutes et qu'elle distingue le résultat cliniquement important (VHC) de celui qui n'est que statistique (l'âge).

14.3 Études de cas intégrées

14.3.1 Hépatologie : un test diagnostique non invasif de la cirrhose

Lupușoru et ses collaborateurs ont évalué l'échographie de contraste avec imagerie paramétrique du temps d'arrivée (CEUS-PAT) pour le diagnostic non invasif de la cirrhose hépatique [2]. Ont été inclus 64 sujets : 37 patients atteints de cirrhose et 27 volontaires sains, d'âge moyen 58,98 ± 8,90 ans, dont 56,3% d'hommes. L'indice CEUS-PAT était de 0,83 ± 0,09 chez les patients cirrhotiques et de 0,49 ± 0,11 chez les volontaires sains (p < 0,0001) et il était corrélé à l'élastographie impulsionnelle (VCTE) avec r = 0,81. Au seuil > 0,7, l'aire sous la courbe ROC était de 0,98 (p < 0,001), avec Se = 89,19%, Sp = 100%, VPP = 100% et VPN = 87,1%.

À titre de comparaison, Moga et ses collaborateurs ont évalué l'élastographie time-harmonic chez 165 patients, avec la VCTE comme méthode de référence, et ont proposé deux seuils [3] : un seuil d'exclusion de la cirrhose (< 1,61 m/s), avec Se = 90,7%, Sp = 66,6%, VPP = 55,7%, VPN = 93,9%, et un seuil de confirmation (> 1,83 m/s), avec Se = 65,1%, Sp = 96,7%, VPP = 90,3%, VPN = 85,7%, tous deux avec une AUROC de 0,90 (IC à 95% 0,82-0,93) ; dans l'échantillon, 28,1% des patients avaient une cirrhose (F4). L'accord inter-examinateurs a été évalué par l'ICC (0,94 ; IC à 95% 0,89-0,97).

Questions d'interprétation, avec réponses types :

1. Quel est le schéma de l'étude CEUS-PAT et quel biais la menace ? Réponse type : il s'agit d'une étude d'exactitude diagnostique de type cas-témoins : on a comparé des patients dont la maladie était confirmée à des volontaires sains. Ce schéma expose au biais de spectre : la différence entre une cirrhose constituée et un foie sain est bien plus grande que la différence entre une cirrhose débutante et une fibrose avancée, qui est la situation réelle dans laquelle le médecin a besoin du test. Les performances (AUC 0,98, Sp 100%) sont, presque à coup sûr, meilleures que celles que l'on obtiendra chez des patients atteints d'hépatopathie chronique de sévérité variable.
2. Reconstituez le tableau 2 x 2 à partir des chiffres publiés. Réponse type : 37 malades, Se 89,19%, donc 33 vrais positifs et 4 faux négatifs (33/37 = 0,8919). 27 sujets sains, Sp 100%, donc 27 vrais négatifs et 0 faux positif. VPP $= \frac{33}{33} = 100\%$; VPN $= \frac{27}{27+4} = \frac{27}{31} =$ 87,1%, soit exactement la valeur rapportée. La reconstitution confirme la cohérence des chiffres.

3. Une spécificité de 100% signifie-t-elle que le test ne donne jamais de résultats faux positifs ? Réponse type : non. Elle signifie que, chez les 27 volontaires, aucun faux positif n'est apparu. La règle de 3 donne pour la borne supérieure de l'IC à 95% de la proportion de faux positifs environ 3/27 = 11%, de sorte que la spécificité réelle pourrait aussi se situer autour de 89%. Avec 27 sujets, "0 sur 27" est compatible avec une spécificité loin de la perfection.
4. La prévalence de la cirrhose dans l'échantillon est de 37/64 = 57,8%. Comment les valeurs prédictives changeraient-elles pour une prévalence de 20% ? Réponse type : en supposant une Se de 89,2% et une spécificité plus prudente de 89% (la borne ci-dessus), pour une prévalence de 20% la VPP serait d'environ 67% et la VPN d'environ 97% ; à la prévalence de l'échantillon, les mêmes Se et Sp donneraient une VPP de 92% et une VPN de 86%. Les valeurs prédictives dépendent de la prévalence (chapitres 11-12) : la VPP = 100% de l'article ne se transpose donc pas automatiquement à la population d'un cabinet de gastro-entérologie.
5. Quel test a probablement été utilisé pour comparer le CEUS-PAT entre les deux groupes et quelle information aurait encore été utile ? Réponse type : deux groupes indépendants, une variable numérique avec des écarts-types faibles par rapport aux moyennes : le test t pour échantillons indépendants (ou le test de Welch si les variances diffèrent). Il aurait été utile de disposer de la différence des moyennes avec son IC à 95% et des intervalles de confiance de l'AUC, de la Se et de la Sp.
6. Que signifie r = 0,81 avec la VCTE ? Réponse type : une forte corrélation positive ; $R^2 = 0{,}66$, c'est-à-dire qu'environ deux tiers de la variabilité d'une méthode sont "expliqués" linéairement par l'autre. Corrélation ne veut pas dire accord : deux méthodes peuvent être parfaitement corrélées et donner pourtant des valeurs systématiquement différentes. Pour évaluer l'accord, on utilise l'ICC ou la méthode de Bland-Altman [102].
7. Pourquoi l'étude sur l'élastographie time-harmonic propose-t-elle deux seuils différents ? Réponse type : un seuil unique ne peut pas avoir simultanément une sensibilité et une spécificité maximales. Le seuil d'exclusion a une Se et une VPN élevées (90,7% et 93,9%) : un résultat inférieur au seuil rend la cirrhose peu probable. Le seuil de confirmation a une Sp et une VPP élevées (96,7% et 90,3%) : un résultat supérieur au seuil rend la cirrhose très probable. Entre les deux seuils subsiste une "zone grise", dans laquelle d'autres examens sont nécessaires. Le RV+ au seuil de confirmation est de 0,651/(1 - 0,967) = 19,7, et le RV- au seuil d'exclusion de (1 - 0,907)/0,666 = 0,14.

14.3.2 Nutrition : habitudes alimentaires et poids chez les étudiants

Trois études menées par des collègues de la discipline ont suivi des étudiants de Timişoara au début de leur vie universitaire. Moleriu et ses collaborateurs ont analysé de manière transversale 921 étudiants âgés de 18 à 24 ans [4] : 24,1% étaient en surpoids ou obèses (19,4% en surpoids et 4,7% obèses). Pour deux des facteurs comportementaux analysés, le résumé rapporte des rapports de cotes ajustés de 7,20 (IC à 95% 4,26-12,17) et de 1,98 (IC à 95% 1,26-3,10) pour le surpoids. Un score composite de risque alimentaire (DRS, 0-4) a mis en évidence une relation dose-réponse : la proportion de surpoids est passée de 19,8% pour DRS = 0 à 50,0% pour DRS = 4, avec un OR ajusté de 1,43 par point (IC à 95% 1,23-1,66 ; p < 0,001).

Piroş et ses collaborateurs ont étudié de manière transversale 693 étudiants, sous l'angle du risque de prédiabète [6]. La prévalence du prédiabète était de 21,1% (celle du diabète de 1,4%). Selon les profils alimentaires obtenus par analyse en clusters, le prédiabète concernait 15,4% des étudiants à alimentation saine, 20,0% de ceux à alimentation mixte et 27,3% de ceux orientés vers

la restauration rapide (χ^2, p = 0,003), avec un OR de 1,78 par catégorie (IC à 95% 1,38-2,30 ; p < 0,001) et un p de tendance < 0,001.

Muntean et ses collaborateurs ont suivi de manière prospective une cohorte de 921 étudiants au cours de leur première année universitaire [5]. La différence moyenne d'IMC était de -0,45 ± 1,07 kg/m² (IC à 95% -0,52 à -0,38 ; p < 0,001) : 48,2% des étudiants ont vu leur IMC diminuer, 38,7% sont restés stables et 13,1% ont pris du poids ; une prise de poids cliniquement significative est survenue chez 7,2% des étudiants. L'un des comportements de chrononutrition était associé à une augmentation a IMC $\left(\beta = +\frac{0{,}22\ \text{kg}}{m^2};\ p = 0{,}004\right)$.

Questions d'interprétation, avec réponses types :

1. Calculez l'IC à 95% de la prévalence du surpoids (24,1% sur 921). Réponse type : ES $= \sqrt{\frac{0{,}241 \times 0{,}759}{921}} = 0{,}0141$; IC 95% $= 24{,}1\% \pm 1{,}96 \times 1{,}41\% = 21{,}3\%$ à 26,9%. Avec près de 1 000 étudiants, l'estimation est précise (±2,8 points de pourcentage).
2. Comment interprétez-vous un OR ajusté (aOR) = 7,20 (IC à 95% 4,26-12,17) ? Réponse type : après ajustement sur les autres facteurs du modèle, la cote de surpoids est environ 7 fois plus élevée chez les étudiants exposés au facteur en question. L'intervalle est large (de 4 à 12), mais il est loin de 1 : l'association est donc forte et significative. Attention : avec une prévalence du critère de jugement de 24%, l'OR surestime le rapport de prévalences ; un OR de 7 ne signifie pas que le surpoids est 7 fois plus fréquent.
3. Que signifie un OR = 1,43 par point de DRS et quelle est la valeur de l'OR entre DRS = 4 et DRS = 0 ? Réponse type : chaque point supplémentaire du score de risque augmente la cote de surpoids de 43%. L'effet se multiplie, il ne s'additionne pas : entre 4 points et 0 point, OR $= 1{,}43^4 = 4{,}18$ (en supposant une relation log-linéaire, comme l'impose le modèle logistique).
4. Les auteurs peuvent-ils conclure que l'alimentation axée sur la commodité est la cause de l'obésité ? Réponse type : non, pas sur la base d'un schéma transversal. L'exposition et le critère de jugement sont mesurés au même moment, nous ne savons donc pas lequel est apparu en premier (peut-être les étudiants en surpoids modifient-ils leur alimentation) ; il peut exister des facteurs de confusion non mesurés (sommeil, stress, revenu). La relation dose-réponse et la force de l'association plaident, selon les critères de Bradford Hill [108], en faveur d'une relation causale possible, mais ne la prouvent pas.
5. Vérifiez l'IC à 95% de la différence moyenne d'IMC de l'étude prospective. Réponse type : ES $= \frac{1{,}07}{\sqrt{921}} = 0{,}0353$; IC 95% $= -0{,}45 \pm 1{,}96 \times 0{,}0353 = -0{,}45 \pm 0{,}069$, c'est-à-dire -0,52 à -0,38, exactement ce que rapportent les auteurs. Le test approprié est le test t pour données appariées (le même étudiant, au début et à la fin de l'année).
6. Comment concilier une faible baisse moyenne (-0,45 kg/m²) avec le fait que 13,1% des étudiants ont pris du poids ? Réponse type : la moyenne décrit le centre de la distribution des variations, et non chaque étudiant. L'écart-type (1,07) est plus de deux fois supérieur à la variation moyenne : les variations individuelles sont donc très dispersées, dans les deux sens. C'est pourquoi les auteurs ont complété l'analyse linéaire de l'IMC par une analyse logistique de la prise de poids cliniquement significative : la même variable, considérée de façon continue puis dichotomisée, répond à des questions différentes.
7. Dans l'étude sur le prédiabète, pourquoi rapporte-t-on aussi un "p de tendance" ? Réponse type : les trois profils alimentaires suivent un ordre logique (sain, mixte, restauration rapide) et les proportions augmentent de façon monotone (15,4%, 20,0%, 27,3%). Le khi-deux habituel (p = 0,003) teste seulement si les proportions diffèrent, sans tenir compte de

l'ordre ; le test de tendance teste spécifiquement une augmentation monotone et possède, en règle générale, une plus grande puissance pour une telle hypothèse.

14.3.3 COVID-19 et diabète : trois études, trois types d'analyse

Cioca et ses collaborateurs ont comparé de manière rétrospective l'évolution de la COVID-19 chez 486 patients : 200 sans diabète, 62 atteints de diabète de type 1 et 224 de diabète de type 2 [8]. Les formes sévères représentaient 33,5% dans le diabète de type 2 contre 25,8% dans le type 1, la mortalité 11,6% contre 8,1%, la durée d'hospitalisation 11,6 ± 7,0 jours contre 9,1 ± 5,8 jours, la glycémie à l'admission 178,3 ± 34,7 contre 164,8 ± 39,6 mg/dL ; les rapports de cotes étaient de 2,24 pour l'admission en réanimation et de 2,46 pour la ventilation mécanique.

Albai et ses collaborateurs ont validé le score ISARIC-4C pour la sévérité et la mortalité chez 159 patients atteints de diabète de type 2 et de COVID-19 [75]. La mortalité était de 25,15%. L'aire sous la courbe ROC était de 0,875 (IC à 95% 0,820-0,930) ; le score a correctement classé 77,36% des patients, avec une Se de 87,80% (IC à 95% 87,66-87,93), une Sp de 73,72% (IC à 95% 73,48-73,96), une VPP de 53,73% et une VPN de 94,56%.

Dănilă et ses collaborateurs ont comparé 168 patients atteints de COVID-19 et de diabète à 168 patients sans diabète, appariés sur le sexe, l'IMC et la sévérité [105]. Pour l'indice dNLR, au seuil de 2,685 : Se 70,312%, Sp 65,978%, AUC 0,624 ($p < 0,001$) ; pour l'ALRI, au seuil de 0,812 : Se 76,429%, Sp 69,541%, AUC 0,752. Pour l'indice SII à 7 jours, un rapport des risques instantanés (hazard ratio) de 2,62 (IC à 95% 1,29-5,04) a été rapporté.

Questions d'interprétation, avec réponses types :

1. Que nous apprend une durée d'hospitalisation de 11,6 ± 7,0 jours sur sa distribution, et comment aurait-elle dû être décrite ? Réponse type : l'écart-type représente 60% de la moyenne, et la moyenne moins deux écarts-types donne une valeur négative (11,6 - 14,0 = -2,4 jours), ce qui est impossible. La distribution est asymétrique à droite (quelques patients avec des hospitalisations très longues). La description appropriée est la médiane avec l'EIQ, et la comparaison entre groupes relève du test de Mann-Whitney (ou de Kruskal-Wallis pour trois groupes).
2. Que manque-t-il dans la présentation de l'OR = 2,24 pour l'admission en réanimation ? Réponse type : l'intervalle de confiance (et la valeur p). Sans IC, nous ne savons pas si l'association est estimée avec précision ou si elle est compatible avec une absence d'effet. Il n'est précisé ni le groupe de référence, ni si l'OR est brut ou ajusté.
3. Avec trois groupes et une variable numérique comme la glycémie à l'admission, quel test utiliseriez-vous ? Réponse type : une ANOVA à un facteur si les distributions sont approximativement normales et les variances homogènes (test de Levene), suivie de comparaisons post-hoc (Tukey) ; sinon, un test de Kruskal-Wallis avec des comparaisons de Mann-Whitney corrigées. La comparaison du seul type 1 au type 2 se fait par le test t pour échantillons indépendants ou par le test de Mann-Whitney.
4. L'intervalle de confiance rapporté pour la sensibilité du score ISARIC-4C (87,66-87,93) est-il plausible ? Réponse type : non. Avec une mortalité de 25,15% chez 159 patients, environ 40 patients sont décédés. L'intervalle de Wald pour une proportion de 87,8% sur 40 patients est $87{,}8\% \pm 1{,}96 \times \sqrt{\frac{0{,}878 \times 0{,}122}{40}} = 87{,}8\% \pm 10{,}1\%$, soit environ 77,7-97,9%. L'intervalle publié, large de 0,27 point de pourcentage, correspondrait à un échantillon de plusieurs dizaines de milliers de patients. Il a probablement été calculé autrement (par exemple, par des méthodes de rééchantillonnage mal rapportées). L'exercice montre pourquoi il vaut la peine de vérifier les intervalles.

5. Vérifiez la VPP rapportée pour l'ISARIC-4C à partir de la Se, de la Sp et de la prévalence. Réponse type : $\text{VPP} = \frac{0{,}2515 \times 0{,}878}{0{,}2515 \times 0{,}878 + 0{,}7485 \times 0{,}2628} = \frac{0{,}2208}{0{,}4175} = 0{,}529$, soit environ 53%, proche des 53,73% rapportés. $\text{VPN} = \frac{0{,}7485 \times 0{,}7372}{0{,}7485 \times 0{,}7372 + 0{,}2515 \times 0{,}122} = 0{,}947$, proche de 94,56%. Le score est surtout utile pour identifier les patients à faible risque (VPN élevée).
6. Comparez l'AUC de 0,624 (dNLR) à celles de 0,752 (ALRI) et de 0,875 (ISARIC-4C). Réponse type : une AUC de 0,624 indique une discrimination médiocre (0,5 correspondrait au hasard) ; même si elle diffère significativement de 0,5 ($p < 0{,}001$), le dNLR n'est pas utile en clinique comme test isolé. L'ALRI a une discrimination acceptable, l'ISARIC-4C une bonne discrimination. La signification statistique d'une AUC ne dit rien de son utilité.
7. Quel problème pose la présentation d'un rapport des risques instantanés dans une étude dite transversale, ainsi que celle de la Se avec trois décimales ? Réponse type : le rapport des risques instantanés est issu de l'analyse du temps jusqu'à l'événement et suppose un suivi des patients dans le temps, donc un schéma longitudinal ; un schéma purement transversal ne peut pas produire de HR. Le chiffre de 70,312% suggère une précision illusoire : avec 168 patients par groupe, la troisième décimale n'a pas de sens ; on rapporte 70,3%, voire 70%, accompagné de l'IC à 95%.

14.3.4 Médecine d'urgence : sexe du patient, délais de thrombolyse et biomarqueurs de l'AVC

Popa et ses collaborateurs ont analysé de manière rétrospective 213 patients (100 femmes, 113 hommes) pris en charge au service des urgences pour une suspicion d'infection bactérienne [9]. Les femmes étaient plus âgées (médiane 77,0 contre 67,0 ans, $p < 0{,}001$), et l'origine urinaire de l'infection était plus fréquente chez les femmes (51,0% contre 34,5%, $p = 0{,}015$; OR 2,14, IC à 95% 1,20-3,81, $p = 0{,}010$; p ajusté selon Benjamini-Hochberg = 0,039). Une évolution sévère est survenue chez 14 femmes sur 100 (14,0%) et 21 hommes sur 113 (18,6% ; OR 0,71, IC à 95% 0,34-1,49) ; la mortalité hospitalière a concerné 7 femmes (7,0%) et 9 hommes (8,0% ; OR 0,87, IC à 95% 0,31-2,43) ; après ajustement, OR 0,80 (IC à 95% 0,27-2,35 ; $p = 0{,}690$).

Dans le même service, une étude ayant le même premier auteur a suivi les patients victimes d'un AVC ischémique arrivés dans les 4,5 premières heures [62] : sur 355 patients initiaux, 258 étaient éligibles et 46 ont reçu une thrombolyse intraveineuse (taux de 17,83%). Dans les comparaisons entre groupes, le résumé rapporte une médiane de 100 minutes pour le délai entre le début des symptômes et l'arrivée aux urgences ($p < 0{,}001$) et de 15,5 minutes pour le délai entre l'arrivée et le scanner ($p = 0{,}009$), ainsi qu'une variable clinique dont les valeurs étaient de $9 \pm 4{,}94$ contre $13{,}85 \pm 2{,}41 (p < 0{,}001)$. Une troisième étude a évalué les biomarqueurs GFAP et UCH-L1 chez 85 patients ayant fait l'objet d'une alerte AVC (69 avec AVC, 16 sans ; 37 ischémiques, 32 hémorragiques) [61] : GFAP $142{,}91 \pm 102{,}19$ pg/mL contre $37{,}76 \pm 19{,}92$ pg/mL$(p < 0{,}001)$, UCH-L1 $1\,307{,}68 \pm 967{,}54$ contre $189{,}81 \pm 92{,}69$ pg/mL$(p < 0{,}001)$; la GFAP a distingué l'AVC hémorragique de l'AVC ischémique avec une exactitude de 94,2% au seuil de 77,15 pg/mL.

Questions d'interprétation, avec réponses types :

1. Vérifiez l'OR de l'évolution sévère (14/100 femmes, 21/113 hommes). Réponse type : les cotes sont de 14/86 = 0,163 chez les femmes et de 21/92 = 0,228 chez les hommes ; $\text{OR} = \frac{0{,}163}{0{,}228} = 0{,}71$, soit exactement la valeur rapportée. De même, pour la mortalité, (7/93)/(9/104) = 0,075/0,087 = 0,87.

2. Comment interprétez-vous un OR de 0,71 (IC à 95% 0,34-1,49) ? Réponse type : l'estimation ponctuelle suggère une cote d'évolution sévère plus faible chez les femmes, mais l'intervalle contient 1 et s'étend d'une réduction des deux tiers à une augmentation de près de 50%. Les données sont compatibles aussi bien avec un effet protecteur important qu'avec un risque accru. La conclusion correcte n'est pas "le sexe n'influence pas le pronostic", mais "l'étude n'est pas assez précise pour répondre" ; avec seulement 35 événements sévères, la puissance est faible.
3. Pourquoi rapporte-t-on un p ajusté selon Benjamini-Hochberg de 0,039 en plus du p = 0,015 ? Réponse type : les auteurs ont comparé les femmes et les hommes sur de nombreuses caractéristiques. La correction pour comparaisons multiples (ici, le contrôle du taux de fausses découvertes) augmente la valeur p afin de tenir compte du nombre de tests. Le fait que la différence concernant l'origine urinaire reste inférieure à 0,05 après correction la rend plus crédible.
4. Pourquoi les délais ont-ils été rapportés sous forme de médianes ? Réponse type : les délais jusqu'à un événement sont presque toujours asymétriques à droite (quelques patients arrivent très tard). La médiane n'est pas influencée par ces valeurs extrêmes. La comparaison de deux groupes indépendants se fait alors par le test de Mann-Whitney.
5. Calculez l'IC à 95% du taux de thrombolyse. Réponse type : $p = \frac{46}{258} = 0{,}178$; $\text{ES} = \sqrt{\frac{0{,}178 \times 0{,}822}{258}} = 0{,}0238$; $\text{IC}\,95\% = 17{,}8\% \pm 4{,}7\%$, soit de 13,2% à 22,5%.
6. Que suggèrent les écarts-types des biomarqueurs GFAP et UCH-L1 chez les patients victimes d'un AVC ? Réponse type : l'écart-type est presque aussi grand que la moyenne (102,19 contre 142,91 pour la GFAP ; 967,54 contre 1 307,68 pour l'UCH-L1). Pour une variable qui ne peut pas être négative, cela indique une distribution fortement asymétrique. Le test t peut être inadapté ; le test de Mann-Whitney ou la transformation logarithmique des valeurs (chapitre 3) conviennent mieux, et la description devrait inclure la médiane et l'EIQ.
7. Exactitude de 94,2% pour distinguer l'AVC hémorragique de l'AVC ischémique : que faudrait-il encore rapporter ? Réponse type : 94,2% de 69 patients correspondent à 65 patients correctement classés. L'exactitude seule ne suffit pas : il faut la Se et la Sp (dans quelle mesure les hémorragies et les ischémies sont identifiées séparément), avec leur IC à 95%, et le seuil de 77,15 pg/mL, dérivé sur le même échantillon, doit être validé sur un échantillon indépendant, car les seuils choisis de façon "optimale" sur ses propres données surestiment les performances.

14.3.5 Validation d'un questionnaire et analyse de scores

Nous avons discuté au chapitre 13 de la validation du questionnaire TAM sur la télésurveillance chez les patients cirrhotiques [10] et de la validation de la version roumaine du PSPSQ 2.0 en dermatologie [11]. Nous les reprenons ici avec une troisième étude, qui utilise un questionnaire déjà validé pour répondre à une question clinique. Frandeș et ses collaborateurs ont étudié l'engagement numérique dans la prise en charge du diabète chez 304 adultes diabétiques [7]. L'échelle utilisée avait un α de Cronbach de 0,89 ; 35,5% des participants avaient un engagement numérique élevé. Les scores ont été rapportés sous forme de médianes et de quartiles, par exemple 1,71 [1,71-2,39] contre 3,86 [3,86-4,43] entre les groupes, et une régression bêta multivariable, avec une régression linéaire comme analyse de sensibilité, a fourni des coefficients tels que β = -0,147 (IC à 95% -0,219 à -0,075 ; p < 0,001), β = 0,996 (IC à 95% 0,859-1,134 ; p < 0,001) et β = 0,069 (IC à 95% 0,006-0,132 ; p = 0,033).

Questions d'interprétation, avec réponses types :

1. Dans l'étude TAM, α = 0,934 et ω = 0,950. Que disent ces chiffres et que ne disent-ils pas ? Réponse type : ils disent que les items des échelles sont fortement intercorrélés (très bonne cohérence interne). Ils ne disent pas si l'échelle mesure l'acceptation de la télésurveillance (validité), ni si les patients répondraient de la même façon lors d'une deuxième passation (fidélité test-retest).
2. Un modèle d'AFC présente un CFI = 0,999 et un RMSEA = 0,109 (PSPSQ 2.0). Comment est-ce possible ? Réponse type : ces indices mesurent des aspects différents de l'ajustement. Le CFI compare le modèle à un modèle d'indépendance totale des items ; lorsque les items sont très corrélés entre eux (α = 0,978), presque n'importe quel modèle structuré paraît bien meilleur que l'indépendance, et le CFI est donc très élevé. Le RMSEA mesure l'erreur d'approximation par degré de liberté et pénalise les modèles ayant peu de degrés de liberté et des échantillons de taille modérée ; 0,109 dépasse le seuil usuel de 0,08. Rapporter les deux indices et discuter la contradiction est la bonne pratique.
3. Dans le modèle TAM, "le modèle a expliqué 69,5% de la variation de l'intention d'utilisation". À quelle notion du chapitre 9 ce chiffre se rattache-t-il ? Réponse type : au coefficient de détermination R². Dans une régression simple, R² est le carré de la corrélation ; ici, dans un modèle structurel à plusieurs prédicteurs, $R^2 = 0{,}695$ signifie que les prédicteurs expliquent ensemble environ 70% de la variabilité de l'intention. Les 30% restants relèvent de facteurs non inclus et de l'erreur de mesure.
4. Que signifie une médiane égale au premier quartile, comme dans 1,71 [1,71-2,39] ? Réponse type : toutes les valeurs situées entre le 25e et le 50e percentile, c'est-à-dire celles d'au moins un quart des participants de ce groupe, sont égales à 1,71, très probablement une valeur limite du score. La distribution présente un effet plancher et elle est fortement asymétrique. C'est exactement la situation dans laquelle la moyenne et l'écart-type décriraient mal les données, et le choix des auteurs de rapporter des médianes et d'utiliser une régression adaptée aux scores bornés (la régression bêta) est justifié.
5. Comment interprétez-vous β = 0,069 (IC à 95% 0,006-0,132 ; p = 0,033) ? Réponse type : l'association est positive et significative au seuil de 0,05, et l'intervalle ne contient pas 0. Sa borne inférieure, 0,006, est cependant très proche de 0 : l'effet pourrait être négligeable. En comparaison, β = -0,147 (IC -0,219 à -0,075) est estimé plus précisément et se situe loin de 0. Ce sont l'ampleur et la précision, et pas seulement p, qui distinguent ces deux résultats.
6. Pourquoi effectue-t-on une "analyse de sensibilité" par régression linéaire ? Réponse type : pour vérifier si la conclusion dépend du choix du modèle. Si la régression bêta et la régression linéaire donnent des coefficients de même sens et de même significativité, le résultat est robuste à ce choix méthodologique.

14.3.6 Insuffisance cardiaque : comparaisons entre trois groupes, corrélations et confusion

Stoichescu-Hogea et ses collaborateurs ont comparé le strain longitudinal global (GLS) dans trois groupes : 62 patients atteints d'insuffisance cardiaque à fraction d'éjection préservée (HFpEF), 46 patients hypertendus et 40 témoins du même âge [90]. Le GLS différait entre les trois groupes (p < 0,001) : -20,2 ± 1,4% chez les témoins, -18,4 ± 3,0% chez les hypertendus (p = 0,031 par rapport aux témoins) et -17,6 ± 2,3% chez les patients HFpEF (p < 0,001 par rapport aux témoins). Le GLS était corrélé à une variable clinique ordinale avec un rho de Spearman = 0,423 (p < 0,001), et le seuil optimal de -19,35% avait une AUC de 0,833 (p < 0,001).

Buleu et ses collaborateurs ont comparé 60 patients atteints d'insuffisance cardiaque ischémique à fraction d'éjection réduite (< 40%) à 60 témoins sains du même âge [87]. L'ostéoprotégérine (OPG) était de 4,7 ± 0,25 contre $1{,}3 \pm 0{,}67$ ng/mL, et la vitamine 25-OH-D3 de 20,49 ± 7,31 contre $37{,}09 \pm 4{,}59$ ng/mL ($p < 0{,}001$ dans les deux cas). L'OPG était corrélée au volume télédiastolique du ventricule gauche (r = 0,862), à la fraction d'éjection (r = -0,832), à la vitesse de l'onde de pouls (r = 0,833), à la vitamine D (r = -0,636) et au débit de filtration glomérulaire (r = -0,488), toutes avec $p < 0{,}001$.

Enfin, l'étude sur la HRR discutée au chapitre 13 [23] a montré une association avec le pronostic dans l'analyse de Kaplan-Meier (log-rank p = 0,027), qui a disparu après ajustement par le modèle de Cox sur l'âge, le sexe, le NT-proBNP, la fraction d'éjection et le débit de filtration (p = 0,240).

Questions d'interprétation, avec réponses types :

1. Quel test a probablement été utilisé pour comparer le GLS entre trois groupes et que faudrait-il vérifier ? Réponse type : une ANOVA à un facteur, avec des comparaisons post-hoc deux à deux. Les écarts-types diffèrent beaucoup (1,4 chez les témoins, 3,0 chez les hypertendus), de sorte que l'homogénéité des variances (test de Levene) est douteuse ; une variante robuste (ANOVA de Welch avec post-hoc de Games-Howell) ou le test de Kruskal-Wallis serait préférable si le test de Levene rejette l'homogénéité.
2. Pourquoi compare-t-on les groupes deux à deux avec un test post-hoc et non avec trois tests t ? Réponse type : trois tests t, chacun réalisé à $\alpha = 0{,}05$, portent la probabilité globale d'une erreur de type I à environ $1 - 0{,}95^3 = 14\%$. Les tests post-hoc (Tukey, Bonferroni) contrôlent cette erreur à l'échelle de l'ensemble des comparaisons.
3. Quelle part de la variabilité un rho = 0,423 explique-t-il et en quoi l'interprétation diffère-t-elle de celle du r = 0,862 de l'étude sur l'OPG ? Réponse type : $\text{rho}^2 \approx 0{,}18$ (environ 18% ; pour le coefficient de Spearman, l'interprétation du carré n'est qu'approximative) ; r = 0,862 donne $R^2 = 0{,}74$. La première est une corrélation modérée, la seconde une corrélation forte. Toutes deux sont significatives, mais la force de l'association diffère beaucoup.
4. L'OPG est fortement corrélée à la fraction d'éjection (r = -0,832). Pouvons-nous dire que l'OPG diminue la fraction d'éjection ? Réponse type : non. La corrélation mesure l'association linéaire, et non le sens de la causalité. Il est possible que la dysfonction ventriculaire augmente l'OPG, que les deux soient déterminées par un troisième facteur (l'athérosclérose, l'inflammation, la fonction rénale, elle-même corrélée à l'OPG) ou que la relation soit bidirectionnelle.
5. L'étude sur l'OPG a un schéma cas-témoins avec des témoins sains. Quelles en sont les implications pour les corrélations calculées sur l'ensemble de l'échantillon ? Réponse type : si les corrélations sont calculées sur les 120 sujets, elles combinent deux groupes très différents (OPG 4,7 contre 1,3). La différence entre les groupes peut à elle seule créer une forte corrélation dans l'échantillon combiné, même si, à l'intérieur de chaque groupe, la corrélation est faible (chapitre 9, effet des groupes distincts sur r). Il vaut la peine de vérifier si les corrélations ont été calculées uniquement chez les patients.
6. Quelle leçon commune tirer de l'étude sur la HRR et de celle sur l'OPG ? Réponse type : une association brute (log-rank p = 0,027 ; corrélations fortes) peut être produite ou amplifiée par des facteurs de confusion. L'ajustement multivariable (Cox, régression multiple) est l'étape qui sépare l'association propre au marqueur de celle qu'il hérite d'autres facteurs. Dans l'étude sur la HRR, après ajustement, l'association n'était plus significative.

14.3.7 Pharmacologie chez le sujet âgé : prescriptions potentiellement inappropriées

Baneu et ses collaborateurs ont analysé de manière transversale 1 498 prescriptions électroniques établies pour des personnes âgées vivant en milieu urbain, dans les départements de Timiș et d'Arad, selon les critères STOPP/START version 2 [38]. Parmi les prescriptions, 57% concernaient des femmes ; l'âge moyen des patients était de 74,1 ± 6,95 ans et le nombre moyen de médicaments par prescription de 4,7 ± 1,51. Parmi les classes médicamenteuses, le résumé mentionne les statines, avec 47,4%. La prévalence des médicaments potentiellement inappropriés était de 18,58%, et celle des omissions potentielles de prescription de 49,2%.

Comme terme de comparaison pour des données appariées en pharmacologie, Gherbon et ses collaborateurs ont suivi de manière rétrospective 200 patients atteints de diabète de type 2 pendant un an après l'instauration d'un inhibiteur du SGLT-2 [59] : la glycémie a baissé de 180,00 mg% (EIQ 154,50-207,00) à 130,00 mg% (EIQ 117,50-150,00), l'HbA1c de 8,40% (EIQ 7,98-9,15) à 7,30% (EIQ 6,90-7,95), p < 0,001 dans les deux cas, et le tour de taille de 107,05 (± 16,39) à $102{,}50(\pm 15{,}11)$ cm$(p = 0{,}042)$.

Questions d'interprétation, avec réponses types :

1. Quelle est l'unité d'analyse dans l'étude STOPP/START et pourquoi est-ce important ? Réponse type : la prescription, et non le patient. Si un patient a plusieurs prescriptions, les observations ne sont pas indépendantes, et des intervalles de confiance calculés comme pour 1 498 patients indépendants seront trop étroits. L'interprétation correcte est "18,58% des prescriptions", et pas nécessairement "18,58% des personnes âgées".
2. Calculez l'IC à 95% de la prévalence des médicaments potentiellement inappropriés. Réponse type : $\text{ES} = \sqrt{\frac{0{,}1858 \times 0{,}8142}{1498}} = 0{,}0100$; IC $95\% = 18{,}6\% \pm 2{,}0\%$, soit de 16,6% à 20,5% (sous réserve de l'indépendance évoquée à la question 1).
3. La description "4,7 ± 1,51 médicaments par prescription" est-elle appropriée ? Réponse type : le nombre de médicaments est une variable numérique discrète (0, 1, 2...). Avec une moyenne de 4,7 et un ÉT de 1,5, la distribution n'est probablement pas très asymétrique, de sorte que la moyenne et l'ÉT sont acceptables ; la médiane et l'EIQ compléteraient utilement la description. La moyenne moins deux ÉT (1,7) reste plausible.
4. De quel type d'étude s'agit-il et quelles conclusions permet-elle ? Réponse type : une étude descriptive transversale. Elle permet d'estimer des prévalences et de décrire les pratiques de prescription. Elle ne permet pas de conclure sur les conséquences cliniques des prescriptions inappropriées (hospitalisations, chutes), car elle ne suit pas les patients dans le temps.
5. Dans l'étude sur les SGLT-2, pourquoi rapporte-t-on des médianes avec l'EIQ et quel test convient ? Réponse type : des données appariées (le même patient avant et après un an) de distribution asymétrique se décrivent par des médianes et se comparent par le test de Wilcoxon pour échantillons appariés. Le résumé emploie le mot "mean" pour des valeurs qui sont en réalité des médianes ; à la lecture, les chiffres accompagnés de l'EIQ doivent être lus comme des médianes.
6. Absence de groupe témoin dans l'étude sur les SGLT-2 : quelles explications alternatives peut-on proposer pour la baisse de la glycémie ? Réponse type : la régression vers la moyenne (les patients commencent un nouveau traitement précisément lorsque leur glycémie est la plus élevée), d'autres modifications du traitement ou du mode de vie au cours de l'année, une meilleure observance chez ceux qui sont restés suivis. Le schéma avant-après montre un changement, mais ne permet pas de l'attribuer avec certitude au médicament.

14.3.8 Oncologie : la pandémie et le cancer colorectal, les mutations BRCA et la survie

Feier et ses collaborateurs ont comparé la prise en charge du cancer du côlon avant et pendant la pandémie de COVID-19, chez 867 patients opérés dans un hôpital universitaire sur quatre ans [55]. Pendant la pandémie, les interventions programmées ont significativement diminué ($p < 0,001$), les interventions en urgence ont augmenté ($p < 0,001$) et, au cours de la deuxième année de pandémie, la proportion de cas au stade IV a augmenté ($p = 0,045$). Une corrélation inverse a été rapportée entre l'âge du patient et la durée de l'intervention ($r = -0,088$, $p = 0,01$).

Petruşan et ses collaborateurs ont suivi de manière prospective, pendant 24 mois, 133 patientes atteintes d'un carcinome ovarien séreux de haut grade [117]. Des mutations BRCA pathogènes ont été identifiées chez 39,1% des patientes. Le résumé rapporte des différences de proportions entre groupes (50% contre 77,77%, $p = 0,001$; 32,69% contre 51,85%, $p = 0,017$). Dans l'analyse de Cox univariée, le statut BRCA était associé à la survie sans progression (HR 0,52, IC à 95% 0,27-0,99, $p = 0,048$), mais dans le modèle multivariable l'association n'était plus significative (HR 0,57, $p = 0,124$).

Questions d'interprétation, avec réponses types :

1. La corrélation $r = -0,088$ est significative ($p = 0,01$). Est-elle aussi importante ? Réponse type : $r^2 = 0,0077$, c'est-à-dire que l'âge explique moins de 1% de la variabilité de la durée de l'intervention. La significativité vient de la grande taille de l'échantillon (867 patients) : lorsque n est grand, même des corrélations négligeables deviennent significatives. C'est l'exemple classique de la différence entre signification statistique et signification clinique.
2. Quel test a probablement été utilisé pour comparer la proportion d'interventions en urgence entre les périodes ? Réponse type : le test du khi-deux pour tableaux de contingence (période x type d'intervention), ou le test de Fisher si certains effectifs attendus sont faibles. Pour une comparaison portant sur plus de deux périodes, un khi-deux sur un tableau k x 2.
3. Le résultat concernant le stade IV ($p = 0,045$) doit être lu avec prudence. Pourquoi ? Réponse type : il s'agit d'un test parmi beaucoup d'autres (plusieurs périodes, plusieurs variables), et la valeur est proche du seuil. Sans correction pour comparaisons multiples, un tel résultat peut être un faux positif. L'ampleur de la différence et l'IC aideraient à l'interpréter.
4. Comment lire un HR = 0,52 (IC à 95% 0,27-0,99) ? Réponse type : à tout moment du suivi, le taux instantané de progression chez les patientes porteuses d'une mutation BRCA est environ la moitié de celui des patientes sans mutation, soit une réduction relative de 48%. La borne supérieure de l'intervalle, 0,99, est cependant très proche de 1 : les données sont aussi compatibles avec une réduction de seulement 1%. Et $p = 0,048$ est à la limite.
5. Pourquoi la significativité disparaît-elle dans le modèle multivariable (HR 0,57, $p = 0,124$) ? Réponse type : deux explications, qui ne s'excluent pas. La première est la confusion : une partie de l'association brute est due à d'autres facteurs (le stade, le type de chirurgie, la chimiothérapie) liés à la fois au statut BRCA et au pronostic. La seconde est la précision : le HR ajusté (0,57) est très proche du HR brut (0,52), l'effet n'a donc pas disparu, mais l'ajout des covariables a élargi l'intervalle de confiance. Avec 133 patientes, le modèle multivariable a une puissance limitée. La formulation correcte est la suivante : l'association ne reste pas significative après ajustement, et son ampleur demeure incertaine.
6. Quelle méthode descriptive et quel test correspondent à l'analyse de la survie sans progression ? Réponse type : les courbes de Kaplan-Meier pour la description, le test du

log-rank pour comparer deux courbes et le modèle de Cox pour estimer le HR avec ajustement. Toutes ces méthodes tiennent compte des patientes censurées (qui n'avaient pas progressé à la fin du suivi ou qui sont sorties de l'étude), ce que ne fait pas une simple comparaison de proportions.

14.3.9 Une étude de type ECR : Aloe ferox et une application pour l'asthme

Nous avons présenté au chapitre 13 l'étude contrôlée contre placebo portant sur Aloe ferox chez des personnes obèses, rapportée selon CONSORT [85], et l'étude pilote contrôlée de l'application mobile pour l'asthme [94]. Nous les utilisons maintenant pour récapituler les principes des essais cliniques. Rappelons les données principales : dans la première étude, 20 personnes obèses ont reçu un régime et de l'Aloe ferox (2 gélules par jour, 460 mg), et 20 témoins appariés ont reçu un régime et un placebo, pendant 3 mois ; le résumé rapporte des valeurs p pour l'IMC (0,03 et 0,002), le cholestérol total (0,032), le LDL (0,01 et 0,039), la glycémie à jeun (0,018 et < 0,001) et la pression diastolique (0,006). Dans la seconde, 93 patients asthmatiques ont été comparés par le test de Mann-Whitney ($p < 0,001$), par le test de Friedman avec correction de Bonferroni ($p < 0,001$) et par le khi-deux pour la proportion d'exacerbations (10,30% contre 46,30%, $\chi^2(1) = 13,707$, $p < 0,001$).

Questions d'interprétation, avec réponses types :

1. Quels éléments de CONSORT devriez-vous trouver dans l'article sur Aloe ferox ? Réponse type : le mode de génération de la séquence de randomisation, la dissimulation de l'allocation, qui était en insu (les patients, les évaluateurs), la définition du critère de jugement principal, le calcul de la taille de l'échantillon, le diagramme de flux avec les perdus de vue, le Tableau 1 par groupe, l'effet principal avec son ampleur et son IC à 95%, les effets indésirables et l'enregistrement de l'essai [56].
2. Que signifie "témoins appariés" dans une étude contre placebo et quelle influence cela a-t-il sur l'analyse ? Réponse type : les témoins ont été choisis pour ressembler aux patients traités sur certaines caractéristiques (âge, sexe, IMC). Si l'allocation s'est faite par appariement et non par randomisation, la comparabilité n'est garantie que pour les variables d'appariement. Si l'appariement s'est fait par paires, l'analyse devrait tenir compte de cette structure appariée.
3. Pourquoi deux séries de valeurs p pour les mêmes variables (IMC 0,03 et 0,002) ? Réponse type : une série correspond probablement à la comparaison entre groupes, et l'autre à la comparaison avant-après au sein d'un groupe. Seule la comparaison entre groupes à 3 mois (ou celle des variations entre groupes) teste l'effet d'Aloe ferox, puisque le régime était commun aux deux groupes. Une baisse significative dans le groupe traité, sans comparaison avec le placebo, peut être l'effet du régime.
4. Avec 20 patients par groupe, quelle taille d'effet l'étude pouvait-elle détecter avec une puissance de 80% ? Réponse type : d'après la formule pour deux moyennes, $d = (1,96 + 0,84) \times \sqrt{\frac{2}{20}} = 2,8 \times 0,316 = 0,89$. L'étude n'avait une puissance suffisante que pour des effets importants (près d'un écart-type). Des effets modérés, cliniquement pertinents, pouvaient passer inaperçus.
5. Calculez le risque relatif d'exacerbation dans l'étude sur l'application, si 10,30% est la proportion observée dans le groupe utilisant l'application. Réponse type : $\text{RR} = \frac{10,30}{46,30} = 0,22$, soit une réduction relative du risque de 78% et une réduction absolue de 36 points de pourcentage (NST ≈ 3). L'interprétation doit cependant s'accompagner de l'IC à 95% et

d'une vérification du mode d'allocation : dans une étude pilote non randomisée, les patients qui acceptent l'application peuvent être plus observants dès le départ.

6. Pourquoi la correction de Bonferroni était-elle nécessaire pour le test de Friedman ? Réponse type : le test de Friedman global indique seulement qu'au moins un temps de mesure diffère des autres. Pour savoir lequel, on effectue des comparaisons entre paires de temps de mesure (par exemple, par le test de Wilcoxon) ; avec plusieurs comparaisons, le seuil de chacune est divisé par leur nombre, afin de maintenir l'erreur de type I globale à 0,05.

14.3.10 Réanalyse complète du jeu de données mg2026

Nous terminons les études de cas avec le jeu de données que vous avez utilisé tout au long du semestre. Les questions ci-dessous parcourent l'ensemble de la discipline sur les mêmes données ; les réponses reprennent les chiffres calculés dans les chapitres précédents, afin que vous voyiez comment ils s'articulent.

1. Décrivez l'âge, l'IMC et les triglycérides en justifiant le choix des indicateurs. Réponse type : l'âge a une distribution symétrique (asymétrie 0,01 ; Shapiro-Wilk p = 0,687) : moyenne 56,29 ans, ÉT 8,77. L'IMC est asymétrique à droite (asymétrie 1,42 ; aplatissement 3,68 ; Shapiro-Wilk W = 0,904, p < 0,001) : médiane 28,25 kg/m², EIQ 26,2-31,4. Les triglycérides sont asymétriques (asymétrie 0,95 ; p < 0,001) : médiane 151,5 mg/dL, EIQ 107-203,25 ; après transformation logarithmique, la distribution devient presque normale (asymétrie -0,28 ; p = 0,200), et la moyenne géométrique est de 144,0 mg/dL (IC à 95% 132,95-156,02).
2. Estimez la moyenne de la PAS dans la population et interprétez-la. Réponse type : moyenne 131,13 mmHg, ÉT 13,88, ES 1,097 ; IC à 95% 128,96-133,30 mmHg. Avec une confiance de 95%, la moyenne de la PAS de la population dont provient l'échantillon se situe entre ces bornes. L'intervalle des valeurs individuelles est autre chose : environ 95% des patients ont une PAS comprise entre 103,9 et 158,3 mmHg.
3. La PAS diffère-t-elle entre les femmes et les hommes ? Réponse type : femmes 131,37 (ÉT 14,47 ; n 86), hommes 130,85 (ÉT 13,25 ; n 74) ; Levene p = 0,529 ; test t de Welch t = -0,24, ddl 157,4, p = 0,81 ; différence M-F -0,5 mmHg (IC à 95% -4,9 à 3,8) ; d = -0,04. Nous ne rejetons pas H0. L'intervalle montre qu'une différence supérieure à environ 5 mmHg, dans un sens ou dans l'autre, est peu probable, une information plus utile que le simple "non significatif".
4. La glycémie a-t-elle baissé en 6 mois ? Réponse type : de 105,4 à 95,2 mg/dL ; différence moyenne 10,19 mg/dL (ÉT 11,57 ; IC à 95% 8,38-11,99) ; les différences suivent une distribution normale (Shapiro-Wilk p = 0,725), d'où un test t apparié : t = 11,14, ddl 159, p < 0,001 ; dz = 0,88. Attention : la baisse globale ne peut pas être attribuée à l'intervention ; pour juger de l'effet de l'intervention, nous comparons les bras.
5. La baisse de la glycémie diffère-t-elle entre les bras ? Réponse type : ANOVA F(2 ; 157) = 22,93, p < 0,001, $\eta^2 = 0{,}226$; Tukey : B-A 8,1 mg/dL (IC 3,4-12,8), C-A 13,3 (8,6-18,0), C-B 5,2 (0,6-9,9). Tous les bras diffèrent entre eux ; l'effet le plus important est celui de l'association application-téléconsultation.
6. L'IMC est-il associé à la PAS ? Réponse type : r = 0,405, p < 0,001 (rho de Spearman 0,372, préférable étant donné l'asymétrie de l'IMC) ; $R^2 = 0{,}164$. Droite de régression : PAS = 89,47 + 1,435 x IMC (IC à 95% de la pente 0,925-1,944) : chaque kg/m² supplémentaire est associé à une PAS plus élevée de 1,4 mmHg ; pour un IMC de 30, la PAS prédite est de 132,5 mmHg. Dans la régression multiple avec l'âge, le sexe et le

tabagisme, le coefficient de l'IMC reste presque inchangé (1,389, p < 0,001), et le modèle explique 30,8% de la variabilité de la PAS (R^2 ajusté 0,286) ; à IMC, âge et sexe égaux, les fumeurs ont une PAS plus élevée de 6,7 mmHg que les anciens fumeurs (p = 0,012).

7. L'obésité est-elle associée au diabète ? Réponse type : prévalence du diabète de 33,3% chez les obèses contre 10,4% chez les non-obèses ; rapport de prévalences 3,21 (IC à 95% 1,64-6,31) ; OR 4,32 (IC à 95% 1,86-10,03) ; χ^2 = 12,70, p < 0,001. Après ajustement sur l'âge et le sexe (régression logistique), OR 6,01 (IC à 95% 2,35-15,39). Schéma transversal : association, et non causalité.
8. Quelle est l'utilité du FINDRISC et comment la probabilité de diabète change-t-elle selon son résultat ? Réponse type : AUC 0,852 (IC à 95% 0,762-0,943). Au seuil ≥ 12 : Se 75,9%, Sp 74,8%, RV+ 3,01, RV- 0,32. En partant d'une prévalence de 18,1% (cote pré-test 0,181/0,819 = 0,221), un score positif donne une cote post-test de 0,221 x 3,01 = 0,665, soit une probabilité de 0,665/1,665 = 40%, identique à la VPP ; un score négatif donne une cote de 0,221 x 0,32 = 0,071, soit une probabilité de 6,6%, c'est-à-dire 1 - VPN. Le théorème de Bayes (chapitre 11) et le tableau 2 x 2 (chapitre 12) disent la même chose. Dans une population où la prévalence est de 5%, le même test n'aurait qu'une VPP de 13,7%.

14.4 Comment rédiger le paragraphe des résultats

Un résultat statistique bien rapporté répond à quatre questions : ce qui a été comparé (groupes, variable, unité), l'ampleur de la différence ou de l'association, la précision de l'estimation et son degré de compatibilité avec le hasard. L'ordre naturel est le suivant : la description dans chaque groupe, la taille de l'effet avec son IC à 95%, le test avec sa statistique, les degrés de liberté et la valeur p. La conclusion clinique vient ensuite, dans la discussion, et non mêlée aux chiffres.

Voici quelques paires d'exemples, tous fondés sur les données mg2026.

Exemple 1, version faible : "La glycémie a baissé significativement dans le bras C (p < 0,05)." Les problèmes : la phrase ne dit pas par rapport à quoi (par rapport à l'inclusion ou par rapport à un autre bras ?), elle ne donne aucun chiffre sur l'ampleur de la baisse, "p < 0,05" masque une valeur bien plus faible, et le test n'est pas nommé.

Exemple 1, version correcte : "À 6 mois, la baisse moyenne de la glycémie à jeun était de 3,0 mg/dL (ÉT 8,8) dans le bras A, de 11,1 mg/dL (ÉT 11,2) dans le bras B et de 16,3 mg/dL (ÉT 10,6) dans le bras C (ANOVA, $F(2;\ 157) = 22{,}93$, $p < 0{,}001$; $\eta^2 = 0{,}23$). Par rapport au bras A, la baisse était plus importante de 8,1 mg/dL dans le bras B (IC à 95% 3,4-12,8) et de 13,3 mg/dL dans le bras C (IC à 95% 8,6-18,0) (test de Tukey)."

Exemple 2, version faible : "Il n'y a pas de différence de PAS entre les sexes (p = 0,81), donc le sexe n'influence pas la pression artérielle." Les problèmes : un résultat non significatif ne prouve pas l'absence de différence, et la seconde partie de la phrase avance une affirmation causale qu'une étude transversale ne peut pas étayer.

Exemple 2, version correcte : "La PAS moyenne était de 131,4 mmHg (ÉT 14,5) chez les femmes et de 130,9 mmHg (ÉT 13,3) chez les hommes ; la différence de -0,5 mmHg (IC à 95% -4,9 à 3,8 ; test t de Welch, p = 0,81) n'était pas statistiquement significative. Les données sont incompatibles avec des différences entre les sexes supérieures à environ 5 mmHg."

Exemple 3, version faible : "L'obésité multiplie le risque de diabète par 4,32 (p < 0,001)." Les problèmes : l'OR n'est pas un risque relatif, "multiplie" suggère une causalité, l'IC manque.

Exemple 3, version correcte : "La prévalence du diabète confirmé par HGPO était de 33,3% (18/54) chez les patients obèses et de 10,4% (11/106) chez les patients non obèses (rapport de

prévalences 3,21 ; IC à 95% 1,64-6,31 ; χ^2 = 12,70, p < 0,001). L'association persistait après ajustement sur l'âge et le sexe (OR 6,01 ; IC à 95% 2,35-15,39)."

Exemple 4, version faible : "La corrélation entre l'IMC et la PAS était très significative (p < 0,001), donc l'IMC détermine la PAS." Les problèmes : "très significative" parle de p, et non de la force de la corrélation ; la corrélation ne démontre pas la causalité.

Exemple 4, version correcte : "L'IMC était corrélé positivement, de façon modérée, à PAS($r = 0{,}41,\ p < 0{,}001;\ R^2 = 0{,}16$). Chaque augmentation de l'IMC de 1 kg/m² était associée à une PAS plus élevée de 1,4 mmHg (IC à 95% 0,9-1,9)."

Exemple 5, version faible : "Le FINDRISC a une exactitude de 75%, c'est donc un bon test." Les problèmes : l'exactitude globale dépend de la prévalence et masque le compromis entre Se et Sp ; "bon" n'a pas de sens sans préciser à quoi sert le test.

Exemple 5, version correcte : "L'aire sous la courbe ROC du score FINDRISC pour le diabète confirmé par HGPO était de 0,852 (IC à 95% 0,762-0,943). Au seuil ≥ 12, la sensibilité était de 75,9% et la spécificité de 74,8% ; à la prévalence de l'échantillon (18,1%), la VPP était de 40,0% et la VPN de 93,3%."

Quelques règles de forme complètent ces exemples. Rapportez les valeurs p exactes (p = 0,034), sauf lorsqu'elles sont très faibles (p < 0,001) ; n'écrivez pas "p = 0,000", que certains logiciels affichent. Utilisez un nombre raisonnable de décimales : un pourcentage calculé sur un échantillon de 50 ne s'écrit pas avec trois décimales. Le nom du test apparaît une fois, lors de sa première utilisation. L'unité de mesure apparaît chaque fois qu'une valeur est donnée. La moyenne est accompagnée de l'ÉT lorsque vous décrivez les données et de l'IC lorsque vous estimez un paramètre.

> **Étude de cas : un résumé avec de nombreux rapports de cotes et sans intervalles de confiance**
>
> Ursoiu et ses collaborateurs ont analysé l'association entre la sévérité clinique de l'infection par le VIH et les troubles psychiatriques chez 101 patients sous traitement antirétroviral dans l'ouest de la Roumanie [84]. Le résumé rapporte sept rapports de cotes, tous significatifs : OR = 5,81 (p < 0,001), 3,11 (p = 0,016), 4,31 (p = 0,024), 2,81 (p = 0,038), 4,79 (p = 0,006), 8,72 (p < 0,001) et 4,66 (p = 0,0013). À titre d'exercice de rédaction, observez ce qui manque et comment on pourrait le compléter. Les intervalles de confiance manquent : avec 101 patients, un OR de 8,72 a presque à coup sûr un intervalle très large, et le lecteur ne sait pas si la borne inférieure est de 1,2 ou de 5. Il manque la précision indiquant si les OR sont bruts ou ajustés. De nombreuses associations sont testées sur le même échantillon, de sorte que le risque que l'une d'elles soit un faux positif n'est pas négligeable ; plusieurs valeurs p (0,024 ; 0,038) ne résisteraient pas à une correction de Bonferroni pour sept tests (seuil 0,05/7 = 0,007). Enfin, le schéma transversal ne permet pas de dire si la maladie avancée conduit aux troubles psychiatriques ou l'inverse. Un paragraphe de résultats réécrit donnerait, pour chaque association, les fréquences dans les deux groupes, l'OR avec son IC à 95%, la valeur p et, pour finir, une mention des comparaisons multiples. Ce n'est pas une critique des auteurs : les contraintes de place d'un résumé imposent des compromis, et le texte intégral peut contenir les informations manquantes. Pour le lecteur, la règle demeure toutefois : un OR sans intervalle de confiance est une information incomplète.

14.5 Quarante affirmations : vrai ou faux

Les affirmations ci-dessous couvrent l'ensemble du cours. Essayez de trancher par vous-mêmes avant de lire la justification ; à l'examen QCM, de nombreuses propositions fausses sont construites précisément sur ces confusions.

1. La valeur p est la probabilité que H0 soit vraie. Faux. La valeur p est la probabilité d'obtenir un résultat au moins aussi extrême que celui observé, en supposant que H0 est vraie ; ce n'est pas une probabilité portant sur l'hypothèse [70].
2. Un IC à 95% d'une différence de moyennes qui ne contient pas 0 correspond à un test bilatéral significatif à $\alpha = 0{,}05$. Vrai. L'intervalle et le test utilisent la même erreur standard et la même valeur critique.
3. La médiane est moins influencée par les valeurs extrêmes que la moyenne. Vrai. Dans mg2026, une seule valeur de PAS de 1 540 mmHg (au lieu de 154) déplace la moyenne de 8,66 mmHg, alors que la médiane reste à 131.
4. L'écart-type diminue lorsque la taille de l'échantillon augmente. Faux. L'écart-type estime la variabilité dans la population et reste, en moyenne, le même ; c'est l'erreur standard qui diminue.
5. L'erreur standard de la moyenne diminue proportionnellement à $\sqrt{n}$. Vrai. ES $= \frac{s}{\sqrt{n}}$; pour la diviser par deux, n doit être multiplié par 4.
6. Si nous doublons l'échantillon, la largeur de l'intervalle de confiance est divisée par deux. Faux. La largeur diminue d'un facteur $\sqrt{2}$, soit d'environ 29% ; pour la diviser par deux, il faut un échantillon 4 fois plus grand.
7. Le test t apparié convient pour comparer la PAS entre les femmes et les hommes. Faux. Les femmes et les hommes forment des groupes indépendants ; on utilise le test t pour échantillons indépendants (ou le test de Welch, le test de Mann-Whitney).
8. Une ANOVA significative indique quels groupes diffèrent entre eux. Faux. L'ANOVA dit seulement qu'au moins une moyenne diffère ; les paires sont identifiées par des tests post-hoc (Tukey, Bonferroni).
9. Le test du khi-deux exige des effectifs attendus suffisamment grands, en règle générale au moins 5 dans chaque case. Vrai. Sinon, on utilise le test exact de Fisher (pour les tableaux 2 x 2).
10. Un coefficient de corrélation r = 0,9 montre que X cause Y. Faux. La corrélation mesure l'association linéaire ; le sens de la causalité ou l'intervention d'un troisième facteur ne peuvent pas être exclus.
11. $R^2 = 0{,}164$ pour la régression de la PAS sur l'IMC signifie que 16,4% de la variabilité de la PAS est expliquée linéairement par l'IMC. Vrai. Les 83,6% restants relèvent d'autres facteurs et de la variabilité individuelle.
12. Le coefficient de Spearman convient aux variables ordinales. Vrai. Il travaille sur les rangs et n'exige pas la normalité.
13. Dans une étude cas-témoins, nous pouvons calculer directement le risque relatif. Faux. La proportion de malades est fixée par le chercheur (nombre de cas et de témoins), les risques ne peuvent donc pas être estimés ; on calcule l'OR.
14. Lorsque la maladie est rare, l'OR est une bonne approximation du RR. Vrai. Lorsque a et c sont petits par rapport à b et d, les cotes sont proches des risques.
15. Le NST est l'inverse de la réduction relative du risque. Faux. NST = 1 / réduction absolue du risque.
16. Un RR = 1,02 avec un IC à 95% de 0,80-1,30 prouve que l'exposition n'a aucun effet. Faux. Les données sont compatibles avec une réduction de 20% ou avec une augmentation de 30% ; nous n'avons pas de preuve d'un effet, mais pas non plus la preuve de son absence [79].
17. À sensibilité et spécificité constantes, la VPP augmente lorsque la prévalence augmente. Vrai. Pour FINDRISC ≥ 12, la VPP passe de 5,8% pour une prévalence de 2% à 75,1% pour une prévalence de 50%.
18. La sensibilité d'un test varie directement avec la prévalence de la maladie. Faux. La Se et la Sp sont des propriétés du test vis-à-vis des malades et des sujets sains, respectivement ; elles peuvent changer avec le spectre de la maladie, mais pas avec la prévalence en tant que telle.

19. Un résultat négatif à un test de très grande sensibilité aide à exclure la maladie. Vrai. Le test manque peu de malades, un résultat négatif a donc une VPN élevée.
20. $RV^+ = \frac{Se}{1-Sp}$. Vrai. Pour FINDRISC ≥ 12 : 0,759 / 0,252 = 3,01.
21. Le taux d'erreur de classement est (FP + FN) / N. Vrai. Pour FINDRISC ≥ 12 : (33 + 7)/160 = 25%, le complément de l'exactitude de 75%.
22. Une aire sous la courbe ROC de 0,5 correspond à un test dépourvu de valeur discriminante. Vrai. Le test ne fait pas mieux qu'un tirage à pile ou face.
23. La randomisation garantit que les bras sont identiques à l'inclusion. Faux. Elle les équilibre en moyenne ; des différences dues au hasard subsistent (dans mg2026, 9,4% d'anciens fumeurs dans le bras A contre 31,5% dans le bras C).
24. L'insu et la dissimulation de l'allocation sont la même chose. Faux. La dissimulation de l'allocation protège l'inclusion, l'insu protège le traitement et l'évaluation.
25. L'analyse en intention de traiter inclut tous les patients randomisés, dans le bras qui leur a été attribué. Vrai. C'est ainsi que l'on préserve la comparabilité assurée par la randomisation.
26. Une étude transversale peut établir ce qui est apparu en premier, l'exposition ou la maladie. Faux. L'exposition et la maladie sont mesurées au même moment.
27. La puissance d'un test augmente avec la taille de l'échantillon. Vrai. L'erreur standard diminue, si bien que les effets réels sont détectés plus facilement.
28. L'erreur de type I consiste à rejeter une H0 vraie. Vrai. Sa probabilité est α.
29. Si p = 0,06, H0 est vraie. Faux. Nous disons seulement que nous ne rejetons pas H0 au seuil de 0,05 ; H0 peut être fausse, et l'étude manquer de puissance.
30. Un test de Shapiro-Wilk non significatif prouve que les données sont normales. Faux. Nous ne rejetons pas la normalité ; sur de petits échantillons, le test a une faible puissance et peut manquer des écarts importants.
31. La transformation logarithmique peut rendre symétrique une distribution à asymétrie positive. Vrai. Les triglycérides de mg2026 passent d'une asymétrie de 0,95 à -0,28.
32. La moyenne ± 1,96 ÉT est l'intervalle de confiance de la moyenne. Faux. C'est l'intervalle dans lequel se trouvent environ 95% des valeurs individuelles ; l'IC de la moyenne utilise l'erreur standard.
33. Le test de Kruskal-Wallis est l'alternative non paramétrique de l'ANOVA à un facteur. Vrai. Il compare trois groupes indépendants ou plus à partir des rangs.
34. Le test de Friedman s'utilise pour des groupes indépendants. Faux. Il s'applique à trois mesures ou plus effectuées chez les mêmes sujets (données dépendantes).
35. Dans un très grand échantillon, une différence cliniquement négligeable peut s'avérer statistiquement significative. Vrai. La corrélation r = -0,088 entre l'âge et la durée de l'intervention, chez 867 patients, a un p = 0,01 [55].
36. Un α de Cronbach de 0,93 prouve la validité d'un questionnaire. Faux. L'alpha mesure la cohérence interne (la fidélité), et non la validité.
37. La pente d'une droite de régression a la même unité de mesure que la variable dépendante. Faux. Son unité est celle de Y par unité de X (par exemple, des mmHg par kg/m²).
38. Le coefficient r ne change pas si nous exprimons la taille en mètres au lieu de centimètres. Vrai. r n'a pas d'unité de mesure et il est invariant par transformation linéaire des variables.
39. Dans le Tableau 1 d'un ECR, les différences entre les bras doivent obligatoirement être testées avec des valeurs p. Faux. CONSORT recommande de ne pas les tester ; les différences sont, par construction, dues au hasard [56].

40. Le test du khi-deux significatif pour une baisse ≥ 5% dans les trois bras de mg2026 (p = 0,034) montre que tous les bras diffèrent entre eux. Faux. Il montre qu'au moins un bras diffère ; B et C ont des proportions presque identiques (RR de C par rapport à B 0,98, IC à 95% 0,55-1,75).

14.6 Tableau récapitulatif des formules

Le Tableau 14.2 rassemble les formules de travail du cours, avec un renvoi au chapitre dans lequel elles ont été expliquées. Notations : n la taille de l'échantillon, m la moyenne, s l'écart-type, p la proportion, a, b, c, d les cases du tableau 2 x 2 (a exposés malades ou test positif malades, b exposés sains ou test positif sains, c non exposés malades ou test négatif malades, d non exposés sains ou test négatif sains).

Tableau 14.2. Principales formules du cours.

Indicateur	Formule	Chapitre
Moyenne	$m = \frac{\sum x}{n}$	2
Écart-type	$s = \sqrt{\left[\frac{\sum(x-m)^2}{n-1}\right]}$	2
Coefficient de variation	CV = 100 x s / m (%)	2
Score z	$z = \frac{x-m}{s}$	3
Intervalle des valeurs individuelles (95%)	$m \pm 1{,}96 \times s$	3
Erreur standard de la moyenne	$\mathrm{ES} = \frac{s}{\sqrt{n}}$	4
IC à 95% de la moyenne	$m \pm \frac{t(n-1;\ 0{,}975) \times s}{\sqrt{n}}$	4
IC à 95% d'une proportion (Wald)	$p_i \pm 1{,}96 \times \sqrt{\left[\frac{p_i(1-p_i)}{n}\right]}$	4
Taille d'échantillon pour une moyenne avec une marge E	$n = \left(\frac{1{,}96 \times \sigma}{E}\right)^2$	4
Taille d'échantillon pour une proportion avec une marge E	$n = \frac{1{,}96^2 \times p_i(1-p_i)}{E^2}$	4
Taille d'échantillon par groupe, deux moyennes	$n = \frac{2 \times \left(z\left(1-\frac{\alpha}{2}\right) + z(1-\beta)\right)^2 \times \sigma^2}{\Delta^2}$	4, 13
Test t pour échantillons indépendants	$t = \frac{m_1 - m_2}{\sqrt{s^2 p \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	6
Test t apparié	$t = \frac{\mathrm{md}}{\frac{\mathrm{sd}}{\sqrt{n}}}$, ddl = n - 1	6
d de Cohen	$d = \frac{m_1 - m_2}{\mathrm{sp}}$	6
ANOVA	$F = \frac{\frac{\mathrm{SSB}}{k-1}}{\frac{\mathrm{SSW}}{N-k}}$	6
Êta carré	$\eta^2 = \frac{\mathrm{SSB}}{\mathrm{SSB} + \mathrm{SSW}}$	6
Effectif attendu (khi-deux)	E = total de la ligne x total de la colonne / N	7
Khi-deux	$\chi^2 = \frac{\sum(O-E)^2}{E}$, ddl = (r - 1)(c - 1)	7
Test de r	$t = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}}$	9

Indicateur	Formule	Chapitre
Coefficient de détermination	$R^2 = r^2$ (régression simple)	9
Droite de régression	Y = a + b x X ; $b = \frac{\sum(x-m_x)(y-m_y)}{\sum(x-m_x)^2}$; a = my - b x mx	9
Risque relatif (cohorte, ECR)	$\mathrm{RR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$	10
Rapport de cotes	$\mathrm{OR} = \frac{a \times d}{b \times c}$	10
IC à 95% du RR	$\exp\left[\ln \mathrm{RR} \pm 1{,}96 \times \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}\right]$	10
IC à 95% de l'OR	$\exp\left[\ln \mathrm{OR} \pm 1{,}96 \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right]$	10
Différence de risques	$\mathrm{DR} = \frac{a}{a+b} - \frac{c}{c+d}$	10
NST (NNH)	$\mathrm{NNT} = \frac{1}{\mathrm{DR}}$, arrondi à l'entier supérieur	10, 13
Fraction attribuable chez les exposés	$\mathrm{FA} = \frac{\mathrm{RR} - 1}{\mathrm{RR}}$	10
Cote et probabilité	cote = P / (1 - P) ; P = cote / (1 + cote)	11
Théorème de Bayes (VPP)	$\mathrm{VPP} = \frac{\mathrm{Se} \times P}{\mathrm{Se} \times P + (1 - \mathrm{Sp}) \times (1 - P)}$	11
Cote post-test	cote post-test = cote pré-test x RV	11
Sensibilité	$\mathrm{Se} = \frac{a}{a+c}$	12
Spécificité	$\mathrm{Sp} = \frac{d}{b+d}$	12
Valeur prédictive positive	$\mathrm{VPP} = \frac{a}{a+b}$	12
Valeur prédictive négative	$\mathrm{VPN} = \frac{d}{c+d}$	12
Exactitude	$\frac{a+d}{N}$	12
Taux d'erreur	$\frac{b+c}{N}$	12
Rapports de vraisemblance	$\mathrm{RV}^+ = \frac{\mathrm{Se}}{1-\mathrm{Sp}}$; $\mathrm{RV}^- = \frac{1-\mathrm{Se}}{\mathrm{Sp}}$	12
Indice de Youden	$J = \mathrm{Se} + \mathrm{Sp} - 1$	12
Alpha de Cronbach	$\alpha = \frac{k}{k-1} \times \left[1 - \frac{\sum s^2(\mathrm{item})}{s^2(\mathrm{total})}\right]$	13
Effet de plan (grappes)	1 + (m - 1) x ICC	13

Ce tableau ne s'apprend pas par cœur comme une liste. Chaque formule obéit à une logique : l'erreur standard est la variabilité divisée par la racine de n ; une statistique de test est un effet divisé par son erreur standard ; un intervalle de confiance est l'estimation plus ou moins une valeur critique multipliée par l'erreur standard ; les indicateurs des tests diagnostiques sont des proportions calculées sur les lignes ou sur les colonnes d'un même tableau 2 x 2. Qui comprend ces quatre idées peut reconstituer presque tout le tableau.

14.7 Problèmes résolus

Problème résolu 14.1. Tableau de fréquences avec valeurs manquantes

Énoncé. Dans un questionnaire administré à 200 patients d'un cabinet de médecine générale, la question sur le niveau d'études a reçu les réponses suivantes : collège 30, lycée ou école professionnelle 74, universitaire 62, post-universitaire 22, sans réponse 12. Construisez le tableau de fréquences, avec les pourcentages du total et les pourcentages valides, et précisez la catégorie médiane.

Solution. Le nombre de réponses valides est de 200 - 12 = 188. Le pourcentage du total se calcule sur 200, le pourcentage valide sur 188, et le pourcentage cumulé à partir des pourcentages valides.

Tableau 14.3. Distribution du niveau d'études (n = 200, 12 valeurs manquantes).

Niveau d'études	Fréquence	% du total	% valide	% valide cumulé
Collège	30	15,0	16,0	16,0
Lycée ou école professionnelle	74	37,0	39,4	55,3
Universitaire	62	31,0	33,0	88,3
Post-universitaire	22	11,0	11,7	100,0
Manquant	12	6,0	-	-
Total	200	100,0	100,0	

La catégorie médiane est celle dans laquelle le pourcentage valide cumulé dépasse pour la première fois 50% : lycée ou école professionnelle (55,3%). La variable est ordinale : une médiane a donc un sens, mais pas une moyenne. Les pourcentages valides décrivent les répondants ; si l'absence de réponse n'est pas due au hasard (par exemple, si les personnes moins instruites évitent la question), les pourcentages valides peuvent être faussés, et le nombre de valeurs manquantes doit toujours être rapporté.

Problème résolu 14.2. Une étude cas-témoins : contraceptifs oraux et polyarthrite rhumatoïde

Énoncé. Une étude cas-témoins inclut 120 femmes atteintes de polyarthrite rhumatoïde et 240 femmes indemnes de la maladie, d'âge comparable. Ont utilisé des contraceptifs oraux par le passé 42 des cas et 54 des témoins. Calculez la mesure d'association appropriée, avec son IC à 95%, et testez l'association.

Solution. (a) Hypothèse de recherche : l'utilisation de contraceptifs oraux est associée à la polyarthrite rhumatoïde. (b) H0 : OR = 1 (pas d'association) ; H1 : OR ≠ 1. (c) Schéma cas-témoins, la mesure est donc l'OR ; le test est le khi-deux pour un tableau 2 x 2. Le tableau : a = 42 (cas exposés), b = 54 (témoins exposés), c = 78 (cas non exposés), d = 186 (témoins non exposés).

$$\text{OR} = \frac{42\times186}{54\times78} = \frac{7812}{4212} = 1{,}85;\ \text{ES}(\ln \text{OR}) = \sqrt{\frac{1}{42}+\frac{1}{54}+\frac{1}{78}+\frac{1}{186}} = \sqrt{0{,}0238+0{,}0185+0{,}0128+0{,}0054} = \sqrt{0{,}0605} = 0{,}246 \ln 1{,}85 = 0{,}618\ ;\ 0{,}618 \pm 1{,}96 \times 0{,}246 = 0{,}618 \pm 0{,}482, \text{ soit } 0{,}136 \text{ à } 1{,}100;\ \text{IC}\,95\% = e^{0{,}136} \text{ à } e^{1{,}100} = 1{,}15 \text{ à } 3{,}00$$

Khi-deux (sans correction) = 6,39, ddl = 1, p = 0,011. (e) Résultat : OR 1,85 (IC à 95% 1,15-3,00). (f) Conclusion statistique : nous rejetons H0 ; l'intervalle ne contient pas 1. (g) Conclusion de recherche : la cote d'utilisation de contraceptifs oraux est environ 1,9 fois plus élevée chez les femmes atteintes de polyarthrite rhumatoïde que chez les témoins. Une étude cas-témoins ne

permet pas de calculer le risque relatif, et l'association peut être influencée par un biais de mémorisation (les cas se souviennent différemment de leurs expositions passées) et par des facteurs de confusion.

Problème résolu 14.3. Une cohorte : temps passé devant l'écran et acuité visuelle

Énoncé. Une cohorte de 400 salariés est suivie pendant 3 ans. Parmi les 150 qui travaillent au moins 6 heures par jour devant un écran, 45 présentent une baisse de l'acuité visuelle ; parmi les 250 qui y travaillent moins longtemps, 30. Calculez le RR, la DR, le NNH et la fraction attribuable chez les exposés.

Solution. Risque chez les exposés : 45/150 = 0,30 ; chez les non-exposés : 30/250 = 0,12.

$\text{RR} = \frac{0{,}30}{0{,}12} = 2{,}5;\ \text{ES}(\ln \text{RR}) = \sqrt{\frac{1}{45} - \frac{1}{150} + \frac{1}{30} - \frac{1}{250}} = \sqrt{0{,}0222 - 0{,}0067 + 0{,}0333 - 0{,}0040} = \sqrt{0{,}0449} = 0{,}212;\ \text{IC}\,95\%\ \text{RR} = \exp(\ln 2{,}5 \pm 1{,}96 \times 0{,}212) = \exp(0{,}916 \pm 0{,}415) = 1{,}65$ à 3,79 DR = 0,30 - 0,12 = 0,18 ; $\text{NNH} = \frac{1}{0{,}18} = 5{,}6$, soit 6; $\text{FA} = \frac{2{,}5-1}{2{,}5} = 0{,}60$

Khi-deux = 19,94, p < 0,001. OR-ul ar fi $\frac{(45\times220)}{105\times30} = 3{,}14$, supérieur au RR, comme c'est le cas lorsque l'événement est fréquent. Interprétation : le risque de baisse de l'acuité est 2,5 fois plus élevé chez les exposés ; sur 6 salariés exposés, il survient un cas de plus que parmi autant de non-exposés ; 60% des cas survenus chez les exposés pourraient être attribués à l'exposition, si la relation est causale. S'agissant d'une cohorte observationnelle, les salariés exposés peuvent aussi différer par l'âge ou par d'autres caractéristiques : une analyse ajustée est donc nécessaire avant toute conclusion causale.

Problème résolu 14.4. Reconstitution du tableau d'un test diagnostique

Énoncé. Un test de dépistage est appliqué à 500 personnes, dont 20% sont atteintes de la maladie. La sensibilité du test est de 85% et la VPP de 68%. Reconstituez le tableau 2 x 2 et calculez la Sp, la VPN, l'exactitude, le taux d'erreur et les rapports de vraisemblance.

Solution. Malades : 0,20 x 500 = 100 ; sujets sains : 400. Vrais positifs : 0,85 x 100 = 85 ; faux négatifs : 15. De $\text{VPP} = \frac{\text{VP}}{\text{VP+FP}} = 0{,}68$ on déduit $\text{VP} + \text{FP} = \frac{85}{0{,}68} = 125$, donc FP = 40. Vrais négatifs : 400 - 40 = 360.

Tableau 14.4. Tableau reconstitué du test de dépistage.

	Malades	**Sains**	**Total**
Test positif	85	40	125
Test négatif	15	360	375
Total	100	400	500

$\text{Sp} = \frac{360}{400} = 90\%$; $\text{VPN} = \frac{360}{375} = 96\%$; exactitude = (85 + 360)/500 = 89% ; taux d'erreur = (40 + 15)/500 = 11% ; $\text{RV}^{+} = \frac{0{,}85}{0{,}10} = 8{,}5$; $\text{RV}^{-} = \frac{0{,}15}{0{,}90} = 0{,}17$. Vérification par le théorème de Bayes : cote pré-test 0,20/0,80 = 0,25 ; cote post-test pour un résultat positif 0,25 x 8,5 = 2,125 ; probabilité 2,125/3,125 = 68%, soit exactement la VPP de l'énoncé.

Problème résolu 14.5. Choix du test sur le jeu de données mg2026

Énoncé. Pour chaque question, choisissez le test, justifiez votre choix et donnez le résultat obtenu sur le jeu de données mg2026 : a) la glycémie initiale diffère-t-elle entre les hommes et les femmes ? b) le niveau d'études est-il associé à la satisfaction ? c) la PAS initiale diffère-t-elle entre fumeurs, anciens fumeurs et non-fumeurs ? d) le poids a-t-il changé dans le bras A ? e) le tabagisme est-il associé à l'HTA (PAS ≥ 140) à l'inclusion ?

Solution. a) Variable numérique, deux groupes indépendants ; la glycémie s'écarte légèrement de la normalité (SW p = 0,037), mais avec 74 et 86 patients le test t de Welch est robuste : 106,2 contre 104,8 mg/dL, t = 0,54, p = 0,589 ; nous ne rejetons pas H0. b) Deux variables ordinales : corrélation de Spearman, rho = 0,125, p = 0,115 ; il n'y a pas d'association significative. c) Variable numérique, trois groupes indépendants ; la normalité est acceptable dans chaque groupe (SW p 0,37 ; 0,86 ; 0,23), Levene p = 0,114 : ANOVA, F = 7,85, p < 0,001, $\eta^2 = 0{,}09$; Tukey : les fumeurs ont une PAS plus élevée de 9,5 mmHg que les anciens fumeurs (IC à 95% 2,5-16,5) et de 8,5 mmHg que les non-fumeurs (IC à 95% 2,9-14,2). d) Données appariées (les mêmes patients à 0 et à 6 mois), un seul groupe ; test de Wilcoxon : médiane des différences 1,1 kg, p = 0,052, résultat limite qui ne permet pas de rejeter H0 à α = 0,05. e) Deux variables dichotomiques, schéma transversal : khi-deux ; 21/53 (39,6%) des fumeurs contre 20/107 (18,7%) des non-fumeurs et anciens fumeurs ont une HTA ; $\chi^2 = 8{,}15$, p = 0,004 ; rapport de prévalences 2,12 (IC à 95% 1,26-3,55), OR 2,85 (IC à 95% 1,37-5,95).

14.8 Exercices proposés

1. Dans une étude portant sur 250 adultes, le statut tabagique était le suivant : non-fumeurs 120, fumeurs 70, anciens fumeurs 40, sans réponse 20. Construisez le tableau de fréquences avec les pourcentages du total et les pourcentages valides.
2. Une étude cas-témoins inclut 80 patients atteints de BPCO et 160 témoins. Sont fumeurs 56 des cas et 64 des témoins. Calculez l'OR avec son IC à 95% et le khi-deux.
3. Dans une cohorte, sur 200 sujets exposés à un facteur professionnel, 30 tombent malades, et sur 300 non-exposés, 15. Calculez le RR avec son IC à 95%, la DR et le NNH.
4. Un test de Se 90% et de Sp 95% est appliqué en dépistage à 10 000 personnes, avec une prévalence de la maladie de 2%. Reconstituez le tableau 2 x 2 et calculez la VPP et la VPN. Commentez.
5. Chez 36 patients, la moyenne du cholestérol est de 220 mg/dL et l'ÉT de 18 mg/dL. Calculez l'erreur standard et l'IC à 95% de la moyenne (t(35 ; 0,975) = 2,03).
6. Chez 27 patients, la corrélation entre deux variables numériques est r = 0,6. Calculez R^2 et la statistique t, et précisez si r est significatif à α = 0,05 (valeur critique t(25) ≈ 2,06).
7. À l'aide de la droite de régression de mg2026 (PAS = 89,47 + 1,435 x IMC), calculez la PAS prédite chez un patient dont l'IMC est de 27 kg/m². Pourquoi une prédiction pour un IMC de 15 ne serait-elle pas correcte ?
8. La probabilité pré-test d'une maladie est de 30%, et le RV+ d'un test est de 6. Calculez la probabilité post-test après un résultat positif.
9. Choisissez le test pour : a) comparer l'HbA1c entre les trois bras d'un ECR, avec des distributions normales ; b) comparer un score de douleur (1-10) avant et après une intervention, chez les mêmes patients ; c) l'association entre le sexe et la présence d'un diabète ; d) la relation entre l'âge et la PAS ; e) comparer une variable asymétrique entre deux groupes indépendants ; f) l'évolution d'un score à 0, 3 et 6 mois chez les mêmes patients, avec une distribution asymétrique.

10. Rédigez le paragraphe de résultats pour l'analyse suivante de mg2026 : HDL chez les femmes 53,92 mg/dL(ÉT 11,70; $n = 86$), chez les hommes 44,41 mg/dL(ÉT 9,66; $n = 74$) ; Levene p = 0,034 ; test t de Welch t = 5,63, ddl 157,7, p < 0,001 ; différence 9,5 mg/dL (IC à 95% 6,2-12,9).
11. Réécrivez correctement l'affirmation : "Le test du khi-deux a montré que l'obésité cause le diabète (p < 0,001), le risque étant 4,32 fois plus élevé."
12. Un article rapporte : "durée d'hospitalisation de 6,2 ± 5,9 jours dans le groupe A et de 4,1 ± 4,4 jours dans le groupe B (test t, p = 0,04)". Que remarquez-vous et qu'auriez-vous recommandé ?

Pièges fréquents

- Sauter la partie méthodes et ne lire que la conclusion de l'article.
- Interpréter "non significatif" comme "il n'y a pas d'effet", sans regarder l'intervalle de confiance.
- Confondre signification statistique et importance clinique, surtout dans les grands échantillons.
- Lire un OR comme un risque relatif lorsque l'événement est fréquent.
- Transposer la VPP et la VPN d'une étude à forte prévalence à une population de dépistage à faible prévalence.
- Croire à une spécificité de 100% obtenue sur quelques dizaines de volontaires sains.
- Ne pas vérifier si les intervalles de confiance publiés sont plausibles pour le n rapporté.
- Tirer des conclusions causales d'études transversales ou de corrélations.
- Décrire par la moyenne ± ÉT des variables visiblement asymétriques (ÉT proche de la moyenne pour une variable positive).
- Changer de test en fonction du résultat ou ne rapporter que les comparaisons significatives.
- Oublier qu'un HR suppose un suivi dans le temps et ne peut pas provenir d'un schéma purement transversal.

À retenir

- La carte de la discipline : question, schéma d'étude, données, description, estimation, test, association, décision, rédaction.
- Le choix du test dépend du type de variable du critère de jugement, du nombre de groupes et de leur indépendance.
- Un résultat complet comporte trois parties : la taille de l'effet, l'intervalle de confiance et la valeur p.
- L'intervalle de confiance montre la précision ; un intervalle large signifie une étude qui n'a pas tranché grand-chose.
- Les valeurs prédictives dépendent de la prévalence ; la Se, la Sp et les RV sont plus stables d'une population à l'autre.
- On combat la confusion par la randomisation (lors de la conception) ou par l'ajustement (lors de l'analyse).
- Beaucoup de chiffres publiés peuvent être vérifiés avec les formules simples du cours ; faites cet exercice chaque fois que quelque chose semble trop beau.

Questions d'autoévaluation

Le QCM récapitulatif ci-dessous comporte 30 questions dans le style du concours de résidanat : les questions 1 à 15 ont une seule réponse correcte (QRU), les questions 16 à 30 ont entre deux et quatre réponses correctes (QRM). Les réponses se trouvent dans le fichier de réponses du chapitre.

1. (QRU) Pour décrire l'IMC dans le jeu de données mg2026 (asymétrie 1,42, Shapiro-Wilk p < 0,001), les indicateurs les plus appropriés sont :
    a) la moyenne et l'écart-type
    b) la médiane et l'écart interquartile
    c) le mode
    d) l'étendue
    e) le coefficient de variation
2. (QRU) L'intervalle de confiance à 95% de la moyenne de la PAS dans mg2026 est de 128,96-133,30 mmHg. L'affirmation correcte est :
    a) 95% des patients ont une PAS comprise entre 128,96 et 133,30 mmHg
    b) la moyenne de l'échantillon se trouve dans l'intervalle avec une probabilité de 95%
    c) la procédure utilisée produit des intervalles qui contiennent la moyenne de la population dans 95% des échantillons possibles
    d) la PAS est significativement différente de 130 mmHg
    e) la distribution de la PAS est normale
3. (QRU) Pour comparer la glycémie à l'inclusion et à 6 mois chez les mêmes 160 patients, avec des différences de distribution normale, on utilise :
    a) le test t pour échantillons indépendants
    b) le test t pour échantillons appariés
    c) l'ANOVA à un facteur
    d) le test du khi-deux
    e) le test de Mann-Whitney
4. (QRU) La satisfaction (item de Likert 1-5) se compare entre les trois bras de l'étude mg2026 au moyen :
    a) de l'ANOVA à un facteur
    b) du test de Kruskal-Wallis
    c) du test t apparié
    d) du test de Friedman
    e) de la corrélation de Pearson
5. (QRU) Une étude suit pendant 10 ans 500 fumeurs et 500 non-fumeurs, en enregistrant la survenue d'une BPCO. Le schéma d'étude est :
    a) une étude cas-témoins
    b) une étude transversale
    c) une cohorte prospective
    d) un essai contrôlé randomisé
    e) une étude écologique
6. (QRU) Dans une étude cas-témoins, la mesure appropriée de l'association entre l'exposition et la maladie est :
    a) le risque relatif
    b) le rapport de cotes
    c) le nombre de sujets à traiter
    d) la prévalence
    e) l'incidence cumulée

7. (QRU) La corrélation r = -0,832 entre l'ostéoprotégérine et la fraction d'éjection indique :
    a) une corrélation faible
    b) une corrélation inverse forte
    c) que l'ostéoprotégérine diminue la fraction d'éjection
    d) un coefficient de détermination de 0,832
    e) une association non significative
8. (QRU) La réduction absolue du risque d'un événement est de 10 points de pourcentage. Le NST est de :
    a) 1
    b) 5
    c) 10
    d) 18
    e) 100
9. (QRU) Un test de Se 90% et de Sp 80% est appliqué à 1 000 personnes, avec une prévalence de la maladie de 15%. La VPP est d'environ :
    a) 15%
    b) 44%
    c) 80%
    d) 90%
    e) 98%
10. (QRU) Le taux d'erreur de classement d'un test diagnostique, avec les notations du tableau 2 x 2, est :

a) $\frac{a+d}{N}$

b) $\frac{b+c}{N}$

c) $\frac{a}{a+c}$

d) $\frac{d}{b+d}$

e) $\frac{a}{a+b}$

11. (QRU) Une ANOVA à trois groupes donne p < 0,001. L'étape suivante correcte est :

a) des tests t entre toutes les paires, sans correction

b) des tests post-hoc, par exemple le test de Tukey

c) le test du khi-deux

d) la conclusion que tous les groupes diffèrent entre eux

e) la répétition de l'ANOVA sur un autre échantillon avant toute interprétation

12. (QRU) Dans un essai randomisé de supériorité, l'analyse principale recommandée est :

a) l'analyse per protocole

b) l'analyse en intention de traiter

c) l'analyse limitée aux patients ayant terminé l'étude

d) l'analyse limitée aux patients observants

e) l'analyse en sous-groupes

13. (QRU) Pour le critère secondaire PAS de mg2026, l'ANOVA prévue au protocole donne p = 0,056. La formulation correcte est :

a) on accepte H0, l'intervention n'influence pas la PAS

b) on ne rejette pas H0 ; les données n'apportent pas de preuves suffisantes d'une différence entre les bras

c) il existe un effet significatif de l'intervention

d) on utilise le test de Kruskal-Wallis, qui donne $p < 0,05$, et c'est ce résultat que l'on rapporte

e) le résultat prouve l'équivalence des bras

14. (QRU) Dans une cohorte, 18 sujets exposés sur 150 et 10 non-exposés sur 250 développent la maladie. Le risque relatif est de :

a) 1,8

b) 3,0

c) 3,27

d) 0,33

e) 8,0

15. (QRU) Au seuil FINDRISC ≥ 12, la VPP est de 40,0% pour une prévalence de 18,1%. Dans une population où la prévalence est de 5%, à Se et Sp égales, la VPP sera :

a) supérieure à 40%

b) inférieure à 40%

c) égale à 40%

d) 95%

e) impossible à estimer

16. (QRM) À propos de la moyenne et de la médiane, sont vraies les affirmations suivantes :

a) la médiane est robuste aux valeurs extrêmes

b) dans une distribution à asymétrie positive, la moyenne est généralement supérieure à la médiane

c) dans une distribution normale, la moyenne et la médiane coïncident

d) la médiane ne se calcule que pour les variables nominales

e) la moyenne est l'indicateur recommandé pour un item de Likert isolé

17. (QRM) À propos de la valeur p, sont vraies les affirmations suivantes :

a) c'est la probabilité d'obtenir un résultat au moins aussi extrême si H0 est vraie

b) c'est la probabilité que H0 soit vraie

c) elle dépend de la taille de l'échantillon

d) une valeur très faible garantit un effet cliniquement important

e) elle se compare au seuil de signification α

18. (QRM) À propos des erreurs statistiques, sont vraies les affirmations suivantes :

a) l'erreur de type I consiste à rejeter une H0 vraie

b) α est la probabilité de l'erreur de type I

c) la puissance du test est $1 - \beta$

d) l'augmentation de la taille de l'échantillon diminue la puissance

e) l'erreur de type II consiste à ne pas rejeter une H0 fausse

19. (QRM) Les conditions d'application du test t pour échantillons indépendants sont :

a) la normalité dans chaque groupe ou des échantillons suffisamment grands

b) l'indépendance des observations

c) des variances égales, sinon on utilise la variante de Welch

d) des variables nominales

e) les mêmes patients mesurés deux fois

20. (QRM) À propos du test du khi-deux, sont vraies les affirmations suivantes :

a) il s'applique aux variables catégorielles

b) il exige des effectifs attendus suffisamment grands

c) lorsque les effectifs attendus sont faibles dans un tableau 2 x 2, on utilise le test de Fisher

d) il compare des moyennes

e) effectif attendu = total de la ligne x total de la colonne / N

21. (QRM) Pour la droite de régression PAS = 89,47 + 1,435 x IMC de mg2026, sont vraies les affirmations suivantes :

a) la PAS augmente en moyenne de 1,435 mmHg pour chaque kg/m² d'IMC

b) pour un IMC de 30, la PAS prédite est d'environ 132,5 mmHg

c) l'ordonnée à l'origine de 89,47 mmHg a une interprétation clinique directe

d) avec $R^2 = 0{,}164$, l'IMC explique linéairement 16,4% de la variabilité de la PAS

e) l'équation prouve qu'un IMC élevé cause l'hypertension

22. (QRM) À propos du RR et de l'OR, sont vraies les affirmations suivantes :

a) le RR se calcule dans les études de cohorte et dans les ECR

b) l'OR est la mesure appropriée dans les études cas-témoins

c) lorsque RR > 1 et que l'événement est fréquent, l'OR est supérieur au RR

d) RR = 1 indique l'absence d'association

e) un IC à 95% qui contient 1 indique une association significative

23. (QRM) À propos des indicateurs d'un test diagnostique, sont vraies les affirmations suivantes :

a) $\text{Se} = \frac{a}{a+c}$

b) $\text{Sp} = \frac{d}{b+d}$

c) la VPP dépend de la prévalence

d) la sensibilité augmente directement avec la prévalence

e) $\text{RV}^- = \frac{1-\text{Se}}{\text{Sp}}$

24. (QRM) À propos de la courbe ROC, sont vraies les affirmations suivantes :

a) elle représente la sensibilité en fonction de 1 - spécificité

b) une aire de 0,5 indique une absence de discrimination

c) l'indice de Youden est égal à Se + Sp - 1

d) le seuil qui maximise l'indice de Youden maximise la VPP

e) l'aire de 0,852 du score FINDRISC indique une bonne discrimination

25. (QRM) À propos du raisonnement bayésien, sont vraies les affirmations suivantes :

a) cote post-test = cote pré-test x RV

b) probabilité = cote / (1 + cote)

c) un RV+ élevé augmente la probabilité post-test après un résultat positif

d) la probabilité pré-test n'influence pas la probabilité post-test

e) un RV- inférieur à 1 diminue la probabilité de la maladie après un résultat négatif

26. (QRM) Pour la perte de poids ≥ 5% dans mg2026 (χ^2 = 6,77, ddl 2, p = 0,034), sont vraies les affirmations suivantes :

a) H0 est rejetée à α = 0,05
b) au moins un bras diffère des autres
c) les trois bras diffèrent tous entre eux
d) le RR de C par rapport à B de 0,98 (IC à 95% 0,55-1,75) ne montre pas de différence entre C et B
e) les effectifs attendus (minimum 12,6) permettent d'appliquer le test

27. (QRM) À propos des schémas d'étude, sont vraies les affirmations suivantes :
a) la randomisation équilibre en moyenne aussi les facteurs de confusion inconnus
b) l'étude de cohorte permet d'estimer l'incidence
c) l'étude transversale estime la prévalence
d) l'étude cas-témoins estime directement l'incidence de la maladie
e) l'ajustement multivarié peut corriger partiellement la confusion dans les études observationnelles

28. (QRM) À propos de la distribution normale, sont vraies les affirmations suivantes :
a) environ 68% des valeurs se situent dans l'intervalle $m \pm s$
b) environ 95% des valeurs se situent dans l'intervalle $m \pm 1{,}96s$
c) elle est symétrique par rapport à la moyenne
d) un test de Shapiro-Wilk avec p < 0,05 conduit à rejeter l'hypothèse de normalité
e) toutes les variables biologiques suivent une distribution normale

29. (QRM) À propos de la corrélation, sont vraies les affirmations suivantes :
a) r prend des valeurs comprises entre -1 et 1
b) le coefficient de Spearman convient aux variables ordinales ou asymétriques
c) r = 0 montre qu'il n'existe aucune relation d'aucune sorte entre les variables
d) une seule valeur extrême peut modifier fortement la valeur de r
e) dans un très grand échantillon, une corrélation faible peut être statistiquement significative

30. (QRM) À propos de la présentation des études, sont vraies les affirmations suivantes :
a) CONSORT s'utilise pour les essais randomisés
b) STROBE s'utilise pour les études observationnelles
c) STARD s'utilise pour les études d'exactitude diagnostique
d) le Tableau 1 d'un ECR doit contenir des valeurs p pour chaque caractéristique
e) l'effet principal est rapporté avec son ampleur et son intervalle de confiance à 95%

Partie II. Travaux pratiques

Les travaux pratiques suivent, semaine après semaine, les thèmes du cours et utilisent les outils prévus dans la fiche de la discipline : Google Forms et Google Sheets (compte Google de l'université), JASP version 0.98.1 ou plus récente [47], MedCalc 24.0 ou ses calculateurs en ligne [50] et Epi Info 7.2.7 [51], qui fonctionne aussi hors ligne. Le CDC a annoncé l'arrêt du développement et de l'assistance technique d'Epi Info en septembre 2025 ; le logiciel reste utilisable, mais ne reçoit plus de nouvelles versions, c'est pourquoi les travaux pratiques indiquent chaque fois aussi la variante dans JASP ou MedCalc. Les fichiers de travail se téléchargent sur la plateforme Moodle de la discipline (Emediqual, moodle.umft.ro) : le jeu de données pédagogique "mg2026" en deux variantes, mg2026_brut.csv et mg2026_curat.csv, ainsi que les fichiers classiques de la discipline utilisés pour les exercices supplémentaires. Le dictionnaire du jeu de données mg2026 se trouve à l'annexe C. Les travaux rendus sont déposés sur Moodle sous le nom LPx_Nom_Prenom.

Avec les paramètres régionaux France (ou Roumanie), Google Sheets utilise la virgule comme séparateur décimal et le point-virgule comme séparateur des arguments des fonctions. Les formules des travaux pratiques sont écrites sous cette forme, par exemple =AVERAGE(M2:M161). JASP et MedCalc affichent les résultats avec un point décimal ; dans les rapports, on écrit avec une virgule.

Chaque séance de travaux pratiques indique les résultats attendus sur le jeu de données mg2026. Ils ne sont pas donnés pour être recopiés, mais pour que l'étudiant puisse vérifier lui-même ses étapes : si la moyenne de la pression systolique n'est pas de 131,13 mmHg, quelque chose s'est produit lors de l'importation ou du nettoyage, et il vaut la peine de le trouver avant d'aller plus loin. Pour les devoirs et pour l'examen pratique, on utilise le format de présentation a-g : l'hypothèse de recherche, les hypothèses statistiques, la méthode choisie avec sa justification, la démarche de résolution, le résultat avec le graphique approprié, la conclusion statistique et la conclusion de recherche formulée en mots.

Travaux pratiques 1. Introduction à la biostatistique et aux outils numériques : Google Sheets, JASP et MedCalc

Objectif

La première séance de travaux pratiques a un but très concret : à la fin, chaque étudiant doit disposer d'un fichier de données propre, correctement organisé, accompagné d'un dictionnaire des données, et savoir ouvrir ce fichier dans les logiciels qu'il utilisera tout au long du semestre. Les calculs statistiques proprement dits restent simples pour l'instant. Ce qui compte maintenant, c'est la rigueur de travail : une ligne par patient, une colonne par variable, des valeurs cohérentes, aucune erreur cachée. L'expérience des dernières années montre que la plupart des résultats erronés à l'examen pratique ne viennent pas d'un test mal choisi, mais d'un fichier mal importé ou non nettoyé.

À la fin, l'étudiant doit être capable :

- de construire dans Google Sheets un jeu de données médicales de démonstration, avec des règles de validation et un dictionnaire des données ;
- d'importer correctement un fichier CSV dans Google Sheets, en tenant compte des paramètres régionaux (virgule décimale et séparateur ";") ;
- de détecter et de corriger les erreurs de saisie d'un fichier brut, en documentant chaque modification ;
- d'ouvrir le jeu de données dans JASP et dans MedCalc, de définir le type des variables et d'obtenir les statistiques descriptives de base ;
- de connaître le rôle du logiciel Epi Info et de déposer les résultats sur la plateforme Moodle.

Outils

Logiciels : Google Sheets (compte Google de l'université), JASP 0.98.1 [47], MedCalc 24.0 ou ses calculateurs en ligne [50], Epi Info 7.2.7 [51]. Fichiers sur Moodle : mg2026_curat.csv (160 patients) et mg2026_brut.csv (161 lignes, avec des erreurs laissées intentionnellement). Le dictionnaire du jeu de données mg2026 se trouve à l'annexe C et, sous forme résumée, dans le tableau 1.1 du chapitre 1. Durée indicative : 100 minutes.

Dans les fichiers mg2026, les colonnes sont dans l'ordre suivant : A id, B brat, C sex, D varsta, E mediu, F educatie, G fumat, H activitate, I inaltime, J greutate0, K imc0, L talie, M tas0, N tad0, O fc0, P colesterol, Q hdl, R trigliceride, S glicemie0, T glicemie3, U glicemie6, V hba1c0, W findrisc, X diabet_ogtt, Y greutate6, Z tas6, AA scadere5, AB satisfactie. Les données occupent les lignes 2-161 dans le fichier nettoyé et 2-162 dans le fichier brut ; la ligne 1 contient les noms des variables.

Déroulement

L1.1 Création d'un jeu de données médicales de démonstration dans Google Sheets

Le premier exercice n'utilise pas le jeu de données mg2026, mais un petit jeu que vous construisez vous-mêmes, afin de comprendre dès le départ comment on conçoit un tableau de données. Imaginez que vous êtes interne et que vous voulez noter quelques données de base pour 12 patients fictifs vus au cours d'une journée de garde au cabinet de médecine générale.

1. Ouvrez Google Sheets à partir de votre compte universitaire et créez une nouvelle feuille, nommée LP1_Nom_Prenom. Vérifiez les paramètres régionaux : Fichier > Paramètres >

Paramètres régionaux (File > Settings > Locale) doit être réglé sur "France" (ou "Roumanie"). Les intitulés des menus peuvent varier légèrement selon la langue de l'interface.

2. Renommez le premier onglet "donnees" et écrivez dans la ligne 1 les noms des variables, sans espaces et sans diacritiques : id, sex, varsta, inaltime, greutate, imc, tas, glicemie, fumat, educatie.
3. Créez un deuxième onglet, nommé "dictionnaire", avec les colonnes : nom, description, type, unité ou codes, intervalle plausible. Remplissez une ligne pour chaque variable, sur le modèle du tableau 2.2 du chapitre 2. Par exemple : tas ; pression artérielle systolique, mesurée en position assise après 5 minutes de repos ; numérique entier ; mmHg ; 70-260. Pour educatie : ordinal ; 1 collège, 2 lycée ou école professionnelle, 3 universitaire, 4 post-universitaire ; 1-4.
4. Ajoutez des règles de validation, qui empêchent la saisie de valeurs absurdes. Sélectionnez la colonne B (sex), à partir de B2 vers le bas, puis Données > Validation des données > Ajouter une règle (Data > Data validation > Add rule). Choisissez le critère "Drop-down" (liste déroulante) avec les éléments F et M et cochez l'option de rejet des valeurs non valides ("Reject the input"). Pour la colonne G (tas), choisissez le critère "Is between" (est compris entre) 70 et 260 ; pour H (glicemie), entre 40 et 400 ; pour D (inaltime), entre 130 et 210 ; pour E (greutate), entre 30 et 250 ; pour I (fumat), une liste déroulante avec nefumator, fost fumator, fumator ; pour J (educatie), une liste avec 1, 2, 3, 4. Les règles de validation transposent dans la feuille de calcul les intervalles plausibles du dictionnaire.
5. Calculez l'IMC avec une formule, et non à la main. Dans la cellule F2, écrivez =ROUND(E2/(D2/100)^2;1) et recopiez la formule vers le bas en faisant glisser le coin de la cellule. Remarquez le séparateur ";" entre les arguments, propre aux paramètres régionaux France (ou Roumanie).
6. Saisissez les données de 12 patients fictifs, avec les codes P01-P12. Essayez délibérément de saisir une pression de 1540 ou le sexe "f" : la règle de validation doit rejeter la valeur. Si elle ne la rejette pas, vérifiez le paramétrage de la règle.
7. Calculez, sous les données, quelques statistiques simples : le nombre de patients =COUNT(G2:G13), la moyenne de la PAS =AVERAGE(G2:G13), la médiane =MEDIAN(G2:G13), l'écart-type =STDEV(G2:G13), le minimum et le maximum =MIN(G2:G13) et =MAX(G2:G13), le nombre de femmes =COUNTIF(B2:B13;"F").

Ce que vous devez obtenir : une feuille à deux onglets, "donnees" et "dictionnaire", dans laquelle aucune valeur extérieure aux intervalles plausibles ne peut être saisie et dans laquelle l'IMC se recalcule automatiquement lorsque vous modifiez le poids ou la taille.

L1.2 Importation du fichier mg2026_curat.csv

Un fichier CSV (angl. comma-separated values) est un fichier texte dans lequel les valeurs sont séparées par des virgules et les décimales écrites avec un point (73.4). Avec les paramètres régionaux France (ou Roumanie), Google Sheets attend une virgule décimale, et une valeur comme 5.1 peut être mal interprétée, par exemple comme une date (5 janvier). C'est pourquoi l'importation se fait dans un ordre précis.

1. Créez une nouvelle feuille, mg2026_curat. Avant l'importation, réglez temporairement Fichier > Paramètres > Paramètres régionaux sur "États-Unis" (United States).
2. Fichier > Importer > Importer (File > Import > Upload) et choisissez le fichier mg2026_curat.csv téléchargé sur Moodle. Pour "Emplacement de l'importation",

choisissez "Remplacer la feuille de calcul" ou "Insérer de nouvelles feuilles", pour le type de séparateur "Détecter automatiquement" ou "Virgule", et laissez cochée la conversion du texte en nombres.

3. Vérifiez l'importation : la feuille compte 161 lignes (1 ligne d'en-tête et 160 patients) et 28 colonnes (A-AB). Les nombres doivent être alignés à droite ; une colonne alignée à gauche contient du texte.
4. Revenez aux paramètres régionaux "France" (ou "Roumanie"). Les valeurs numériques restent des nombres, mais elles s'affichent désormais avec une virgule (73,4), et les nouvelles formules utilisent le séparateur ";". Figez la première ligne (Affichage > Figer > 1 ligne ; View > Freeze > 1 row), afin que les noms des variables restent visibles lorsque vous faites défiler la feuille.

Ce que vous devez obtenir (autovérification) : =COUNTA(A2:A161) donne 160 ; =AVERAGE(M2:M161) donne 131,13 (moyenne de la PAS) ; =STDEV(M2:M161) donne 13,88 ; =MEDIAN(M2:M161) donne 131 ; =AVERAGE(D2:D161) donne 56,29 (âge moyen) ; =AVERAGE(K2:K161) donne 29,04 (IMC) ; =AVERAGE(S2:S161) donne 105,44 (glycémie) ; =COUNTIF(C2:C161;"F") donne 86 ; =COUNTIF(G2:G161;"fumator") donne 53 ; =COUNTIF(E2:E161;"urban") donne 95. Si la moyenne de la PAS n'est pas de 131,13, n'allez pas plus loin avant d'en avoir trouvé la cause : le plus probable est qu'une partie des valeurs a été importée sous forme de texte ou de dates.

L1.3 Nettoyage du fichier mg2026_brut.csv

Le fichier brut simule ce que vous recevez en réalité d'un collègue qui a saisi les données à partir des fiches d'observation : les mêmes variables, mais avec des erreurs. Il contient 11 erreurs laissées intentionnellement et une ligne en double. Votre tâche consiste à les trouver, à les corriger et à documenter chaque correction. Travaillez toujours sur une copie : conservez un onglet "brut_original" intact et faites les corrections dans un onglet "brut_travail".

1. Importez mg2026_brut.csv dans une nouvelle feuille, selon la même procédure qu'en L1.2 (paramètres régionaux États-Unis lors de l'importation, puis France). Vérifiez : les données occupent les lignes 2-162, soit 161 lignes de données, une de plus que prévu.
2. Ajoutez un onglet "journal", avec les colonnes : n°, id patient, variable, valeur trouvée, mode de détection, valeur corrigée, source de la correction. Chaque modification apportée aux données doit y figurer. C'est le principe de traçabilité : n'importe qui doit pouvoir refaire le chemin du fichier brut au fichier nettoyé.
3. Cherchez les erreurs dans les variables qualitatives. Dans une cellule libre, écrivez =SORT(UNIQUE(C2:C162)) pour le sexe, puis faites de même pour E (mediu) et G (fumat). La fonction UNIQUE affiche chaque valeur distincte une seule fois et distingue majuscules et minuscules. Vous devriez voir, au lieu de deux catégories pour le sexe, quatre : F, f, Feminin, M ; pour mediu, urban, Urban, rural ; pour fumat, en plus des trois catégories correctes, également "Fumator " (avec une majuscule et une espace finale). Attention : la fonction COUNTIF ne distingue pas les majuscules des minuscules, de sorte que =COUNTIF(C2:C162;"F") compte aussi la valeur "f" sans vous avertir ; JASP, en revanche, les traiterait comme deux catégories différentes. Pour les espaces invisibles, la fonction =LEN(G85) indique un caractère de plus que la longueur attendue.
4. Cherchez les erreurs dans les variables numériques. Pour chaque colonne numérique, calculez =MIN(...), =MAX(...), =COUNT(...) et =COUNTBLANK(...). Comparez le minimum et le maximum à l'intervalle plausible du dictionnaire. Vous trouverez une pression systolique maximale de 1540 mmHg (colonne M), une taille minimale de 1,72

(colonne I) et un poids minimal de 7,26 kg (colonne J). La fonction COUNT ne compte que les valeurs numériques : pour la glycémie initiale (colonne S), elle donne 160 au lieu de 161, et il en va de même pour l'HbA1c (colonne V), signe qu'il existe dans chaque cas une valeur stockée sous forme de texte ; =ISNUMBER(S65) donne FALSE pour la cellule contenant "87 mg/dl", et =ISNUMBER(V90) donne FALSE pour l'HbA1c "5,2", écrite avec une virgule dans un fichier où toutes les autres décimales sont notées avec un point. La fonction COUNTBLANK donne 1 pour l'âge (colonne D) et 1 pour la glycémie à 6 mois (colonne U).

5. Cherchez les doublons. =COUNTUNIQUE(A2:A162) donne 160, alors qu'il y a 161 lignes : un code patient apparaît deux fois. Dans une colonne auxiliaire, écrivez en AC2 la formule =COUNTIF(A$2:A$162;A2) et recopiez-la vers le bas ; les lignes ayant la valeur 2 sont les doublons. Vous trouverez le patient P091 deux fois, avec des valeurs identiques dans toutes les colonnes. Supprimez la deuxième ligne (la dernière ligne du fichier) ou utilisez Données > Nettoyage des données > Supprimer les doublons (Data > Data cleanup > Remove duplicates).
6. Corrigez les erreurs en suivant la règle du chapitre 1 : une valeur suspecte ne se corrige pas à l'intuition, elle se vérifie dans le document source. Dans cette séance, le rôle du document source est tenu par le fichier mg2026_curat.csv, dans lequel les valeurs ont été reprises des fiches d'observation. Le tableau L1.1 résume les 12 problèmes et leurs corrections.

Tableau L1.1. Erreurs de mg2026_brut.csv, mode de détection et correction.

N°	Patient	Variable	Valeur trouvée	Détection	Correction et source
1	P005	sex	f	UNIQUE	F (harmonisation ; confirmé dans la fiche source)
2	P021	sex	Feminin	UNIQUE	F (harmonisation ; confirmé dans la fiche source)
3	P034	tas0	1540	MAX	154 (chiffre en trop ; confirmé dans la fiche source)
4	P042	varsta	manquante	COUNTBLANK	70 (fiche source)
5	P057	inaltime	1,72	MIN	172 (mètres convertis en centimètres ; vérifié à partir de l'IMC et du poids)
6	P064	glicemie0	"87 mg/dl"	COUNT, ISNUMBER	87 (suppression de l'unité ; confirmé dans la fiche source)
7	P084	fumat	"Fumator "	UNIQUE, LEN	fumator (TRIM et minuscules ; confirmé dans la fiche source)
8	P089	hba1c0	"5,2" (texte, avec virgule)	COUNT, ISNUMBER	5,2 en tant que nombre (confirmé dans la fiche source)
9	P102	greutate0	7,26	MIN	72,6 (virgule

N°	Patient	Variable	Valeur trouvée	Détection	Correction et source
					déplacée ; vérifié à partir de l'IMC et de la taille)
10	P121	mediu	Urban	UNIQUE	urban (harmonisation)
11	P134	glicemie6	manquante	COUNTBLANK	135 (fiche source)
12	P091	toutes	ligne en double	COUNTUNIQUE, COUNTIF	suppression de la deuxième ligne

7. Discutez avec votre voisin de ce que vous auriez corrigé si vous n'aviez pas disposé du document source. Chaque erreur du fichier est une déformation de la valeur vraie, et pour la plupart d'entre elles il existe une hypothèse très probable : "f" et "Feminin" signifient F, "Urban" signifie urban, "Fumator " devient fumator après suppression de l'espace (fonction TRIM) et passage en minuscules (fonction LOWER), "87 mg/dl" devient 87 après suppression de l'unité, "5,2" devient le nombre 5,2. Pour deux valeurs numériques, l'hypothèse peut être vérifiée de façon indépendante, car le fichier contient une variable dérivée redondante, l'IMC : pour P057, la taille est de $\sqrt{\frac{77,7}{26,3}} = \sqrt{2,954} = 1,719m$, soit 172 cm, exactement la valeur 1,72 convertie de mètres en centimètres ; pour P102, le poids est de 27,0 x 1,64² = 27,0 x 2,6896 = 72,6 kg, donc 7,26 est la valeur correcte avec la virgule déplacée. La valeur 1540 est très probablement 154 avec un chiffre doublé, mais elle aurait tout aussi bien pu être 150 ou 140 mal saisis ; rien dans le fichier ne permet de le vérifier, et la confirmation dans la fiche source est donc ici obligatoire. Les deux valeurs manquantes (l'âge du patient P042 et la glycémie à 6 mois du patient P134) ne peuvent être déduites d'aucune manière : seul le document source permet de les compléter (respectivement 70 ans et 135 mg/dL). Lorsque la source n'est pas disponible, la valeur est laissée manquante et cela est noté dans le journal ; une valeur inventée est plus dangereuse qu'une valeur manquante, même lorsque l'hypothèse semble évidente.
8. Après les corrections, vérifiez le résultat. L'onglet "brut_travail" doit contenir 160 lignes de données, et les statistiques doivent coïncider avec celles du fichier nettoyé : =AVERAGE(M2:M161) donne 131,13 ; =COUNT(S2:S161) donne 160 ; =MIN(I2:I161) donne 145 cm ; =MAX(M2:M161) donne 169 mmHg ; =COUNTBLANK(D2:D161) donne 0 ; =SORT(UNIQUE(C2:C161)) n'affiche que F et M. Si toutes les vérifications sont satisfaites, votre jeu de données nettoyé est identique à mg2026_curat.csv.

Observez à quel point les erreurs auraient modifié les résultats si elles n'avaient pas été trouvées. Sur le fichier brut non nettoyé (161 lignes, avec la valeur 1540 et le doublon), =AVERAGE(M2:M162) donne 139,70 mmHg au lieu de 131,13, et =STDEV(M2:M162) donne 111,89 mmHg au lieu de 13,88 : un seul chiffre mal saisi a déplacé la moyenne du groupe à la limite de l'hypertension et multiplié l'écart-type par près de huit. La médiane, =MEDIAN(M2:M162), reste à 131 mmHg, exactement comme dans le fichier nettoyé, ce qui illustre la robustesse de la médiane évoquée au chapitre 2. Lorsque la moyenne et la médiane d'une variable symétrique diffèrent autant et que l'écart-type est du même ordre que la moyenne, le premier suspect est une erreur de saisie.

L1.4 Découverte de l'interface de JASP

JASP est le logiciel dans lequel vous ferez la plupart des analyses du semestre. L'interface comporte trois zones : les données (à gauche, comme une feuille de calcul), le panneau d'options

de l'analyse en cours (au milieu) et les résultats (à droite), qui se mettent à jour dès que vous modifiez une option.

1. Ouvrez JASP. Dans le menu principal (le bouton à trois traits en haut à gauche), choisissez Open > Computer > Browse et sélectionnez mg2026_curat.csv. Les données apparaissent à gauche.
2. Vérifiez le type de chaque variable. À côté du nom de chaque colonne apparaît une icône : une règle pour les variables numériques (Scale), trois barres en escalier pour les variables ordinales (Ordinal) et trois cercles pour les variables nominales (Nominal). JASP devine le type lors de l'importation, mais pas toujours correctement : educatie, activitate et satisfactie sont importées comme des nombres, bien qu'elles soient ordinales ; diabet_ogtt et scadere5 sont importées comme des nombres, bien qu'elles soient nominales dichotomiques. Cliquez sur l'icône et changez le type. Cette étape n'est pas cosmétique : JASP propose des analyses différentes selon le type déclaré de la variable.
3. Choisissez Descriptives > Descriptive Statistics. Placez dans la zone Variables les variables varsta, imc0, tas0 et glicemie0. Dans la section Statistics, cochez, en plus des options par défaut (Valid, Missing, Mean, Std. deviation, Minimum, Maximum), les options Median, Quartiles, Skewness, Kurtosis et S.E. mean.
4. Comparez les résultats avec ceux de Google Sheets et du tableau 2.12 du chapitre 2. Vous devez obtenir : varsta 56,29 ± 8,77 ans, médiane 56 ; imc_0 $29{,}04 \pm \frac{3{,}91\ \text{kg}}{m^2}$, médiane 28,25 ; tas_0 $131{,}13 \pm 13{,}88$ mmHg, médiane 131, erreur standard 1,097 ; glicemie_0 $105{,}44 \pm 15{,}76$ mg/dL, médiane 104,5. JASP affiche les résultats avec un point décimal ; dans les rapports, vous les retranscrivez avec une virgule.
5. Placez la variable sex dans la zone Split. Les résultats s'affichent séparément pour les femmes et les hommes ; pour la PAS, vous devez voir 131,37 ± 14,47 mmHg chez les femmes (n = 86) et 130,85 ± 13,25 mmHg chez les hommes (n = 74).
6. Pour une variable nominale, par exemple fumat, cochez l'option Frequency tables (dans la section Tables ou directement sous la liste des variables, selon la version). Le tableau de fréquences doit indiquer 74 non-fumeurs, 33 anciens fumeurs et 53 fumeurs.
7. Enregistrez l'analyse sous forme de fichier JASP (File > Save As, LP1_Nom_Prenom.jasp). Le fichier JASP conserve ensemble les données, les options et les résultats, ce qui rend l'analyse reproductible. Les résultats peuvent aussi être exportés en PDF ou en HTML (File > Export Results).

L1.5 Découverte de l'interface de MedCalc

MedCalc possède une interface de type feuille de calcul, avec des menus pour les analyses et les graphiques. C'est le logiciel de prédilection pour les courbes ROC et pour l'analyse des tests diagnostiques, que vous utiliserez lors des séances 11-12.

1. Ouvrez MedCalc et importez le fichier mg2026_curat.csv par File > Open, en choisissant le type de fichier texte/CSV. Vérifiez dans la fenêtre d'importation que le séparateur est la virgule et que le symbole décimal est le point.
2. Choisissez Statistics > Summary statistics, sélectionnez la variable tas0 et lancez l'analyse. MedCalc affiche le nombre d'observations, la moyenne, l'intervalle de confiance à 95% de la moyenne (128,96-133,30 mmHg, que vous étudierez au chapitre 4), l'écart-type, l'erreur standard, la médiane, les percentiles et, en option, les tests de normalité.
3. Choisissez Graphs > Box-and-whisker plot pour tas0, éventuellement avec sex comme variable de regroupement. Comparez le graphique à la figure 2.3D du chapitre 2.

4. Si aucune licence MedCalc n'est disponible sur votre ordinateur, ouvrez les calculateurs en ligne gratuits du site de l'éditeur. Essayez le calculateur de l'intervalle de confiance d'une proportion : saisissez 29 patients diabétiques sur 160 et notez le résultat ; vous l'interpréterez au chapitre 4 (la méthode de Wilson donne 12,9-24,8%).

L1.6 Epi Info, en bref

Epi Info 7 (dernière version, 7.2.7) a été développé par le CDC pour l'épidémiologie de terrain et fonctionne entièrement sans connexion internet. Le CDC a arrêté le développement et l'assistance du logiciel en septembre 2025 ; le logiciel reste utilisable, mais ne reçoit plus de mises à jour. Il comporte trois composants principaux : Form Designer (conception des formulaires de recueil des données), Enter (saisie des données) et Analysis (analyse, en deux variantes : Classic Analysis, fondée sur des commandes, et Visual Dashboard, fondée sur des menus).

1. Ouvrez Epi Info et choisissez Classic Analysis. Avec la commande READ (bouton Read dans la liste des commandes), importez le fichier mg2026_curat.csv, en choisissant le type de source texte/CSV.
2. Exécutez la commande FREQ pour la variable fumat et la commande MEANS pour la variable tas0. Vérifiez que vous obtenez les mêmes fréquences (74, 33, 53) et la même moyenne (131,13 mmHg) que dans les autres logiciels.
3. Remarquez qu'Epi Info affiche aussi le programme généré par les commandes, qui peut être enregistré et exécuté de nouveau. Vous reviendrez à Epi Info lors des séances consacrées à l'analyse du risque, où les tableaux 2x2 avec risque relatif et rapport de cotes s'obtiennent directement.

L1.7 Dépôt des résultats sur Moodle

1. Téléchargez la feuille Google Sheets sous forme de fichier Excel : Fichier > Télécharger > Microsoft Excel (.xlsx). Le fichier doit contenir les onglets : donnees (le jeu de démonstration), dictionnaire, brut_original, brut_travail et journal.
2. Préparez un document court (une page, PDF) contenant : le tableau des 12 problèmes trouvés et corrigés, les statistiques de vérification (n, moyenne et écart-type de la PAS avant et après nettoyage) et deux ou trois phrases sur ce que vous a appris l'exercice de nettoyage.
3. Déposez sur Moodle, dans l'activité LP1, les fichiers LP1_Nom_Prenom.xlsx, LP1_Nom_Prenom.jasp et LP1_Nom_Prenom.pdf, avant la date limite affichée sur la plateforme [52].

Exercices supplémentaires

1. Sur le fichier BD_Studenti disponible sur Moodle (la base de données des étudiants utilisée les années précédentes dans la discipline) : déterminez la taille de l'échantillon (N), classez chaque variable selon son type, rédigez un dictionnaire des données complet (nom, description, type, unité ou codes, intervalle plausible) et vérifiez à l'aide des fonctions MIN, MAX, COUNT, COUNTBLANK et UNIQUE si le fichier contient des valeurs manquantes, des valeurs hors des intervalles plausibles ou des codages incohérents. Ajoutez une colonne calculée avec l'IMC de chaque étudiant. Aucun résultat numérique n'est fourni ; la vérification se fait lors de la séance suivante.
2. Problème avec données dans l'énoncé. Un collègue a saisi les données de huit patients d'une étude sur l'hypertension dans le tableau ci-dessous. Trouvez les erreurs, proposez un mode de vérification pour chacune et calculez la moyenne de la PAS avant et après

correction, en supposant que la fiche source confirme pour P4 la valeur de 152 mmHg et que la valeur manquante du patient P7 ne peut pas être récupérée.

Tableau L1.2. Données saisies pour huit patients (exercice 2).

id	sex	varsta	tas	fumat
P1	F	58	138	nefumator
P2	M	61	145	fumator
P3	m	49	132	nefumator
P4	F	66	1520	fost fumator
P5	F	55	128	Nefumator
P6	M	72	160	fumator
P7	F	63		nefumator
P2	M	61	145	fumator

Solution. Les erreurs sont les suivantes : (1) P3, sexe "m" au lieu de "M" (codage incohérent ; détecté avec UNIQUE ; harmonisé en M après confirmation) ; (2) P4, PAS de 1520 mmHg, biologiquement impossible (détectée avec MAX ; corrigée en 152 d'après la fiche source) ; (3) P5, fumat "Nefumator" avec une majuscule (UNIQUE ; harmonisé en nefumator) ; (4) P7, PAS manquante (COUNTBLANK ; reste manquante et cela est noté dans le journal) ; (5) la ligne P2 en double (COUNTIF sur id ; suppression de la deuxième ligne). Avant correction, la moyenne de la PAS calculée par =AVERAGE sur les 7 valeurs numériques (y compris le doublon et la valeur 1520) est de (138 + 145 + 132 + 1520 + 128 + 160 + 145)/7 = 2 368/7 = 338,3 mmHg, une valeur absurde. Après correction, il reste 7 patients distincts, dont 6 avec une PAS connue : (138 + 145 + 132 + 152 + 128 + 160)/6 = 855/6 = 142,5 mmHg. La fonction AVERAGE ignore la cellule vide, mais le nombre de patients analysés (n = 6 sur 7) doit être rapporté explicitement.

3. Sur le fichier mg2026_curat.csv, créez une nouvelle colonne, "obez", avec la formule =IF(K2>=30;1;0), et comptez les patients obèses avec =SUM(...). Vous devez obtenir 54. Créez ensuite une colonne "tas_crescuta" avec =IF(M2>=140;1;0) ; vous devez obtenir 41 patients. Discutez de l'information perdue du fait de ces dichotomisations (chapitre 2, section 2.1.4).

Ce qu'il faut rendre et barème indicatif

Il faut rendre, sur Moodle, les trois fichiers décrits en L1.7. Barème indicatif, sur 10 points : jeu de démonstration correctement structuré (une ligne par patient, noms de variables sans espaces) et IMC calculé par une formule, 1,5 point ; dictionnaire des données complet, 1,5 point ; règles de validation fonctionnelles, 1 point ; importation correcte du fichier nettoyé, avec les statistiques de vérification, 1 point ; identification des 11 erreurs et du doublon, 2 points (un point pour chaque moitié de la liste) ; journal des modifications, avec la source de chaque correction, 1 point ; fichier JASP avec des types de variables correctement définis et les statistiques descriptives, 1 point ; document PDF avec l'interprétation, 1 point. Un point est retiré si les statistiques finales ne coïncident pas avec celles du fichier nettoyé (par exemple, une moyenne de la PAS différente de 131,13 mmHg).

Variante hors ligne

Si le réseau ne fonctionne pas, la séance peut être réalisée entièrement sur l'ordinateur local. Le jeu de démonstration, le dictionnaire, la validation et le nettoyage se font dans LibreOffice Calc ou Microsoft Excel, avec les mêmes fonctions (dans la version roumaine d'Excel, les noms des fonctions sont traduits, par exemple MEDIE au lieu de AVERAGE, et le séparateur reste ";") ; la validation des données se trouve dans le menu Date > Validare date. JASP et Epi Info fonctionnent sans connexion internet une fois installés. Les fichiers mg2026 se copient à partir de

la clé USB de l'enseignant ou du serveur local du laboratoire. Le dépôt sur Moodle se fait ultérieurement, dans le délai affiché.

Travaux pratiques 2. Description et organisation des données : le questionnaire médical, les tableaux de fréquences et les graphiques

Objectif

Dans cette séance, vous parcourez le chemin complet d'une variable, de la question posée au patient jusqu'au tableau et au graphique d'un rapport. Vous construisez d'abord un court questionnaire médical dans Google Forms et vous voyez comment le type de chaque question détermine le type de la variable que vous analyserez. Ensuite, sur le jeu de données mg2026, vous construisez des tableaux de fréquences, calculez les indicateurs descriptifs et tracez les graphiques appropriés, d'abord dans Google Sheets, puis dans JASP. Tous les résultats attendus sont donnés dans le texte, afin que vous puissiez vous vérifier vous-mêmes.

À la fin, l'étudiant doit être capable :

- de concevoir un court questionnaire dans Google Forms, en choisissant pour chaque question le type qui produit la variable souhaitée et les règles de validation appropriées ;
- de relier le formulaire à une feuille Google Sheets et de recoder les réponses pour l'analyse ;
- de construire des tableaux de fréquences absolues, relatives et cumulées avec COUNTIF, COUNTIFS, FREQUENCY et avec des tableaux croisés dynamiques ;
- de calculer dans Google Sheets et dans JASP la moyenne, la médiane, le mode, l'écart-type, les quartiles, le coefficient de variation, l'asymétrie et l'aplatissement ;
- de construire l'histogramme, la boîte à moustaches, le diagramme en barres et le diagramme circulaire et d'interpréter la forme de la distribution.

Outils

Logiciels : Google Forms et Google Sheets (compte Google de l'université) [43-46], JASP 0.98.1 [47]. Fichiers : mg2026_curat.csv (ou la feuille nettoyée lors des travaux pratiques 1), BD_Studenti pour les exercices supplémentaires. Durée indicative : 100 minutes (environ 30 minutes pour le questionnaire et 70 pour l'analyse descriptive).

Les colonnes du jeu de données mg2026 sont celles des travaux pratiques 1 : C sex, D varsta, E mediu, F educatie, G fumat, H activitate, K imc0, M tas0, R trigliceride, S glicemie0, AB satisfactie ; les données occupent les lignes 2-161.

Déroulement

L2.1 Conception d'un court questionnaire médical dans Google Forms

Le scénario : un cabinet de médecine générale veut savoir à quelle fréquence ses patients hypertendus mesurent leur pression artérielle à domicile et dans quelle mesure ils sont satisfaits du contrôle de leur maladie. Le questionnaire est anonyme, il est rempli en salle d'attente sur la tablette du cabinet et comporte au plus 10 questions.

1. Ouvrez Google Forms à partir de votre compte universitaire et créez un nouveau formulaire, nommé "LP2 Automesure de la pression artérielle". Dans la description, indiquez le but du questionnaire, le fait qu'il est anonyme et volontaire, et la durée nécessaire pour le remplir.

2. La première question est la question de consentement : "J'accepte de remplir ce questionnaire anonyme, et mes réponses peuvent être utilisées à des fins statistiques", de type choix multiples, avec les options "Oui" et "Non", obligatoire. Utilisez le branchement par sections (Accéder à la section en fonction de la réponse) de sorte que la réponse "Non" renvoie directement à la fin du formulaire [148]. Sans consentement, rien n'est recueilli.
3. Ajoutez les questions ci-dessous, en choisissant le type indiqué. Le tableau L2.1 montre le lien entre le type de question dans Google Forms et le type de la variable qui en résultera. C'est le lien le plus important de cette séance : le type de la variable se décide au moment où vous rédigez la question, et non au moment où vous faites l'analyse.

Tableau L2.1. Types de questions dans Google Forms et variables qui en résultent.

Question	Type dans Google Forms	Validation	Variable obtenue
Votre âge (années révolues)	Réponse courte	Nombre, entre 18 et 100	numérique discrète (âge en années)
Sexe	Choix multiples : féminin, masculin	obligatoire	nominale dichotomique
Milieu de résidence	Choix multiples : urbain, rural	obligatoire	nominale dichotomique
Combien de fois avez-vous mesuré votre pression artérielle à domicile au cours de la dernière semaine ?	Réponse courte	Nombre entier, entre 0 et 50	numérique discrète
Dernière valeur de pression systolique mesurée à domicile (mmHg), si vous la connaissez	Réponse courte	Nombre, entre 70 et 260 ; facultative	numérique continue, valeurs manquantes autorisées
Quel appareil utilisez-vous ?	Choix multiples : au bras, au poignet, je n'ai pas d'appareil	obligatoire	nominale
Qu'est-ce qui vous empêche de la mesurer plus souvent ?	Cases à cocher (plusieurs réponses)	-	plusieurs variables dichotomiques (une pour chaque option)
Dans quelle mesure êtes-vous satisfait du contrôle de votre pression artérielle ?	Échelle linéaire 1-5 (1 très insatisfait, 5 très satisfait)	obligatoire	ordinale
Niveau d'études	Liste déroulante : collège, lycée ou école professionnelle, universitaire, post-universitaire	obligatoire	ordinale

4. Définissez les règles de validation. Pour les questions numériques, cliquez sur les trois points dans le coin de la question, choisissez Validation de la réponse et réglez "Nombre", "Entre", avec les limites du tableau, en ajoutant un message d'erreur clair ("Veuillez saisir un nombre entre 70 et 260") [45]. La validation empêche exactement les erreurs que vous avez nettoyées lors des travaux pratiques 1 : le texte "87 mg/dl" au lieu de 87, une pression de 1540, un poids de 7,26 kg, un âge de 5 ans.
5. Ne demandez pas de données d'identification : ni nom, ni adresse e-mail (dans les Paramètres, la collecte des adresses e-mail doit être désactivée), ni date de naissance complète. L'âge en années suffit. C'est l'application directe du principe de minimisation des données (chapitre 1, section 1.8.3).

6. Testez le formulaire : remplissez-le 5 ou 6 fois, comme si vous étiez des patients différents, avec des réponses variées, dont une qui refuse le consentement et une qui ne connaît pas sa pression artérielle. Demandez aussi à votre voisin de le remplir 2 ou 3 fois depuis son téléphone.

L2.2 Liaison du formulaire à Google Sheets et recodage des réponses

1. Dans le formulaire, ouvrez l'onglet Réponses et choisissez Associer à Sheets (Link to Sheets), avec l'option de création d'une nouvelle feuille de calcul [46]. Google crée une feuille dans laquelle chaque réponse est une ligne et chaque question une colonne ; la première colonne contient l'horodatage de l'envoi. La structure est exactement celle recommandée au chapitre 2 : une ligne par répondant, une colonne par variable [147].
2. Observez comment apparaissent les différents types de questions. Les réponses aux choix multiples apparaissent sous forme de texte ("féminin"), l'échelle linéaire sous forme de nombre (1-5), les cases à cocher sous forme d'un texte reprenant toutes les options cochées, séparées par des virgules, dans la même cellule. Ce dernier format n'est pas adapté à l'analyse : chaque option doit être transformée en une colonne distincte, avec 1 (cochée) et 0 (non cochée). Pour l'option "je n'ai pas le temps", par exemple, utilisez la formule =IF(ISNUMBER(SEARCH("je n'ai pas le temps";H2));1;0).
3. Recodez les variables ordinales saisies sous forme de texte. Pour le niveau d'études, dans une nouvelle colonne : =IFS(J2="collège";1;J2="lycée ou école professionnelle";2;J2="universitaire";3;J2="post-universitaire";4). L'analyse ordinale a besoin de codes numériques dans le bon ordre ; le texte serait trié par ordre alphabétique, et si "collège" se retrouvait avant "universitaire", ce ne serait que par hasard.
4. Supprimez les lignes des répondants qui ont refusé le consentement (si le branchement a fonctionné, les autres colonnes sont vides) et renommez les colonnes avec des noms courts, sans espaces : varsta, sex, mediu, nr_masurari, tas_casa, aparat, satisfactie, educatie. Ajoutez un onglet "dictionnaire".

L2.3 Tableaux de fréquences dans Google Sheets sur le jeu de données mg2026

Pour l'analyse descriptive, travaillez sur le jeu de données mg2026, qui compte 160 patients et permet d'obtenir des résultats ayant du sens. Ouvrez la feuille nettoyée lors des travaux pratiques 1 ou importez de nouveau mg2026_curat.csv (selon la procédure de L1.2) et travaillez dans un nouvel onglet, "frequences".

1. Variable nominale : le statut tabagique. Écrivez en A2:A4 les catégories nefumator, fost fumator, fumator. En B2, écrivez =COUNTIF(donnees!$G$2:$G$161;A2) et recopiez vers le bas. En C2, la fréquence relative : =B2/SUM($B$2:$B$4)*100. Vous devez obtenir 74 (46,3%), 33 (20,6%) et 53 (33,1%). Le nom de l'onglet contenant les données ("donnees") est à remplacer par le nom de votre onglet.
2. Variable ordinale : la satisfaction (colonne AB). En A7:A11, écrivez les valeurs 1-5. Fréquence absolue : =COUNTIF(donnees!$AB$2:$AB$161;A7). Fréquence relative : =B7/160*100. Fréquence cumulée absolue : en D7 =B7, en D8 =D7+B8, recopié vers le bas. Fréquence cumulée relative : =D7/160*100. Le résultat doit correspondre au tableau 2.4 du chapitre 2 : fréquences 4, 19, 54, 55, 28 ; fréquences relatives 2,5 ; 11,9 ; 33,8 ; 34,4 ; 17,5% ; fréquences cumulées 2,5 ; 14,4 ; 48,1 ; 82,5 ; 100,0%. La fréquence cumulée dépasse 50% à la valeur 4, la médiane est donc 4.
3. Même démarche pour le niveau d'études (colonne F) : 20, 57, 64, 19 patients (12,5 ; 35,6 ; 40,0 ; 11,9%), avec des fréquences cumulées de 12,5 ; 48,1 ; 88,1 ; 100,0% et une

médiane de 3. Pour l'activité physique (colonne H) : 21, 42, 60, 34, 3 patients, avec une médiane de 3.

4. Variable continue regroupée en classes : la PAS (colonne M). Calculez d'abord le nombre de classes selon Sturges, =1+3,322*LOG10(160), qui donne 8,32, et la largeur =(MAX(donnees!M2:M161)-MIN(donnees!M2:M161))/8, qui donne 8,4 ; arrondissez la largeur à 10 mmHg. En A14:A20, écrivez les limites inférieures 100, 110,..., 160. La fréquence de chaque classe s'obtient avec COUNTIFS : =COUNTIFS(donnees!$M$2:$M$161;">="&A14;donnees!$M$2:$M$161;"<"&A14+10) . Vous devez obtenir 11, 21, 35, 52, 25, 13, 3, c'est-à-dire le tableau 2.5 du chapitre 2.
5. Les mêmes fréquences s'obtiennent avec la fonction FREQUENCY, qui reçoit les données et les limites supérieures des classes. Écrivez en F14:F19 les limites supérieures 109, 119, 129, 139, 149, 159, puis, dans une seule cellule, =FREQUENCY(donnees!M2:M161;F14:F19). La fonction renvoie 7 valeurs (la dernière pour les valeurs supérieures à 159) et les inscrit automatiquement dans les cellules situées en dessous. Vérifiez que la somme des fréquences est égale à 160.
6. Les classes d'IMC utilisées en clinique ont des largeurs inégales : moins de 25 (poids normal), 25-29,9 (surpoids), 30-34,9 (obésité de grade I), 35-39,9 (grade II), ≥ 40 (grade III). Construisez ce tableau avec COUNTIFS sur la colonne K. Résultat attendu : 18, 88, 43, 8 et 3 patients ; au total, 43 + 8 + 3 = 54 obèses (33,8%).

L2.4 Tableaux croisés dynamiques et tableaux de contingence

Un tableau croisé dynamique résume automatiquement les données selon une ou plusieurs variables catégorielles, sans formules [44].

1. Sélectionnez les données (A1:AB161) et choisissez Insertion > Tableau croisé dynamique (Insert > Pivot table), en créant un nouvel onglet.
2. Dans l'éditeur du tableau croisé dynamique, ajoutez en Lignes la variable fumat, en Colonnes la variable sex et en Valeurs la variable id, résumée par COUNTA. Vous devez obtenir le tableau de contingence sex x fumat : femmes, 16 anciennes fumeuses, 30 fumeuses, 40 non-fumeuses ; hommes, 17, 23 et 34 ; totaux 33, 53, 74 et 160.
3. Modifiez l'affichage des valeurs en pourcentages du total de la colonne (option "Afficher sous forme de", Show as > % of column). Comparez la proportion de fumeurs actifs chez les femmes (30/86 = 34,9%) et chez les hommes (23/74 = 31,1%).
4. Construisez un deuxième tableau croisé dynamique, brat x satisfactie. Résultat attendu : bras A, satisfaction 1-5 : 1, 13, 28, 9, 2 ; bras B : 3, 3, 14, 24, 9 ; bras C : 0, 3, 12, 22, 17. Avant même tout test statistique, le tableau montre que les patients du bras C sont plus satisfaits : 39 sur 54 (72,2%) ont choisi 4 ou 5, contre 11 sur 53 (20,8%) dans le bras A. Le test approprié pour cette comparaison, Mann-Whitney ou Kruskal-Wallis, sera appliqué lors des travaux pratiques 7.
5. Le tableau croisé dynamique peut aussi calculer des moyennes : placez brat en Lignes et tas0 en Valeurs, résumée par AVERAGE. Vous devez obtenir 131,53 (A), 132,79 (B) et 129,11 (C) mmHg.

L2.5 Statistiques descriptives dans Google Sheets

Dans un nouvel onglet, "descriptives", construisez un tableau avec les variables en colonnes et les indicateurs en lignes. Les formules pour la PAS (colonne M) sont les suivantes :

Tableau L2.2. Formules Google Sheets pour les statistiques descriptives de la PAS et résultats attendus.

Indicateur	Formule (paramètres France ou Roumanie)	Résultat attendu
n	=COUNT(donnees!M2:M161)	160
Moyenne	=AVERAGE(donnees!M2:M161)	131,13
Médiane	=MEDIAN(donnees!M2:M161)	131
Mode	=MODE(donnees!M2:M161)	138
Écart-type	=STDEV(donnees!M2:M161)	13,88
Variance	=VAR(donnees!M2:M161)	192,66
Erreur standard	=STDEV(donnees!M2:M161)/SQRT(COUNT(donnees!M2:M161))	1,097
Coefficient de variation (%)	=STDEV(donnees!M2:M161)/AVERAGE(donnees!M2:M161)*100	10,6
Minimum, maximum	=MIN(donnees!M2:M161), =MAX(donnees!M2:M161)	102, 169
Q1, Q3	=QUARTILE(donnees!M2:M161;1), =QUARTILE(donnees!M2:M161;3)	122, 140
10e et 90e percentiles	=PERCENTILE(donnees!M2:M161; 0,1), =PERCENTILE(donnees!M2:M161; 0,9)	112; 148,2
Asymétrie	=SKEW(donnees!M2:M161)	0,01
Aplatissement	=KURT(donnees!M2:M161)	-0,10

Remarquez l'argument 0,1 de la fonction PERCENTILE : avec les paramètres France (ou Roumanie), la décimale s'écrit avec une virgule et les arguments sont séparés par des points-virgules. Recopiez les formules pour les autres variables, en changeant la lettre de la colonne. Résultats attendus :

Tableau L2.3. Résultats attendus pour d'autres variables de mg2026.

Variable (colonne)	Moyenne	ÉT	Médiane	Q1	Q3	Asymétrie	Aplatissement
varsta (D)	56,29	8,77	56	50,75	62	0,01	-0,09
imc0 (K)	29,04	3,91	28,25	26,2	31,4	1,42	3,68
glicemie0 (S)	105,44	15,76	104,5	93	116	0,37	-0,23
trigliceride (R)	161,18	76,30	151,5	107	203,25	0,95	1,12

Calculez aussi la moyenne géométrique des triglycérides, =GEOMEAN(donnees!R2:R161), qui doit donner 144,0 mg/dL, et comparez-la à la moyenne arithmétique (161,18) et à la médiane (151,5). Calculez les limites des valeurs extrêmes de l'IMC : =QUARTILE(donnees!K2:K161;3)+1,5*(QUARTILE(donnees!K2:K161;3)-QUARTILE(donnees!K2:K161;1)) donne 39,2 kg/m², et =COUNTIF(donnees!K2:K161;">"&39,2) donne 3 patients.

L2.6 Graphiques dans Google Sheets

1. Diagramme en barres pour fumat. Sélectionnez le tableau de fréquences de L2.3, étape 1 (les catégories et les fréquences absolues), puis Insertion > Graphique. Dans l'éditeur de graphique, choisissez le type "Graphique à colonnes". Dans l'onglet Personnaliser, ajoutez le titre ("Statut tabagique, mg2026, n = 160"), les titres des axes et les étiquettes de données. Vérifiez que l'axe vertical commence à 0.
2. Diagramme circulaire pour le niveau d'études. Sélectionnez le tableau de fréquences du niveau d'études (avec les libellés des catégories en texte, et non les codes 1-4) et choisissez le type "Graphique à secteurs". Affichez les pourcentages sur les secteurs : 12,5 ; 35,6 ; 40,0 ; 11,9%. N'utilisez pas la variante 3D.

3. Histogramme de la glycémie. Sélectionnez la colonne glicemie0 (S1:S161) et choisissez le type "Histogramme". Dans Personnaliser > Histogramme, réglez la taille des intervalles (bucket size) sur 10. Forme attendue : un large pic entre 90 et 120 mg/dL (34 patients dans chacune des classes 90-99, 100-109 et 110-119), avec une courte queue vers la droite, jusqu'à 154 mg/dL. Essayez aussi un intervalle de 5, puis de 20, et observez comment l'aspect change.
4. La boîte à moustaches n'existe pas comme type de graphique standard dans Google Sheets. Vous la construirez dans JASP, à l'étape suivante. Vous pouvez en revanche afficher le résumé en cinq nombres de la PAS dans un petit tableau : 102, 122, 131, 140, 169.

L2.7 Analyse descriptive et graphiques dans JASP

1. Ouvrez mg2026_curat.csv dans JASP et vérifiez les types des variables, comme lors des travaux pratiques 1 : educatie, activitate et satisfactie ordinales ; sex, mediu, fumat, brat, diabet_ogtt nominales.
2. Descriptives > Descriptive Statistics, avec les variables varsta, imc0, tas0, glicemie0, trigliceride. Cochez Mean, Std. deviation, Median, Quartiles, Minimum, Maximum, Skewness, Kurtosis, Shapiro-Wilk test (vous l'interpréterez au chapitre 3). Comparez avec le tableau L2.3 ; les valeurs doivent coïncider.
3. Pour les variables fumat, educatie et satisfactie, cochez Frequency tables. JASP affiche les fréquences, les pourcentages, les pourcentages valides (qui excluent les valeurs manquantes) et les pourcentages cumulés, exactement comme dans le tableau 2.4.
4. Dans la section Basic plots (ou Plots), cochez Distribution plots pour imc0 et trigliceride. Comparez les histogrammes à la figure 2.2 du chapitre 2 : les deux distributions ont une queue vers la droite.
5. Pour la boîte à moustaches, placez tas0 dans Variables et sex dans Split, puis cochez Boxplots (et éventuellement les options Boxplot element, Violin element et Jitter element, qui montrent aussi les points individuels). Le résultat doit ressembler à la figure 2.3D : médiane de 131 mmHg pour les deux sexes, boîtes presque superposées. Dans le tableau, les femmes au $131{,}37 \pm 14{,}47$ mmHg$(n = 86)$, et les hommes $130{,}85 \pm 13{,}25$ mmHg$(n = 74)$.
6. Refaites la boîte à moustaches pour imc0, sans séparation par sexe. Repérez sur le graphique les trois valeurs extrêmes supérieures à 39,2 kg/m² (40,0 ; 43,4 et 48,3).

L2.8 Interprétation de la distribution

Rédigez, pour le rapport à rendre, un paragraphe d'interprétation pour la PAS, l'IMC et les triglycérides. Le modèle pour la PAS : "La pression systolique à l'inclusion avait une distribution approximativement symétrique (asymétrie 0,01, aplatissement -0,10), avec une moyenne de 131,13 ± 13,88 mmHg, pratiquement égale à la médiane (131 mmHg ; Q1-Q3 122-140), et des valeurs comprises entre 102 et 169 mmHg. Pour cette variable, la moyenne et l'écart-type sont les indicateurs appropriés." Pour l'IMC et les triglycérides, utilisez l'asymétrie, la relation entre la moyenne et la médiane et la boîte à moustaches pour justifier une présentation par la médiane (Q1-Q3).

Exercices supplémentaires

1. Sur le fichier BD_Studenti disponible sur Moodle : réalisez les tableaux de fréquences et les graphiques appropriés pour le genre, la couleur des yeux, les heures de sommeil et le type de logement ; calculez les indicateurs de tendance centrale et de dispersion pour l'âge,

l'IMC (calculé par formule à partir du poids et de la taille), les heures de sommeil et les heures d'étude ; comparez, au moyen d'un graphique approprié, la couleur des yeux chez les filles et chez les garçons, ainsi que les heures de sommeil chez les filles et chez les garçons. Justifiez le choix de chaque indicateur et de chaque graphique selon le type de la variable. Aucun résultat numérique n'est fourni ; ils sont discutés lors de la séance suivante.

2. Tableau avec valeurs manquantes : l'HbA1c chez des patients diabétiques. Un médecin a regroupé les valeurs d'HbA1c des 40 patients atteints de diabète de type 2 inscrits sur sa liste. Complétez les valeurs A, B, C, D, E et F et répondez : quelle proportion des patients a atteint la cible HbA1c < 7% ? Quelle proportion a une HbA1c ≥ 8% ?

Tableau L2.4. HbA1c chez 40 patients atteints de diabète de type 2.

HbA1c (%)	**Fréquence absolue**	**Fréquence relative (%)**	**Fréquence cumulée (%)**
< 6,5	5	12,5	12,5
6,5-6,9	9	A	35,0
7,0-7,9	B	30,0	C
8,0-8,9	8	D	85,0
≥ 9,0	E	F	100,0
Total	40	100,0	

Solution. $A = \frac{9}{40} \times 100 = 22{,}5\%$ (vérification : 12,5 + 22,5 = 35,0). B = 0,30 x 40 = 12 patients. C = 35,0 + 30,0 = 65,0%. $D = \frac{8}{40} \times 100 = 20{,}0\%$ (vérification : 65,0 + 20,0 = 85,0). F = 100,0 - 85,0 = 15,0%, et E = 0,15 x 40 = 6 patients. Vérification de la somme : 5 + 9 + 12 + 8 + 6 = 40. La cible HbA1c < 7% a été atteinte par 5 + 9 = 14 patients, soit 35,0% (fréquence cumulée de la classe 6,5-6,9). Une HbA1c ≥ 8% concerne 8 + 6 = 14 patients, soit 20,0 + 15,0 = 35,0%. La classe médiane est 7,0-7,9 : les 20e et 21e valeurs ordonnées se trouvent dans cette classe, car la fréquence cumulée passe de 35,0% (14 patients) à 65,0% (26 patients). Interprétation clinique : seul un tiers des patients est à la cible, et un autre tiers a un mauvais contrôle glycémique, qui impose une intensification du traitement.

3. Tableau avec valeurs manquantes : la satisfaction à l'égard des services d'une consultation externe. Dans une enquête de satisfaction, 120 patients ont répondu sur une échelle allant de 1 (très insatisfait) à 5 (très satisfait). Complétez les valeurs manquantes, déterminez la médiane et le mode et calculez la proportion de patients satisfaits ou très satisfaits.

Tableau L2.5. Satisfaction de 120 patients d'une consultation externe.

Satisfaction	**Fréquence absolue**	**Fréquence relative (%)**	**Fréquence cumulée absolue**	**Fréquence cumulée (%)**
1	6	5,0	6	5,0
2	A	12,5	21	B
3	33	C	54	45,0
4	D	35,0	E	80,0
5	F	20,0	120	100,0
Total	120	100,0		

Solution. A = 0,125 x 120 = 15 patients (vérification : 6 + 15 = 21). $B = \frac{21}{120} \times 100 = 17{,}5\%$ (ou 5,0 + 12,5). $C = \frac{33}{120} \times 100 = 27{,}5\%$ (vérification : 17,5 + 27,5 = 45,0). D = 0,35 x 120 = 42 patients. E = 54 + 42 = 96 (vérification : 96/120 = 80,0%). F = 0,20 x 120 = 24 patients (vérification : 96 + 24 = 120). Médiane : n = 120, la médiane est donc la moyenne des valeurs situées aux rangs 60 et 61 ; la fréquence cumulée est de 54 pour la valeur 3 et de 96 pour la valeur 4, les deux rangs ont donc la valeur 4, et la médiane est 4. Le mode est également 4 (42 patients). Satisfaits ou très satisfaits : 35,0 + 20,0 = 55,0% (66 patients). La présentation correcte pour cette

variable ordinale est "médiane 4 (Q1-Q3 3-4)" ; les quartiles se déterminent de la même manière que la médiane, aux rangs 30-31 (valeur 3) et 90-91 (valeur 4).

4. Construisez dans Google Forms un formulaire de recueil de données pour 5 variables du jeu de données mg2026 (sexe, âge, PAS, glycémie, satisfaction), avec les validations appropriées, et vérifiez que vous ne pouvez saisir aucune des erreurs de mg2026_brut.csv (par exemple, PAS 1540, glycémie "87 mg/dl", sexe "Feminin").

Ce qu'il faut rendre et barème indicatif

À déposer sur Moodle : le lien du formulaire Google Forms (ou une capture d'écran des questions), la feuille Google Sheets téléchargée au format .xlsx (onglets réponses, dictionnaire, frequences, pivot, descriptives), le fichier JASP et un rapport PDF de 1-2 pages contenant les tableaux de fréquences, les statistiques descriptives, au moins quatre graphiques (barres, circulaire, histogramme, boîte à moustaches) et les paragraphes d'interprétation de L2.8. Barème indicatif, sur 10 points : le questionnaire, avec des types de questions correctement choisis, des validations et le consentement, 2 points ; la liaison avec Sheets et le recodage, 1 point ; des tableaux de fréquences corrects (fumat, satisfaction, PAS par classes), 2 points ; les tableaux croisés dynamiques, 1 point ; des statistiques descriptives correctes, avec l'indicateur adapté au type de variable, 2 points ; les graphiques, avec titres, axes et unités, 1 point ; l'interprétation de la distribution, 1 point.

Variante hors ligne

Sans connexion internet, le questionnaire est conçu sur papier (avec les mêmes types de questions et les mêmes règles de validation, écrites sous forme de consignes), et les réponses sont saisies dans une feuille LibreOffice Calc ou Microsoft Excel, avec la validation des données définie à partir du menu Données. Les tableaux de fréquences se construisent avec les mêmes fonctions (COUNTIF, COUNTIFS, FREQUENCY ; dans la version roumaine d'Excel, "CONTARDACĂ", "FRECVENȚĂ"), les tableaux croisés dynamiques à partir du menu Insertion > PivotTable, et les statistiques descriptives et les graphiques (y compris la boîte à moustaches) dans JASP, qui fonctionne hors ligne. Le formulaire Google est construit ultérieurement, dans le cadre du devoir à la maison.

Travaux pratiques 3. Distributions et test de la normalité

Objectif

Cette séance transforme le chapitre 3 en étapes de travail. Nous partons de cinq valeurs physiologiques mesurées à l'inclusion des patients dans l'étude mg2026 (l'âge, la pression systolique, l'IMC, la glycémie à jeun et les triglycérides) et nous nous demandons, pour chacune, si sa distribution peut être considérée comme normale. La réponse ne tient pas à un seul nombre : nous la construisons à partir des graphiques, des indicateurs de forme et des tests de normalité, puis nous l'utilisons pour décider quels indicateurs descriptifs rapporter et quel type de test nous pourrons appliquer lors des séances suivantes. Nous travaillons en parallèle dans JASP, MedCalc et Google Sheets, car à l'examen pratique vous devrez choisir l'un d'eux.

À la fin, l'étudiant doit être capable :

- d'obtenir dans JASP et MedCalc l'asymétrie, l'aplatissement, leurs erreurs standard, le test de Shapiro-Wilk, l'histogramme avec la courbe normale et le graphique Q-Q pour une variable continue ;
- de calculer dans Google Sheets l'asymétrie (SKEW), l'aplatissement (KURT), leurs erreurs standard, les scores z (STANDARDIZE) et les proportions théoriques sous la courbe normale (NORM.DIST, NORM.INV) ;
- de construire à la main, dans Google Sheets, un graphique Q-Q ;
- d'appliquer une transformation logarithmique, d'en vérifier l'effet et de rétrotransformer la moyenne en moyenne géométrique ;
- d'identifier les valeurs extrêmes par le score z et par la règle 1,5 x EIQ ;
- de rédiger une décision argumentée du type "la variable X peut/ne peut pas être traitée comme normale", dans le format exigé pour les devoirs de la discipline.

Outils

Logiciels : JASP 0.98.1 ou plus récent [47] ; MedCalc 24.0 [50] (ou ses calculateurs en ligne) ; Google Sheets avec les paramètres régionaux France ou Roumanie (séparateur décimal virgule, séparateur d'arguments ";") [43] ; en option, Epi Info 7.2.7 [51]. Fichiers : mg2026_curat.csv (160 lignes), sur Moodle, plateforme Emediqual [52] ; pour les exercices supplémentaires, les fichiers classiques de la discipline BD_Studenti et Probleme BPOC (problèmes sur la BPCO). Durée indicative : 100 minutes.

Après l'importation dans Google Sheets, les colonnes utilisées aujourd'hui sont : D = varsta, K = imc0, M = tas0, R = trigliceride, S = glicemie0, V = hba1c0. Les données occupent les lignes 2-161. Les valeurs attendues dans cette séance proviennent du fichier de résultats du jeu de données (calculées avec Python) et coïncident avec celles affichées par JASP et MedCalc ; des différences sur la dernière décimale peuvent provenir des arrondis.

Déroulement

L3.1 Préparation du fichier

1. Ouvrez mg2026_curat.csv dans JASP (File > Open > Computer). Dans la vue des données, vérifiez que varsta, imc0, tas0, glicemie0, trigliceride et hba1c0 sont marquées comme variables continues (icône en forme de règle). Si une variable apparaît comme

nominale (icône à trois cercles), cela signifie généralement qu'une cellule contient du texte ; revenez au chapitre 2 et à mg2026_brut.csv pour voir d'où peut venir le problème.

2. Dans JASP, créez une nouvelle variable contenant le logarithme népérien des triglycérides : cliquez sur le signe "+" à droite de la dernière colonne, nommez-la ln_trig, choisissez "R" comme mode de calcul et écrivez la formule log(trigliceride) (en R, la fonction log est le logarithme népérien). La première valeur doit être 5,1417, car le premier patient a des triglycérides à 171 mg/dL.
3. Importez le même fichier dans Google Sheets (File > Import > Upload ; séparateur virgule). Vérifiez que les décimales de l'IMC ont été lues comme des nombres (alignés à droite) et non comme du texte. Dans la colonne AC, écrivez l'en-tête ln_trig et, en AC2, la formule =LN(R2), recopiée jusqu'à la ligne 161.

L3.2 Forme de la distribution dans JASP

4. Menu Descriptives > Descriptive Statistics. Placez dans la zone Variables : varsta, tas0, imc0, glicemie0, trigliceride et ln_trig.
5. Dans la section Statistics, cochez : Mean, Median, Std. deviation, Minimum, Maximum, Quartiles ; sous Distribution, cochez Skewness, Kurtosis et Shapiro-Wilk test. JASP affichera automatiquement aussi les erreurs standard de l'asymétrie et de l'aplatissement.
6. Dans la section Plots (ou Basic plots, selon la version), cochez Distribution plots (avec l'option d'affichage de la densité), Q-Q plots et Boxplots.
7. Comparez le tableau de JASP au tableau L3.1. Toutes les valeurs doivent coïncider aux arrondis près. L'erreur standard de l'asymétrie est la même pour toutes les variables (0,192), car elle ne dépend que de n ; il en va de même pour l'erreur standard de l'aplatissement (0,381).

Tableau L3.1. Valeurs d'autovérification sur mg2026 (n = 160).

Variable	Moyenne	ÉT	Médiane	Asymétrie	Aplatissement	Shapiro-Wilk W	p
varsta	56,29	8,77	56,0	0,01	-0,09	0,993	0,687
tas0	131,13	13,88	131,0	0,01	-0,10	0,990	0,303
imc0	29,04	3,91	28,25	1,42	3,68	0,904	< 0,001
glicemie0	105,44	15,76	104,5	0,37	-0,23	0,982	0,037
trigliceride	161,18	76,30	151,5	0,95	1,12	0,945	< 0,001
ln_trig	4,97	0,49	5,02	-0,28	-0,28	0,988	0,200

8. Pour chaque variable, calculez les rapports asymétrie / 0,192 et aplatissement / 0,381 et notez-les dans votre cahier. Vous devez obtenir, par exemple, 7,42 et 9,64 pour l'IMC, 4,97 et 2,94 pour les triglycérides, 1,92 et -0,60 pour la glycémie. Signalez les rapports qui dépassent 2 en valeur absolue.
9. Regardez les graphiques Q-Q. Pour tas0 et varsta, les points doivent suivre la diagonale. Pour imc0, les 3-4 derniers points à droite montent nettement au-dessus de la droite (les patients ayant un IMC de 40,0 ; 43,4 et 48,3). Pour les triglycérides, les points forment une courbe concave vers le haut, typique d'une asymétrie positive. Pour ln_trig, la courbe se redresse. Dessinez schématiquement dans votre cahier la forme du Q-Q pour l'IMC et pour les triglycérides.
10. Question de contrôle : comment interprétez-vous la glycémie, pour laquelle le test de Shapiro-Wilk donne p = 0,037, alors que l'asymétrie rapportée à son erreur standard est inférieure à 2 et que la moyenne diffère de la médiane de moins de 1 mg/dL ? La réponse attendue, que vous trouverez argumentée dans la section 3.8.5 du cours, est que l'écart à la normalité est faible, explicable par les patients atteints de diabète non diagnostiqué, et

qu'il n'empêche pas l'utilisation de la moyenne et des tests paramétriques pour n = 160 ; il est toutefois rapporté en toute transparence.

11. En option, vérifiez la normalité par groupe : dans la même fenêtre, placez la variable sex dans la zone Split. Pour la glycémie, vous observerez que l'écart provient surtout du groupe des femmes (asymétrie 0,57, Shapiro-Wilk p = 0,022), alors que chez les hommes la distribution est compatible avec la normalité (p = 0,736). Pour tas0 par bras (Split : brat), les trois tests sont non significatifs (p = 0,782 ; 0,373 ; 0,818). Retenez l'idée : pour comparer des groupes, c'est la normalité dans chaque groupe qui compte.

L3.3 Les mêmes analyses dans MedCalc

12. Dans MedCalc, faites File > Open et choisissez mg2026_curat.csv (lors de l'importation, confirmez que la première ligne contient les noms des variables et que le séparateur est la virgule).
13. Menu Statistics > Summary statistics. Pour Variable, choisissez tas0. Dans la fenêtre d'options, cochez le test de la distribution normale (Shapiro-Wilk, ou à défaut Kolmogorov-Smirnov) et, si elle est disponible, l'option des statistiques de forme. Le résultat affiche la moyenne, l'ÉT, l'erreur standard, l'IC à 95% de la moyenne, la médiane avec son IC, l'asymétrie et l'aplatissement (chacun avec la valeur p d'un test propre) et la conclusion du test de normalité ("accept Normality" ou "reject Normality").
14. Comparez les valeurs au tableau L3.1. Attention à deux détails. Premièrement, la formule "accept Normality" de MedCalc est un raccourci ; dans votre rapport, écrivez "on ne rejette pas l'hypothèse de normalité". Deuxièmement, pour le test de Kolmogorov-Smirnov, vérifiez dans la documentation du logiciel si la correction de Lilliefors est appliquée. Sur l'IMC, le test sans correction donne p ≈ 0,08 et, avec la correction de Lilliefors, p ≈ 0,002 ; le test de Shapiro-Wilk donne p < 0,001. Cette différence montre pourquoi le test classique, sans correction, est trop permissif.
15. Les graphiques : Graphs > Histogram (cochez la superposition de la courbe normale) et Graphs > Normal plot (le graphique Q-Q de MedCalc). Pour les triglycérides, l'histogramme avec la courbe normale superposée montre clairement que le pic réel se situe à gauche du sommet de la courbe et que la queue droite dépasse la courbe.
16. Refaites l'opération pour imc0, glicemie0 et trigliceride. Pour les triglycérides, dans la fenêtre Summary statistics, cochez aussi l'option de transformation logarithmique (si votre version la propose) ou créez dans la feuille de données de MedCalc une nouvelle variable, définie par la formule LN(trigliceride). La moyenne géométrique affichée doit être de 144,0 mg/dL.

L3.4 Indicateurs de forme et graphique Q-Q dans Google Sheets

17. Dans une nouvelle feuille, nommée "Forme", construisez un tableau avec les variables en lignes et les indicateurs en colonnes. Pour la PAS, les formules sont les suivantes :

- n : =COUNT(donnees!M2:M161), résultat 160 ;
- moyenne : =AVERAGE(donnees!M2:M161), résultat 131,13 ;
- médiane : =MEDIAN(donnees!M2:M161), résultat 131 ;
- ÉT : =STDEV(donnees!M2:M161), résultat 13,88 ;
- asymétrie : =SKEW(donnees!M2:M161), résultat 0,01 ;
- aplatissement : =KURT(donnees!M2:M161), résultat -0,10 ;
- Q1 et Q3 : =QUARTILE(donnees!M2:M161; 1) et =QUARTILE(donnees!M2:M161; 3), résultats 122 et 140.

Recopiez la ligne pour les autres variables, en changeant la colonne. Les fonctions SKEW et KURT de Google Sheets donnent les mêmes valeurs corrigées pour l'échantillon que JASP et MedCalc.

18. Erreurs standard de l'asymétrie et de l'aplatissement, avec les formules exactes, pour n placé dans la cellule B2 :

- ES de l'asymétrie : =SQRT(6*B2*(B2-1)/((B2-2)*(B2+1)*(B2+3))), résultat 0,192 ;
- ES de l'aplatissement : =2*C2*SQRT((B2^2-1)/((B2-3)*(B2+5))), où C2 contient l'ES de l'asymétrie, résultat 0,381.

Calculez les rapports et appliquez une mise en forme conditionnelle (Format > Conditional formatting) aux cellules dont la valeur absolue est supérieure à 2, avec un fond rouge. Google Sheets ne dispose pas d'un test de Shapiro-Wilk intégré ; pour la valeur p, utilisez JASP ou MedCalc.

19. Construisez le graphique Q-Q de la PAS. Dans une nouvelle feuille, "QQ_TAS" :

- en A2:A161, copiez les valeurs de la PAS et triez-les par ordre croissant (Data > Sort range) ;
- en B2, écrivez le rang, =ROW()-1, et recopiez vers le bas (de 1 à 160) ;
- en C2, calculez la proportion cumulée, =(B2-0,5)/160 ;
- en D2, calculez le quantile normal théorique, =NORM.S.INV(C2) ; la première valeur est -2,73, la dernière 2,73 ;
- sélectionnez D2:D161 et A2:A161 (D sur l'axe X), Insert > Chart > Scatter chart ; ajoutez une courbe de tendance linéaire.

Les points doivent se trouver très près de la droite ; la corrélation entre les deux colonnes, =CORREL(D2:D161; A2:A161), est de 0,996. Refaites l'opération pour les triglycérides : la corrélation tombe à 0,972, et les points s'incurvent à l'extrémité droite. La statistique W du test de Shapiro-Wilk repose précisément sur cette idée : dans quelle mesure les valeurs ordonnées s'alignent sur les quantiles normaux.

20. L'histogramme : sélectionnez la colonne R (trigliceride), Insert > Chart, Chart type : Histogram chart. Dans Customize > Histogram, essayez des largeurs de classes (Bucket size) de 20 puis de 50 mg/dL et observez comment l'aspect change. C'est la raison pour laquelle nous ne jugeons pas la normalité sur le seul histogramme.

L3.5 Scores z et proportions sous la courbe normale

21. Dans la feuille de données, colonne AD, en-tête z_tas. En AD2 : =STANDARDIZE(M2; AVERAGE($M$2:$M$161); STDEV($M$2:$M$161)), recopiée jusqu'à la ligne 161. Le premier patient a une PAS de 139 mmHg et z = 0,57. Vérifiez que la moyenne de la colonne est de 0 et l'ÉT de 1 (=AVERAGE(AD2:AD161), =STDEV(AD2:AD161)).
22. Comptez les patients dont le score z est supérieur à 2 en valeur absolue : =COUNTIF(AD2:AD161; ">2") + COUNTIF(AD2:AD161; "<-2"). Résultat : 10 patients (6,25%), proche des 4,6% attendus théoriquement pour une loi normale. Aucun patient n'a |z| > 3 ; le score z maximal est de 2,73 et le minimal de -2,10.
23. Score z et percentile d'un patient ayant une PAS de 160 mmHg : =STANDARDIZE(160; 131,13; 13,88) donne 2,08 ; =NORM.S.DIST(2,08; TRUE) donne 0,9812. Le patient se situe au 98e percentile. Vérifiez ce même résultat dans le tableau E.1 de l'annexe E (ligne 2,0, colonne 0,08).
24. Proportion théorique et proportion observée pour une PAS ≥ 140 mmHg :

- théorique : =1-NORM.DIST(140; AVERAGE(M2:M161); STDEV(M2:M161); TRUE), résultat 0,261 (26,1%) ;

- observée : =COUNTIF(M2:M161; ">=140")/COUNT(M2:M161), résultat 0,256 (25,6% ; 41 patients).

25. L'intervalle censé contenir 95% des patients : =NORM.INV(0,025; AVERAGE(M2:M161); STDEV(M2:M161)) et =NORM.INV(0,975;...) donnent 103,9 et 158,3 mmHg. Comptez combien de patients se trouvent effectivement entre ces deux limites avec =COUNTIFS(M2:M161; ">=103,9"; M2:M161; "<=158,3") : le résultat est de 150, soit 93,8%.
26. Refaites les étapes 24-25 pour l'IMC, avec le seuil de 30 kg/m². Le modèle normal prédit 40,3% de patients obèses, alors qu'il y en a 33,8% dans l'échantillon. Notez dans votre cahier pourquoi le modèle normal se trompe ici (l'asymétrie positive tire la moyenne vers le haut et "élargit" la courbe).

L3.6 Transformation logarithmique des triglycérides

27. Dans la feuille "Forme", ajoutez la ligne de ln_trig (colonne AC). L'asymétrie doit baisser de 0,95 à -0,28, et l'aplatissement de 1,12 à -0,28. Dans JASP, le test de Shapiro-Wilk pour ln_trig donne p = 0,200, contre p < 0,001 pour trigliceride.
28. Essayez aussi la transformation racine carrée : colonne AE, =SQRT(R2). L'asymétrie doit être de 0,32. Comparez avec le logarithme et argumentez votre choix (voir le tableau 3.7 du cours). Essayez aussi l'inverse, =1/R2 : l'asymétrie monte à 1,48, ce qui signifie que la transformation est trop forte.
29. La moyenne géométrique, par deux voies : =EXP(AVERAGE(AC2:AC161)) et =GEOMEAN(R2:R161). Les deux doivent donner 144,0 mg/dL. Comparez avec la moyenne arithmétique (161,18) et avec la médiane (151,5).
30. L'intervalle de confiance à 95% de la moyenne géométrique se calcule sur l'échelle logarithmique puis est rétrotransformé :

- borne inférieure : =EXP(AVERAGE(AC2:AC161) - CONFIDENCE.T(0,05; STDEV(AC2:AC161); 160)) ;
- borne supérieure : =EXP(AVERAGE(AC2:AC161) + CONFIDENCE.T(0,05; STDEV(AC2:AC161); 160)).

Sheets affiche 133,44 et 155,42 mg/dL, les valeurs du chapitre 3 du cours. À titre d'expérience, arrondissez les bornes de l'échelle logarithmique à deux décimales avant l'exponentiation et observez de combien chaque borne se déplace ; la conclusion, que vous noterez dans votre cahier : sur l'échelle logarithmique, on n'arrondit qu'à la fin.

31. L'intervalle de référence des triglycérides sur l'échelle logarithmique : =EXP(AVERAGE(AC2:AC161) - 1,96*STDEV(AC2:AC161)) et =EXP(AVERAGE(AC2:AC161) + 1,96*STDEV(AC2:AC161)) donnent 55,3 et 374,9 mg/dL. Comparez avec l'intervalle naïf, calculé sur les valeurs non transformées, 11,6-310,7 mg/dL, et avec les percentiles observés, =PERCENTILE(R2:R161; 0,025) et =PERCENTILE(R2:R161; 0,975), qui donnent 53,8 et 364,0 mg/dL.

L3.7 Valeurs extrêmes

32. Pour l'IMC, calculez les barrières de Tukey : Q1 = 26,2, Q3 = 31,4, EIQ = 5,2 ; barrière inférieure =QUARTILE(K2:K161;1)-1,5*(QUARTILE(K2:K161;3)-QUARTILE(K2:K161;1)) = 18,4 ; barrière supérieure = 39,2. Comptez les valeurs situées au-dessus de la barrière supérieure avec COUNTIF : le résultat est de 3 (40,0 ; 43,4 ; 48,3).
33. Refaites l'opération pour les triglycérides : les barrières sont de -37,4 et 347,6 mg/dL, et 5 valeurs dépassent la barrière supérieure (364, 364, 373, 405, 442). Pour ln_trig, aucune

valeur ne sort des barrières. La conclusion, que vous écrirez dans votre cahier : les "valeurs extrêmes" des triglycérides sont la queue normale d'une distribution log-normale, et non des erreurs.

34. Le score z du patient ayant un IMC de 48,3 est de 4,92. Vérifiez-le avec STANDARDIZE. Décidez, par écrit, ce que vous faites de ce patient (réponse attendue : on vérifie dans la fiche source, on le conserve, car la valeur est plausible, et on utilise la médiane et l'EIQ ou des tests non paramétriques pour l'IMC ; on ne le supprime pas).

L3.8 Rédaction de la conclusion au format a-g

35. Rédigez, pour l'IMC, la résolution au format des devoirs de la discipline. Le modèle attendu :

(a) Hypothèse de recherche : l'IMC des patients à risque cardiométabolique de la population étudiée suit une distribution normale.

(b) Hypothèses statistiques : H0, l'IMC provient d'une population à distribution normale ; H1, la distribution de l'IMC dans la population n'est pas normale.

(c) Méthode : le test de Shapiro-Wilk, recommandé pour les échantillons de taille petite et moyenne, complété par l'asymétrie et l'aplatissement rapportés à leurs erreurs standard et par le graphique Q-Q.

(d) Résolution : JASP, Descriptives > Descriptive Statistics, avec les options des étapes 5-6.

(e) Résultat : asymétrie 1,42 (ES 0,192), aplatissement 3,68 (ES 0,381), W = 0,904, p < 0,001 ; le graphique Q-Q montre des écarts importants dans la queue droite.

(f) Conclusion statistique : p < 0,05, on rejette H0.

(g) Conclusion de recherche : la distribution de l'IMC s'écarte significativement de la normalité, avec une asymétrie positive due à un petit groupe de patients atteints d'obésité sévère. L'IMC sera décrit par la médiane et l'EIQ (28,25 ; 26,2-31,4 kg/m²), et, dans les comparaisons, on utilisera des tests non paramétriques ou l'on rapportera aussi leur résultat.

36. Faites de même pour la PAS (conclusion attendue : on ne rejette pas la normalité, p = 0,303 ; on rapporte la moyenne $\pm$ÉT, $131{,}13 \pm 13{,}88$ mmHg).

Exercices supplémentaires

Les exercices 1-3 se résolvent sur les fichiers classiques de la discipline, disponibles sur Moodle ; les résultats numériques ne sont pas donnés ici, et l'enseignant les vérifie en salle de travaux pratiques.

1. Le fichier BD_Studenti. Pour les variables Varsta, IMC (calculé comme poids / taille², la taille étant exprimée en mètres), Ore_somn et Inaltime, obtenez dans JASP l'asymétrie, l'aplatissement, leurs erreurs standard et le test de Shapiro-Wilk, ainsi que les graphiques Q-Q. Refaites l'analyse séparément par sexe (Split : Gen). Pour chaque variable, écrivez une phrase de décision du type de celles de l'étape 35. Attention aux variables qui comportent peu de valeurs distinctes (heures de sommeil déclarées en heures entières) : quel effet cela a-t-il sur le test de Shapiro-Wilk ?
2. Le fichier Probleme BPOC (problèmes sur la BPCO). Analysez la distribution de la capacité vitale initiale (CVV1), de la fréquence cardiaque initiale (FC1), de la pression artérielle maximale initiale (TMAX1) et du débit expiratoire de pointe initial (DEM1). Pour chacune, indiquez si vous utiliseriez, pour la comparaison entre les sexes, le test t non apparié ou le test de Mann-Whitney, et pourquoi. Pour FC1 et FC2, vérifiez la normalité de la différence FC1 - FC2 (créez une nouvelle variable dans JASP), car c'est elle, et non chaque valeur prise isolément, qui compte pour le test t apparié.

3. Sur l'un ou l'autre des deux fichiers, choisissez la variable la plus asymétrique, appliquez-lui la transformation logarithmique et rapportez l'asymétrie avant et après, la moyenne géométrique et son intervalle de confiance à 95%.
4. Problème avec données dans l'énoncé (avec solution). Dans une population de référence de nouveau-nés à terme, le poids de naissance a une moyenne de 3400 g et un écart-type de 450 g, avec une distribution approximativement normale (données fictives). (a) Quelle proportion de nouveau-nés pèse moins de 2500 g ? (b) Quel poids sépare les 10% de nouveau-nés les plus lourds ? (c) Un nouveau-né pèse 4500 g : calculez le score z et le percentile.

Solution. (a) z = (2500 - 3400) / 450 = -2,00 ; Φ(-2,00) = 1 - 0,9772 = 0,0228. Environ 2,3% des nouveau-nés pèseraient moins de 2500 g. Dans Sheets : =NORM.DIST(2500; 3400; 450; TRUE). (b) Φ(z) = 0,90 pour z = 1,28 ; x = 3400 + 1,28 x 450 = 3976 g (=NORM.INV(0,9; 3400; 450) donne 3977 g). (c) z = (4500 - 3400) / 450 = 2,44 ; Φ(2,44) = 0,9927 ; le nouveau-né se situe au 99,3e percentile. Remarque : en réalité, la distribution du poids de naissance a une queue gauche plus épaisse que celle de la loi normale (prématurés, retard de croissance), de sorte que la proportion réelle de poids inférieurs à 2500 g est généralement plus élevée que ne le prédit le modèle normal ; l'exercice montre une fois de plus que le modèle doit être vérifié sur les données.

5. Problème avec données dans l'énoncé (avec solution). Une variable présente, sur 50 patients, une asymétrie de 0,62 et un aplatissement de 0,40. Les erreurs standard pour n = 50 sont de 0,337 (asymétrie) et 0,662 (aplatissement). La distribution est-elle compatible avec la normalité ?

Solution. Les rapports sont de 0,62 / 0,337 = 1,84 et de 0,40 / 0,662 = 0,60, tous deux inférieurs à 2. Les indicateurs ne signalent pas d'écart significatif, mais une asymétrie de 0,62 n'est pas négligeable en valeur : avec 50 patients, la puissance pour la détecter est limitée. La décision se prend en regardant aussi l'histogramme, le graphique Q-Q et la différence entre la moyenne et la médiane ; si la variable est une concentration sérique ou une durée, il est prudent de rapporter aussi la médiane (EIQ) et de vérifier le résultat par un test non paramétrique.

Ce qu'il faut rendre et barème indicatif

À déposer sur Moodle, sous le nom LP3_Nom_Prenom, un seul fichier (PDF ou document Google exporté) contenant :

- le tableau de forme pour les cinq variables et ln_trig, avec la moyenne, la médiane, l'ÉT, l'asymétrie, l'aplatissement, les rapports aux erreurs standard et le p du test de Shapiro-Wilk (2 points) ;
- les graphiques Q-Q pour la PAS et les triglycérides, l'un issu de JASP ou de MedCalc et l'autre construit à la main dans Google Sheets (2 points) ;
- les calculs de L3.5 : le score z et le percentile pour 160 mmHg, les proportions théorique et observée pour une PAS ≥ 140 et pour un IMC ≥ 30, avec un commentaire de deux phrases (2 points) ;
- les résultats de la transformation logarithmique : l'asymétrie avant et après, la moyenne géométrique et son IC à 95% (1,5 point) ;
- l'analyse des valeurs extrêmes pour l'IMC et les triglycérides, avec une décision argumentée (1 point) ;
- la résolution au format a-g pour l'IMC et la PAS (1,5 point).

Les erreurs qui font le plus souvent baisser la note : "le test a confirmé la normalité" au lieu de "on ne rejette pas l'hypothèse de normalité" ; l'utilisation de l'argument FALSE dans NORM.DIST ; la suppression du patient ayant un IMC de 48,3 ; la présentation de la moyenne ± ÉT pour l'IMC sans aucun commentaire sur l'asymétrie.

Variante hors ligne

En l'absence d'accès à internet, Google Sheets ne peut pas être utilisé, mais JASP et MedCalc fonctionnent en local. Toutes les étapes de L3.2 et L3.3 peuvent être réalisées entièrement hors ligne. Les étapes de L3.4-L3.7 peuvent être refaites dans n'importe quel tableur installé localement (Microsoft Excel ou LibreOffice Calc), avec les mêmes noms de fonctions (SKEW, KURT, STANDARDIZE, NORM.DIST, NORM.INV, NORM.S.INV, GEOMEAN, QUARTILE, PERCENTILE) ; dans les versions roumaines d'Excel, le séparateur d'arguments est également ";". Epi Info 7.2.7 peut calculer, dans le module Classic Analysis, la moyenne, l'ÉT et les quartiles (commande MEANS), mais ne propose pas de tests de normalité. En dernier recours, les scores z et les proportions sous la courbe se calculent à la main, avec le tableau E.1 de l'annexe E, à partir des valeurs du tableau L3.1.

Travaux pratiques 4. Estimation et intervalles de confiance : l'effet de la taille de l'échantillon

Objectif

Cette séance met en pratique le chapitre 4. Nous calculons des intervalles de confiance pour des moyennes et pour des proportions sur le jeu de données mg2026, d'abord avec les logiciels, puis à la main dans Google Sheets, pour constater qu'il n'y a aucune "boîte noire" : chaque nombre affiché par JASP ou MedCalc peut être reconstitué à partir de la moyenne, de l'écart-type et de n. Nous construisons sur les mêmes données les intervalles classiques de 68,3%, 95,4% et 99,7% pour les valeurs et pour la moyenne, nous simulons dans Google Sheets la façon dont l'intervalle change lorsque nous prenons des échantillons de 10, 40 et 160 patients, et nous terminons par le calcul de la taille de l'échantillon dans MedCalc, G*Power et Epi Info. C'est l'une des rares séances où vous "verrez" de vos propres yeux le hasard de l'échantillonnage.

À la fin, l'étudiant doit être capable :

- d'obtenir l'IC à 95% et l'IC à 99% d'une moyenne dans JASP, MedCalc et Google Sheets et de les vérifier par un calcul à la main ;
- de construire et d'interpréter, sur les mêmes données, les intervalles pour les valeurs individuelles et pour la moyenne ;
- de calculer l'IC à 95% d'une proportion par les méthodes de Wald, de Wilson et exacte ;
- d'organiser une simulation simple dans Google Sheets et de décrire comment la largeur et la position de l'intervalle dépendent de la taille de l'échantillon ;
- de calculer la taille de l'échantillon pour une moyenne, une proportion et la comparaison de deux moyennes, avec au moins deux logiciels ;
- de rapporter correctement une estimation : valeur ponctuelle, IC à 95%, unité de mesure, méthode.

Outils

Logiciels : JASP 0.98.1 ou plus récent [47] ; MedCalc 24.0 [50] ; G*Power [77] ; Epi Info 7.2.7, module StatCalc [51] (le CDC a arrêté le développement et l'assistance technique d'Epi Info en septembre 2025 ; le logiciel reste utilisable, et chaque étape a aussi une variante dans MedCalc ou G*Power) ; Google Sheets avec les paramètres régionaux France (ou Roumanie) [43]. Fichiers : mg2026_curat.csv sur Moodle [52] ; pour les exercices supplémentaires, BD_Studenti et Probleme BPOC. Durée indicative : 100 minutes.

Colonnes de Google Sheets utilisées aujourd'hui : M = tas0, P = colesterol, Q = hdl, S = glicemie0, $X = \text{diabet}_{\text{ogtt}}\left(\frac{0}{1}\right)$, D = varsta. Les données occupent les lignes 2-161.

Déroulement

L4.1 Intervalle de confiance de la moyenne dans JASP

1. Ouvrez mg2026_curat.csv dans JASP. Menu T-Tests > One Sample T-Test. Placez tas0 dans la zone Variables. Pour Test value, laissez 0 : ce n'est pas le test qui nous intéresse, mais l'intervalle, et avec une valeur de test égale à 0, la différence affichée est la moyenne elle-même.
2. Cochez "Location parameter" (ou "Mean difference", selon la version) et "Confidence interval", au niveau de 95%. Cochez aussi "Descriptives". Résultat attendu : moyenne

131,13 mmHg, ÉT 13,88, ES 1,097, IC à 95% 128,96-133,30 mmHg. Les valeurs t et p affichées (test par rapport à 0) n'ont pas de sens clinique et sont ignorées.

3. Changez le niveau de confiance à 99% : l'intervalle devient 128,27-133,99 mmHg. Puis à 90% : 129,32-132,95 mmHg. Notez les trois intervalles et leurs largeurs (3,6 ; 4,3 ; 5,7 mmHg).
4. Refaites l'opération pour glicemie0 (IC à 95% attendu : 102,98-107,90 mg/dL) et pour colesterol (203,63-215,15 mg/dL).
5. En option : dans Descriptives > Descriptive Statistics, les versions récentes de JASP proposent directement un intervalle de confiance de la moyenne, et l'option bootstrap permet d'obtenir des intervalles pour d'autres statistiques. Si vous en disposez, vérifiez qu'elles donnent le même IC à 95% pour la PAS.

L4.2 Le même intervalle dans MedCalc

6. Dans MedCalc, Statistics > Summary statistics, variable tas0. Le résultat contient les lignes "Arithmetic mean", "95% CI for the mean", "Standard deviation", "Standard error of the mean" et, plus bas, "Median" avec "95% CI for the median". Comparez : IC de la moyenne 128,96-133,30 mmHg ; l'intervalle de la médiane doit se situer autour de 129-134 mmHg (les logiciels peuvent utiliser des variantes légèrement différentes de la méthode des rangs).
7. Pour les triglycérides, le même menu affiche un IC de la moyenne de 149,26-173,09 mg/dL. Si vous activez l'option de transformation logarithmique (lorsqu'elle est disponible), MedCalc rapporte la moyenne géométrique de 144,0 mg/dL avec un IC à 95% de 133,44-155,42 mg/dL. Notez pourquoi ces deux intervalles ne se contredisent pas (ils estiment des paramètres différents, voir la section 4.9.1).

L4.3 Calcul à la main dans Google Sheets

8. Dans une nouvelle feuille, "IC_moyenne", construisez le tableau pas à pas pour la PAS :

- B2, n : =COUNT(donnees!M2:M161), résultat 160 ;
- B3, moyenne : =AVERAGE(donnees!M2:M161), résultat 131,13 ;
- B4, ÉT : =STDEV(donnees!M2:M161), résultat 13,88 ;
- B5, ES : =B4/SQRT(B2), résultat 1,097 ;
- B6, t critique : =T.INV.2T(0,05; B2-1), résultat 1,975 ;
- B7, marge : =B6*B5, résultat 2,167 ;
- B8, borne inférieure : =B3-B7, résultat 128,96 ;
- B9, borne supérieure : =B3+B7, résultat 133,30.

9. Vérifiez la marge avec la fonction dédiée : =CONFIDENCE.T(0,05; B4; B2) doit donner également 2,167. Calculez aussi =CONFIDENCE.NORM(0,05; B4; B2), qui utilise z = 1,96 et donne 2,151. La différence est faible pour n = 160 ; à l'étape 16, vous la verrez augmenter.
10. Pour l'IC à 99%, remplacez 0,05 par 0,01 en B6 : t = 2,607, marge 2,861, intervalle 128,27-133,99 mmHg.

L4.4 Intervalles pour les valeurs et intervalles pour la moyenne

11. Dans la même feuille, construisez le tableau L4.1 avec des formules (k dans la colonne A : 1, 2, 3) :

- borne inférieure pour les valeurs : =$B$3-A12*$B$4 ; supérieure : =$B$3+A12*$B$4 ;

- borne inférieure pour la moyenne : =$B$3-A12*$B$5 ; supérieure : =$B$3+A12*$B$5.

Tableau L4.1. Valeurs d'autovérification pour la PAS (mg2026).

Confiance	**Valeurs individuelles** $(m \pm k \times s)$	**Moyenne** $(m \pm k \times \text{ES})$
68,3% $(k = 1)$	117,3-145,0 mmHg	130,03-132,23 mmHg
95,4% $(k = 2)$	103,4-158,9 mmHg	128,94-133,33 mmHg
99,7% $(k = 3)$	89,5-172,8 mmHg	127,84-134,42 mmHg

12. Comptez combien de patients se trouvent effectivement dans chaque intervalle pour les valeurs, avec =COUNTIFS(donnees!M2:M161; ">="&C12; donnees!M2:M161; "<="&D12). Résultats attendus : 112 patients (70,0%), 150 (93,8%) et 160 (100%). Comptez ensuite combien de patients se trouvent dans l'intervalle à 95,4% de la moyenne : seulement 25 patients. Écrivez dans votre cahier, avec vos propres mots, pourquoi ce second nombre est si faible et pourquoi ce n'est pas un problème.
13. Refaites le tableau pour l'âge (m = 56,29 ; s = 8,77) : l'intervalle à 95,4% pour les valeurs est de 38,75-73,83 ans, et pour la moyenne de 54,90-57,68 ans. C'est exactement le type de problème posé à l'examen : "estimez avec une confiance de 95,4% les valeurs de l'âge" par opposition à "estimez avec une confiance de 95,4% l'âge moyen dans la population".

L4.5 Intervalle de confiance d'une proportion

14. La proportion de patients atteints d'un diabète confirmé par HGPO. Dans la feuille "IC_prop" :

- B2, k : =SUM(donnees!X2:X161), résultat 29 ;
- B3, n : =COUNT(donnees!X2:X161), résultat 160 ;
- B4, p : =B2/B3, résultat 0,181 ;
- B5, z : =NORM.S.INV(0,975), résultat 1,96 ;
- B6, ES de Wald : =SQRT(B4*(1-B4)/B3), résultat 0,0305 ;
- B7 et B8, bornes de Wald : =B4-B5*B6 et =B4+B5*B6, résultats 0,122 et 0,241.

15. L'intervalle de Wilson, en trois cellules :

- B10, le centre : =(B4+B5^2/(2*B3))/(1+B5^2/B3), résultat 0,189 ;
- B11, la marge : =B5/(1+B5^2/B3)*SQRT(B4*(1-B4)/B3+B5^2/(4*B3^2)), résultat 0,059 ;
- B12 et B13, les bornes : =B10-B11 et =B10+B11, résultats 0,129 et 0,248.

IC à 95% de Wilson : 12,9%-24,8%, contre 12,2%-24,1% pour Wald.

16. L'intervalle exact (Clopper-Pearson) : dans JASP, Frequencies > Binomial Test, variable diabet_ogtt (définie comme nominale), cochez "Confidence interval" ; pour le niveau 1, le résultat est de 12,5%-25,0%. Dans Epi Info, Classic Analysis, la commande FREQ diabet_ogtt affiche les mêmes bornes exactes. Dans MedCalc, le calculateur pour une proportion (Tests > One proportion ou les calculateurs en ligne) donne lui aussi l'intervalle exact.
17. Test de sensibilité : en B2 et B3, saisissez manuellement 2 et 20 (deux complications chez 20 patients). L'intervalle de Wald devient -3,1% à 23,1%, avec une borne négative impossible ; l'intervalle de Wilson est de 2,8%-30,1%. Saisissez ensuite 0 et 50 : Wald donne 0-0%, Wilson 0-7,1%, et la règle de trois donne 3/50 = 6%. Rétablissez les formules initiales à la fin.

L4.6 Simulation : comment l'intervalle dépend de la taille de l'échantillon

Dans cette partie, nous traitons les 160 patients de mg2026 comme une "population" de moyenne connue, 131,13 mmHg, et nous en tirons des échantillons aléatoires, exactement comme dans la figure 4.2 du cours.

18. Dans une nouvelle feuille, "Simulation", écrivez en A1 l'en-tête "echantillon". En A2, écrivez la formule =INDEX(donnees!$M$2:$M$161; RANDBETWEEN(1; 160)) et recopiez-la jusqu'en A161. Chaque cellule choisit au hasard un patient parmi les 160 (avec remise, de sorte que le même patient peut apparaître plusieurs fois). À chaque modification de la feuille, les valeurs changent ; pour forcer un nouveau tirage, appuyez sur Suppr dans une cellule vide ou utilisez File > Settings > Calculation > Recalculation : "On change and every minute".
19. Calculez les intervalles pour les 10 premières valeurs, les 40 premières et l'ensemble des 160 valeurs de la colonne A :

- n = 10, en D2 : =AVERAGE(A2:A11) ; E2 : =CONFIDENCE.T(0,05; STDEV(A2:A11); 10) ; les bornes =D2-E2 et =D2+E2 ; la largeur =2*E2 ;
- n = 40 : les mêmes formules sur A2:A41, avec 40 ;
- n = 160 : sur A2:A161, avec 160.

Dans la colonne suivante, écrivez =IF(AND(131,13>=F2; 131,13<=G2); "OUI"; "NON"), qui indique si l'intervalle contient la moyenne de la "population".

20. Copiez les résultats (Edit > Paste special > Values only) dans un tableau de répétitions, puis provoquez un nouveau tirage et recommencez. Faites au moins 20 répétitions (40 en binôme). Pour chaque n, calculez à la fin la moyenne des largeurs et le nombre d'intervalles qui contiennent 131,13.
21. Ce que vous devez observer. Sur 10 000 répétitions effectuées par ordinateur, la largeur moyenne de l'IC à 95% était de 19,3 mmHg pour n = 10 (avec de fortes variations, d'environ 12 à 27 mmHg selon les échantillons), de 8,8 mmHg pour n = 40 (7,2-10,4) et de 4,3 mmHg pour n = 160 (3,9-4,7). La proportion d'intervalles contenant la moyenne de la population était respectivement de 95,3%, 94,9% et 94,9%, c'est-à-dire très proche de 95% pour les trois tailles. Avec 20 répétitions, vous aurez probablement 18-20 "OUI" pour chaque n ; si un "NON" apparaît, ce n'est pas une erreur, ce sont précisément les 5% annoncés. Comparez aussi avec les échantillons fixes du cours (tableau 4.3) : n = 10, IC 116,38-143,62 mmHg (largeur 27,2) ; n = 40, 123,08-133,67 (10,6) ; n = 160, 128,96-133,30 (4,3).
22. La conclusion, à écrire dans votre cahier en 3-4 phrases : la largeur de l'intervalle diminue approximativement proportionnellement à $\sqrt{n}$ (de 10 à 40 patients, elle est divisée par deux, de même de 40 à 160) ; le niveau de confiance de 95% ne dépend pas de n, il se rapporte à la proportion d'intervalles qui contiennent le paramètre ; dans les petits échantillons, ce n'est pas seulement la largeur, mais aussi la position de l'intervalle qui varie beaucoup d'une répétition à l'autre.
23. Variante avec population normale, en option. Remplacez la formule de A2 par =NORM.INV(RAND(); 131; 14), qui génère des valeurs issues d'une population normale de moyenne 131 et d'ÉT 14 mmHg, et refaites les étapes 19-20, en vérifiant la couverture par rapport à 131. C'est exactement la simulation de la figure 4.2.

L4.7 Calcul de la taille de l'échantillon

24. Pour une moyenne, à la main dans Sheets : =ROUNDUP((1,96*14/3)^2; 0) donne 84 patients (marge ±3 mmHg, σ = 14 mmHg). Essayez les marges de 2 et de 5 mmHg : 189 et 31. Certains logiciels utilisent la loi t au lieu de z et donnent un peu plus (pour ±3 mmHg, 87) ; la différence n'a pas d'importance pratique.
25. Pour une proportion : =ROUNDUP(1,96^2*0,2*0,8/0,05^2; 0) donne 246 (prévalence attendue de 20%, marge de ±5 points de pourcentage). Avec p = 0,5, 385.
26. Epi Info, StatCalc > Population survey. Saisissez : Population size 1000, Expected frequency 20%, Acceptable margin of error 5%, Design effect 1, Clusters 1. Au niveau de confiance de 95%, le logiciel affiche 198 sujets : c'est l'échantillon de 246 corrigé pour une population finie de 1000 (246 / (1 + 245 / 1000) = 197,6). Remplacez Population size par 999999 : le résultat devient 246. Faites le même calcul avec Population size 500 (résultat attendu : 166).
27. G*Power, comparaison de deux moyennes : Test family "t tests", Statistical test "Means: Difference between two independent means (two groups)", Type of power analysis "A priori: Compute required sample size". Saisissez Tail(s) : Two, Effect size d : 0,5, α err prob : 0,05, Power (1-β err prob) : 0,80, Allocation ratio N2/N1 : 1. Résultat : 64 par groupe, 128 au total. Refaites le calcul avec d = 0,3 (176 par groupe) et avec d = 0,8 (26 par groupe). Le bouton "Determine" vous permet de calculer d à partir des moyennes et de l'écart-type : pour une différence de 5 mmHg et un ÉT de 14 mmHg, d = 0,357, et l'échantillon nécessaire est de 125 par groupe.
28. MedCalc, menu Sampling : choisissez le calcul pour la comparaison de deux moyennes, saisissez α = 0,05, β = 0,20 (puissance de 80%), la différence de 5 et l'ÉT de 14, et comparez avec G*Power. Utilisez ensuite le calcul pour une seule proportion (prévalence 20%, précision ±5%).
29. Ajustement pour les perdus de vue : si vous prévoyez 20% de perdus de vue, =ROUNDUP(64/(1-0,2); 0) donne 80 par groupe. Vérifiez pourquoi la formule =64*1,2 (résultat 77) est fausse : calculez combien de patients sur 77 restent après 20% de pertes.
30. Puissance d'un échantillon existant : dans G*Power, Type of power analysis "Post hoc: Compute achieved power", $d = 0{,}357\left(\frac{5\ \text{mmHg}}{14\ \text{mmHg}}\right)$, α = 0,05, Sample size group 1 : 86, group 2 : 74. Le résultat est d'environ 61% : la comparaison de la PAS entre les sexes dans mg2026 avait une puissance modeste pour une différence de 5 mmHg. Retenez : ce calcul se fait pour une différence cliniquement pertinente fixée à l'avance, et non pour la différence observée.

L4.8 Présentation d'une estimation

31. Rédigez, pour la PAS, la glycémie et la prévalence du diabète, une phrase de présentation de type article pour chacune. Modèle : "La PAS moyenne à l'inclusion était de 131,13 mmHg (ÉT 13,88 ; IC à 95% de la moyenne 128,96-133,30 mmHg ; n = 160). La prévalence du diabète confirmé par HGPO était de 18,1% (29/160 ; IC à 95% de Wilson 12,9-24,8%)." Vérifiez que chaque phrase contient : l'estimation, l'unité de mesure, l'intervalle avec son niveau, la méthode (pour les proportions) et n.

Exercices supplémentaires

Les exercices 1-3 se résolvent sur les fichiers classiques de la discipline ; les résultats numériques ne sont pas donnés ici.

1. Le fichier BD_Studenti. (a) Estimez, avec une confiance de 68,3%, les valeurs de la taille des étudiants. (b) Estimez, avec une confiance de 95,4%, la moyenne de la taille idéale dans la population des étudiants. (c) Estimez, avec une confiance de 99,7%, les valeurs de l'IMC des étudiants et, séparément, l'IMC moyen dans la population. Commentez la différence entre ces deux derniers intervalles. (d) Calculez l'IC à 95% avec la loi t pour la moyenne des heures de sommeil, séparément pour les filles et pour les garçons, et dites si les intervalles se chevauchent.
2. Le fichier Probleme BPOC. Calculez l'IC à 95% de la moyenne de la fréquence cardiaque initiale (FC1) et de la moyenne de la différence FC1 - FC2. Calculez l'IC à 95% de Wilson pour la proportion de patients présentant une dyspnée à l'inclusion (DISPN = 1), sur l'ensemble de l'effectif et séparément chez les sujets de moins et de plus de 50 ans.
3. Sur BD_Studenti, tirez dans Google Sheets des échantillons aléatoires de 10 et de 30 étudiants (méthode de l'étape 18) et comparez la largeur de l'IC à 95% de la taille moyenne avec celle obtenue sur l'ensemble du fichier.
4. Problème avec données dans l'énoncé (avec solution). Une étude pilote menée chez 15 patients hypertendus observe une baisse moyenne de la PAS de 9 mmHg, avec un ÉT de 12 mmHg. (a) Calculez l'IC à 95% de la baisse moyenne. (b) Combien de patients seraient nécessaires pour estimer la baisse moyenne avec une marge de ±2 mmHg ?

Solution. (a) ES $= \frac{12}{\sqrt{15}} = \frac{12}{3{,}873} = 3{,}098$ mmHg ; ddl = 14, t = 2,145 ; marge = 2,145 x 3,098 = 6,65 mmHg ; IC à 95% : 9 ± 6,65, soit 2,35-15,65 mmHg. L'intervalle exclut 0, mais il est très large : la baisse réelle pourrait être mineure ou importante. $(b)\ n = \left(\frac{1{,}96 \times 12}{2}\right)^2 = (11{,}76)^2 = 138{,}3$, donc 139 patients (avec la loi t, un logiciel donnera 1 ou 2 patients de plus). Notez que l'ÉT de l'étude pilote est lui-même imprécis sur 15 patients ; un calcul prudent utiliserait une valeur un peu plus élevée.

5. Problème avec données dans l'énoncé (avec solution). Dans une enquête menée au cabinet, 45 des 150 patients vaccinés contre la grippe déclarent une réaction locale. Calculez les IC à 95% de Wald et de Wilson.

Solution. $p = \frac{45}{150} = 0{,}30$; ES $= \sqrt{\frac{0{,}30 \times 0{,}70}{150}} = 0{,}0374$; Wald : 0,30 ± 0,073, soit 22,7%-37,3%. Wilson : $\frac{z^2}{n} = 0{,}0256$; centre = (0,30 + 0,0128) / 1,0256 = 0,305 ; marge = $\left(\frac{1{,}96}{1{,}0256}\right) \times \sqrt{0{,}0014 + 0{,}0000427} = 1{,}911 \times 0{,}0380 = 0{,}0726$; IC 23,2%-37,8%. Pour p = 30% et n = 150, les deux méthodes donnent des résultats très proches ; la différence ne devient importante que pour des proportions extrêmes ou un n faible.

Ce qu'il faut rendre et barème indicatif

À déposer sur Moodle, sous le nom LP4_Nom_Prenom, un fichier contenant :

- les intervalles de confiance à 90%, 95% et 99% de la moyenne de la PAS, avec la capture d'écran de JASP ou de MedCalc et le tableau de calcul de Sheets (2 points) ;
- le tableau L4.1 complété pour la PAS et pour l'âge, avec le nombre de patients dans chaque intervalle et le commentaire de l'étape 12 (2 points) ;
- l'IC à 95% de la proportion de diabétiques par les méthodes de Wald, de Wilson et exacte, ainsi que le test de sensibilité de l'étape 17 (1,5 point) ;
- le tableau de la simulation (au moins 20 répétitions pour n = 10, 40, 160), avec la moyenne des largeurs, le nombre d'intervalles contenant 131,13 et la conclusion de l'étape 22 (2,5 points) ;

- les calculs de taille d'échantillon de L4.7, avec les captures d'écran de G*Power ou de MedCalc et d'Epi Info (1,5 point) ;
- les phrases de présentation de l'étape 31 (0,5 point).

Erreurs fréquentes : utiliser l'écart-type au lieu de l'erreur standard pour l'intervalle de la moyenne (ou l'inverse) ; interpréter l'IC à 95% de la moyenne comme l'intervalle dans lequel se trouvent 95% des patients ; arrondir la taille de l'échantillon à l'entier inférieur ; ajuster pour les perdus de vue en multipliant par 1,2.

Variante hors ligne

JASP, MedCalc, G*Power et Epi Info fonctionnent en local, de sorte que les étapes L4.1, L4.2, L4.5 (variantes JASP et Epi Info) et L4.7 se font sans internet. Pour L4.3, L4.4 et la simulation, on utilise un tableur installé localement (Excel ou LibreOffice Calc), avec les mêmes fonctions : AVERAGE, STDEV, T.INV.2T, CONFIDENCE.T, CONFIDENCE.NORM, NORM.S.INV, NORM.INV, RAND, RANDBETWEEN, INDEX, COUNTIFS. Dans Excel, le recalcul des tirages aléatoires se force avec la touche F9. Si aucun ordinateur n'est disponible, les intervalles se calculent à la main, avec les valeurs critiques des tableaux E.2 et E.3 de l'annexe E, à partir des statistiques du tableau L4.1.

Travaux pratiques 5. Test des hypothèses statistiques : formulation des hypothèses, erreurs et décision

Objectif

Cette séance transforme en gestes pratiques le raisonnement du chapitre 5 : d'une question clinique aux hypothèses H0 et H1, des données à la statistique de test et à la valeur p, de la valeur p à une décision correctement formulée. Nous travaillons sur le test le plus simple, le test t pour une seule moyenne, car chaque étape peut y être vérifiée à la main. La seconde partie de la séance est une simulation : vous générerez dans Google Sheets des centaines d'"études" dans lesquelles l'hypothèse nulle est vraie et vous verrez de vos propres yeux qu'environ 5% d'entre elles sont "significatives". La simulation vous aidera à comprendre, mieux que n'importe quelle définition, ce que signifie α et pourquoi un résultat non significatif ne prouve pas l'absence d'effet.

À la fin, l'étudiant doit être capable :

- de formuler les hypothèses statistiques H0 et H1 (en mots et mathématiquement) pour une question clinique donnée et de justifier le choix d'un test bilatéral ou unilatéral ;
- d'appliquer le test t pour une seule moyenne dans JASP, MedCalc et Google Sheets et de lire dans les résultats la statistique t, les degrés de liberté, la valeur p, l'intervalle de confiance et la taille de l'effet ;
- de calculer à la main la statistique t et d'obtenir la valeur p à partir de la table et avec les fonctions T.DIST.2T et T.INV.2T ;
- de simuler l'erreur de type I et la puissance du test dans Google Sheets et d'interpréter la proportion de résultats significatifs ;
- de calculer, avec une formule simple et avec un logiciel, le nombre de patients nécessaire pour une puissance donnée ;
- de rédiger la résolution d'un problème au format a-g des devoirs de la discipline.

Outils

Logiciels : JASP 0.98.1 [47] ; MedCalc 24.0 [50] ; Google Sheets, avec les paramètres régionaux France ou Roumanie (le séparateur d'arguments est ";" et le séparateur décimal est la virgule) [43] ; en option, G*Power pour le calcul de la puissance [77]. Fichiers : mg2026_curat.csv (160 patients) sur la plateforme Moodle de la discipline [52] ; pour les exercices supplémentaires, les fichiers classiques Date_HbA1c, BD_Studenti et BPOC. Durée indicative : 100 minutes (20 minutes pour les hypothèses, 35 minutes pour le test t dans les trois logiciels, 25 minutes pour la simulation, 20 minutes pour la puissance et le devoir).

Les colonnes de mg2026_curat.csv, importé dans Google Sheets avec l'en-tête en ligne 1 et les données en lignes 2-161, sont : A id, B brat, C sex, D varsta, E mediu, F educatie, G fumat, H activitate, I inaltime, J greutate0, K imc0, L talie, M tas0, N tad0, O fc0, P colesterol, Q hdl, R trigliceride, S glicemie0, T glicemie3, U glicemie6, V hba1c0, W findrisc, X diabet_ogtt, Y greutate6, Z tas6, AA scadere5, AB satisfactie. Vérifiez l'ordre après l'importation ; les formules ci-dessous supposent cette disposition.

Déroulement

L5.1 Formulation des hypothèses

Avant d'ouvrir un quelconque logiciel, écrivez les hypothèses sur papier. C'est la règle qui évite la plupart des erreurs à l'examen pratique : le test se choisit en fonction de la question, et non en fonction de ce qui "sort" dans le logiciel.

1. Lisez la question et identifiez la population, la variable et la comparaison. Exemple : "La PAS moyenne des patients à risque cardiométabolique diffère-t-elle de 125 mmHg ?" La population : les adultes à risque cardiométabolique suivis dans les cabinets de médecine générale. La variable : la PAS à l'inclusion (numérique, mmHg). La comparaison : la moyenne de la population par rapport à une valeur fixe, 125 mmHg.
2. Écrivez l'hypothèse de recherche en langage clinique : "la PAS moyenne des patients à risque cardiométabolique diffère de 125 mmHg".
3. Écrivez H0 en mots et mathématiquement : "la PAS moyenne dans la population est de 125 mmHg", $\mu = 125$. Écrivez H1 : "la PAS moyenne dans la population diffère de 125 mmHg", $\mu \neq 125$.
4. Décidez si le test est bilatéral ou unilatéral. Par défaut, il est bilatéral ; un test unilatéral exige une justification écrite avant l'analyse (par exemple, une étude de non-infériorité).
5. Fixez $\alpha = 0,05$.

Exercice en classe (10 minutes, en binôme). Écrivez les hypothèses H0 et H1, en mots et mathématiquement, pour les questions ci-dessous ; précisez le paramètre (μ, μD, π, ρ) et le type de test (bilatéral ou unilatéral).

- (a) La moyenne de la glycémie à jeun dans mg2026 diffère-t-elle de 100 mg/dL ?
- (b) La moyenne de l'HbA1c diffère-t-elle de 5,7% ?
- (c) La glycémie à jeun change-t-elle entre l'inclusion et 6 mois chez les mêmes patients ?
- (d) La PAS diffère-t-elle entre les femmes et les hommes ?
- (e) La baisse de la glycémie diffère-t-elle entre les trois bras de l'étude ?
- (f) La proportion de patients atteints d'un diabète confirmé par HGPO diffère-t-elle entre obèses et non-obèses ?
- (g) L'IMC et la PAS sont-ils corrélés ?
- (h) Un nouveau glucomètre ne sous-estime pas la glycémie de plus de 5 mg/dL par rapport au laboratoire.

Vérification. Pour (a) : H0 : $\mu = 100$; H1 : $\mu \neq 100$. Pour (c) : H0 : $\mu D = 0$, où D = glicemie0 - glicemie6 ; H1 : $\mu D \neq 0$. Pour (e) : H0 : $\mu A = \mu B = \mu C$; H1 : au moins une moyenne diffère. Pour (g) : H0 : $\rho = 0$; H1 : $\rho \neq 0$. Pour (h) : hypothèses de non-infériorité, unilatérales, avec une marge de 5 mg/dL fixée à l'avance. Les hypothèses s'écrivent toujours avec les paramètres de la population, jamais avec la moyenne de l'échantillon.

L5.2 Test t pour une seule moyenne dans JASP

1. Ouvrez JASP et chargez le fichier : menu principal (les trois traits en haut à gauche) > Open > Computer > Browse > mg2026_curat.csv. Vérifiez dans la vue des données que tas0 et glicemie0 apparaissent comme variables continues (icône en forme de règle) ; sinon, cliquez sur l'icône et choisissez Scale.
2. Dans la barre des modules, choisissez T-Tests > One Sample T-Test (dans la section Classical).
3. Placez tas0 dans la zone Variables. Dans la zone Test value, écrivez 125.
4. Sous Tests, cochez Student et Wilcoxon signed-rank. Pour Alt. Hypothesis, laissez "≠ Test value" (bilatéral).
5. Sous Additional Statistics, cochez : Location parameter (différence moyenne par rapport à la valeur testée), Confidence interval (95%), Effect size (d de Cohen), Descriptives et, si vous voulez un graphique, Descriptives plots.
6. Sous Assumption Checks, cochez Normality (Shapiro-Wilk) et, en option, Q-Q plot.

Ce que vous devez obtenir pour tas0 par rapport à 125 : moyenne 131,13 mmHg, ÉT 13,88, n = 160 ; test de Student t = 5,59, ddl = 159, p < 0,001 ; différence moyenne (location parameter) 6,13 mmHg, avec un IC à 95% de 3,96-8,30 mmHg ; d de Cohen = 0,44 ; Shapiro-Wilk p = 0,303 ; test de Wilcoxon p < 0,001. Remarquez que JASP rapporte l'intervalle de confiance de la différence par rapport à 125, et non celui de la moyenne ; en ajoutant 125, vous obtenez l'intervalle de la moyenne, 128,96-133,30 mmHg, connu depuis le chapitre 4.

7. Remplacez Test value par 130 et lisez de nouveau le résultat. Vous devez obtenir t = 1,03, p = 0,304, un IC à 95% de la différence de -1,04 à 3,30 mmHg (qui contient zéro). Écrivez la décision correcte : "on ne rejette pas H0", et non "on accepte H0".
8. Refaites l'analyse pour glicemie0, avec Test value 100. Vous devez obtenir : moyenne 105,44 mg/dL, ÉT 15,76 ; t = 4,36, ddl = 159, p < 0,001 ; différence moyenne 5,44 mg/dL (IC à 95% 2,98-7,90) ; d = 0,35 ; Shapiro-Wilk p = 0,037 (léger écart à la normalité) ; Wilcoxon p < 0,001. Discutez avec votre voisin : le test de Wilcoxon est-il nécessaire ici ? Avec n = 160 et une faible asymétrie (0,37), le test t est robuste, et les deux tests conduisent à la même conclusion.
9. Refaites l'analyse pour hba1c0 avec Test value 5,7. Résultat attendu : t = -3,48, ddl = 159, p < 0,001 ; la moyenne est inférieure à 5,7%.

L5.3 Le même test dans MedCalc

1. Ouvrez MedCalc et importez le fichier : File > Import > CSV (ou File > Open, avec le type de fichier CSV). Vérifiez le séparateur décimal dans les options d'importation, car le fichier utilise le point décimal.
2. Dans le menu Statistics, choisissez T-tests > One sample t-test.
3. Pour Variable, sélectionnez tas0, et pour Test mean (la valeur de comparaison), écrivez 125. Confirmez par OK.
4. Lisez dans la fenêtre des résultats : la moyenne, l'IC à 95% de la moyenne, la différence par rapport à la valeur testée, la statistique t, les degrés de liberté et la valeur p. Les résultats doivent coïncider avec ceux de JASP (t = 5,59 ; p < 0,0001).
5. MedCalc permet aussi le calcul direct à partir de données résumées, sans fichier : dans le menu Tests, cherchez le test pour une moyenne (Test for one mean), saisissez la moyenne 131,13, l'ÉT 13,88, n = 160 et la valeur 125. Le résultat doit être le même. C'est une méthode utile pour vérifier les chiffres d'un article.

L5.4 Le même test dans Google Sheets, pas à pas

Google Sheets n'a pas de fonction dédiée au test t pour une moyenne, mais vous pouvez le construire à partir de fonctions simples ; c'est précisément pour cela que c'est le meilleur exercice

de compréhension. Travaillez dans une nouvelle feuille (par exemple, "LP5") et écrivez les libellés dans la colonne A et les formules dans la colonne B.

1. B1, moyenne : =AVERAGE(donnees!M2:M161) (si les données se trouvent dans la feuille "donnees"). Résultat : 131,13.
2. B2, écart-type : =STDEV(donnees!M2:M161). Résultat : 13,88.
3. B3, n : =COUNT(donnees!M2:M161). Résultat : 160.
4. B4, valeur testée μ0 : 125.
5. B5, erreur standard : =B2/SQRT(B3). Résultat : 1,097.
6. B6, statistique t : =(B1-B4)/B5. Résultat : 5,59.
7. B7, degrés de liberté : =B3-1. Résultat : 159.
8. B8, valeur p bilatérale : =T.DIST.2T(ABS(B6);B7). Résultat : environ 0,0000001, soit p < 0,001. La fonction exige une valeur positive, c'est pourquoi nous utilisons ABS.
9. B9, valeur critique pour α = 0,05 en bilatéral : =T.INV.2T(0,05;B7). Résultat : 1,975.
10. B10 et B11, bornes de l'IC à 95% de la moyenne : =B1-B9*B5 et =B1+B9*B5. Résultat : 128,96 et 133,30.
11. B12, décision automatique : =IF(B8<0,05;"on rejette H0";"on ne rejette pas H0").
12. B13, taille de l'effet d : =(B1-B4)/B2. Résultat : 0,44.

Une vérification élégante avec la fonction T.TEST. Le test t pour une moyenne équivaut à un test t apparié entre les données et une colonne constante égale à μ0. Écrivez 125 dans les cellules donnees!AD2:AD161 (écrivez 125 en AD2 et faites glisser vers le bas) et calculez =T.TEST(donnees!M2:M161;donnees!AD2:AD161;2;1). Le résultat doit être identique à celui de B8. Le troisième argument (2) demande le test bilatéral, et le quatrième (1) le test apparié.

Changez maintenant la valeur de B4 en 130, puis en 140, et observez comment se modifient simultanément t, p, la décision et la position de μ0 par rapport à l'intervalle de confiance. Complétez le tableau ci-dessous ; il reproduit, dans votre feuille, l'équivalence des trois voies de décision du chapitre 5.

μ0	t	p	μ0 dans l'IC à 95% ?	Décision
125	5,59	< 0,001	non	on rejette H0
130				
132				
140				

L5.5 Valeur p calculée à la main à partir de la statistique t

À l'examen écrit, vous n'avez pas de calculatrice statistique, mais vous devez pouvoir situer la valeur p entre des seuils à l'aide de la table des valeurs critiques (tableau 5.6 du chapitre).

1. Calculez t et les degrés de liberté.
2. Cherchez dans la table la ligne correspondant aux degrés de liberté (ou la ligne inférieure la plus proche, par prudence).
3. Comparez |t| aux valeurs critiques pour 0,05, 0,01 et 0,001. Si |t| est inférieur à la valeur critique pour 0,05, alors p > 0,05 ; s'il se situe entre les valeurs pour 0,05 et 0,01, alors 0,01 < p < 0,05 ; et ainsi de suite.
4. Vérifiez ensuite dans Google Sheets avec =T.DIST.2T(t;ddl).

Exercices d'encadrement (résolvez-les d'abord avec la table, puis vérifiez dans Sheets) : t = 2,10 avec ddl = 20 (le résultat exact doit être 0,049) ; t = 0,83 avec ddl = 9 (0,43) ; t = 3,10 avec ddl = 9 (0,013) ; t = 1,80 avec ddl = 25 (bilatéral 0,084 ; unilatéral =T.DIST.RT(1,8;25) = 0,042) ; t =

4,36 avec ddl = 159 (< 0,001). Pour l'avant-dernier exemple, discutez : pourquoi est-il dangereux de décider, après avoir vu les données, que le test "devrait" être unilatéral ?

Exercice 5 de la séance : calculez à la main t pour un petit échantillon. Prenez les 10 premiers patients du fichier (lignes 2-11) et testez si la PAS moyenne diffère de 125 mmHg. Calculez la moyenne, l'ÉT, l'ES, t, les ddl et situez p à l'aide de la table. Comparez avec le résultat obtenu sur 160 patients et expliquez, en deux phrases, pourquoi la conclusion peut être différente. (Le résultat dépend des patients choisis ; dans le chapitre, un échantillon aléatoire de 10 patients a donné t = 0,83, p = 0,43.)

L5.6 Simulation de l'erreur de type I dans Google Sheets

Le but est de voir ce que signifie α = 0,05. Nous allons générer 200 "études" dans lesquelles deux groupes de 20 patients chacun proviennent de la même population (PAS de moyenne 130 mmHg et d'ÉT 14 mmHg). Par construction, H0 est vraie dans chaque étude. Tout résultat significatif est donc une erreur de type I.

1. Créez une nouvelle feuille, "Simulation". Dans la cellule A1, écrivez "etude", en B1:U1 les libellés du groupe 1 (g1_1... g1_20), en V1:AO1 les libellés du groupe 2 (g2_1... g2_20), en AP1 "p".
2. En A2, écrivez 1, en A3 =A2+1, et faites glisser jusqu'en A201 (200 études).
3. En B2, écrivez =NORM.INV(RAND();130;14) et recopiez la formule dans tout le bloc B2:AO201. Chaque cellule est maintenant une valeur de PAS aléatoire issue d'une population normale de moyenne 130 et d'ÉT 14.
4. En AP2, écrivez =T.TEST(B2:U2;V2:AO2;2;2) et faites glisser jusqu'en AP201. Chaque ligne reçoit la valeur p d'un test t pour échantillons indépendants entre les deux groupes de la même étude.
5. En AR1, écrivez "proportion p < 0,05", et en AR2 =COUNTIF(AP2:AP201;"<0,05")/200. Appliquez le format pourcentage.
6. Modifiez une cellule vide quelconque (par exemple, écrivez quelque chose en AT1 puis effacez-le) pour forcer le recalcul de la fonction RAND. Notez 5 fois la proportion obtenue.

Ce que vous devez obtenir : des proportions autour de 5%, généralement entre 2% et 8% (avec 200 études, la variation aléatoire est d'environ ±3 points de pourcentage). Dans la simulation du chapitre (1 000 études), la proportion était de 4,9%. Construisez aussi un histogramme des valeurs p de la colonne AP (Insert > Chart > Histogram) : il devrait être à peu près plat, avec des valeurs p réparties uniformément entre 0 et 1. La conclusion à écrire dans votre cahier : lorsque H0 est vraie, le test déclare "significatives" environ une proportion α des études ; α est le taux de fausses alertes que nous acceptons.

7. Simulation de la puissance. Remplacez la formule du groupe 2 (V2:AO201) par =NORM.INV(RAND();137;14), c'est-à-dire une différence réelle de 7 mmHg entre les groupes (d = 0,5). H0 est maintenant fausse, et la proportion de résultats significatifs estime la puissance. Ce que vous devez obtenir : environ 34% (la valeur théorique pour d = 0,5 et 20 patients par groupe est de 33,8%). Autrement dit, deux études de ce type sur trois passeraient à côté d'un effet réel de 7 mmHg. C'est l'erreur de type II.
8. Essayez aussi une grande différence, avec une moyenne du groupe 2 égale à 144 mmHg (d = 1) : la puissance devrait monter à environ 87%.
9. Comparaisons multiples. Ramenez la moyenne du groupe 2 à 130 (H0 vraie). Considérez chaque bloc de 10 lignes comme un seul "article" comportant 10 tests. Combien des 20 "articles" ont au moins un p < 0,05 ? En théorie, 1 - 0,95^10 = 40% d'entre eux. Vérifiez, en comptant ou avec =COUNTIF sur des blocs de 10 lignes.

L5.7 Calcul de la puissance et de la taille de l'échantillon

1. La formule approchée du nombre de patients par groupe pour la comparaison de deux moyennes, dans Google Sheets : =2*(NORM.S.INV(0,975)+NORM.S.INV(0,8))^2*σ^2/Δ^2. Pour σ = 14 mmHg et Δ = 7 mmHg (d = 0,5) : le résultat est de 62,8, soit environ 63 par groupe ; le calcul exact, avec la loi t, donne 64 par groupe (la valeur utilisée au chapitre 5). Pour Δ = 5 mmHg : environ 123 par groupe. Pour Δ = 4,2 mmHg (d = 0,3) : environ 175 par groupe (176 en calcul exact).
2. La formule approchée de la puissance, pour un n donné par groupe : =NORM.S.DIST(d*SQRT(n/2)-NORM.S.INV(0,975);TRUE). Pour d = 0,5 et n = 20 : environ 0,35, très proche de la simulation. Pour d = 0,5 et n = 64 : environ 0,81.
3. Dans JASP, le calcul de la puissance se fait à partir du module Power (s'il n'apparaît pas dans la barre des modules, on l'active avec le bouton "+" situé à droite de la barre) : choisissez le test t pour échantillons indépendants, saisissez la taille de l'effet, α et la puissance souhaitée, et lisez n.
4. Dans MedCalc, le menu Sampling propose le calcul de la taille de l'échantillon pour la comparaison de deux moyennes (saisissez la différence, l'ÉT, α et β). Autre possibilité, G*Power : t tests > Means: Difference between two independent means (two groups) > A priori.
5. Complétez le tableau et commentez la façon dont n change lorsque la puissance augmente ou que α diminue.

Δ (mmHg)	σ	d	α	Puissance	n par groupe (formule)
7	14	0,5	0,05	80%	environ 63
7	14	0,5	0,05	90%	
7	14	0,5	0,01	80%	
5	14	0,36	0,05	80%	
10	14	0,71	0,05	80%	

Valeurs d'autovérification (formule approchée, arrondie à l'entier supérieur) : 85 par groupe pour une puissance de 90% ; 94 par groupe pour α = 0,01 ; 123-124 par groupe pour Δ = 5 mmHg (123,07 avant arrondi) ; 31 par groupe pour Δ = 10 mmHg. Le calcul exact avec la loi t donne des valeurs supérieures de 1 ou 2 patients (par exemple, 86 et 96 pour les deux premières lignes).

L5.8 Le devoir au format a-g : un modèle complet

Problème. La moyenne de la glycémie à jeun à l'inclusion dans mg2026 diffère-t-elle du seuil de 100 mg/dL ?

(a) Hypothèse de recherche : la moyenne de la glycémie à jeun dans la population des adultes à risque cardiométabolique suivis dans les cabinets de médecine générale diffère du seuil de 100 mg/dL de l'hyperglycémie modérée à jeun.

(b) Hypothèses statistiques. H0 : la moyenne de la glycémie à jeun dans la population est de 100 mg/dL ; μ = 100. H1 : la moyenne diffère de 100 mg/dL ; μ ≠ 100. Test bilatéral, α = 0,05.

(c) Méthode : une variable numérique, une moyenne comparée à une valeur fixe, σ inconnu ; test t pour une seule moyenne. Conditions : observations indépendantes ; normalité vérifiée par le test de Shapiro-Wilk (p = 0,037, léger écart, acceptable pour n = 160 grâce au théorème central limite) ; vérification complémentaire par le test de Wilcoxon.

(d) Démarche de résolution : JASP, T-Tests > One Sample T-Test, variable glicemie0, Test value 100, avec Student et Wilcoxon, Location parameter, Effect size, Descriptives, Normality ; vérification dans Google Sheets avec les formules de L5.4.

(e) Résultat : $m = 105{,}44$ mg/dL, s = 15,76, n = 160 ; t(159) = 4,36 ; p < 0,001 ; différence moyenne 5,44 mg/dL (IC à 95% 2,98-7,90) ; d = 0,35 ; Wilcoxon p < 0,001. Le graphique

: l'histogramme de la glycémie avec une ligne verticale à 100 mg/dL ou le graphique descriptif de JASP avec la moyenne et l'IC à 95%.

(f) Conclusion statistique : p < 0,001 < 0,05, on rejette H0 ; la différence est extrêmement significative.

(g) Conclusion de recherche : la moyenne de la glycémie à jeun des patients à risque cardiométabolique est supérieure au seuil de 100 mg/dL, d'environ 3-8 mg/dL ; le patient type de l'étude se situe dans la zone de l'hyperglycémie modérée à jeun. L'effet est faible à moyen (d = 0,35), et le résultat décrit la population étudiée, et non la population générale.

Exercices supplémentaires

Résolvez chaque exercice au format a-g, dans JASP et dans Google Sheets. Pour les fichiers disponibles sur Moodle, les résultats ne sont pas fournis ; vérifiez-vous mutuellement entre collègues et comparez les valeurs obtenues dans les deux logiciels.

1. Le fichier Date_HbA1c. La moyenne de l'HbA1c de la première année des données diffère-t-elle de la cible thérapeutique de 7% ? Formulez les hypothèses, vérifiez la normalité et appliquez le test approprié. Refaites l'analyse pour la dernière année. Commentez la différence entre la signification statistique et l'atteinte de la cible au niveau individuel.
2. Le fichier BD_Studenti. La taille moyenne des étudiantes diffère-t-elle d'une valeur de référence choisie par l'enseignant (par exemple, la moyenne nationale des jeunes femmes, communiquée pendant la séance) ? Et celle des étudiants ? Lequel des deux tests a la plus grande puissance, et pourquoi ?
3. Le fichier BPOC. La fréquence cardiaque initiale moyenne des patients diffère-t-elle de 80 battements/min ? Vérifiez la normalité ; le cas échéant, rapportez aussi le test de Wilcoxon.
4. Avec données dans l'énoncé (avec solution complète). Chez 16 patients hypothyroïdiens traités, la TSH a une moyenne de 3,1 mUI/L et un ÉT de 1,2 mUI/L. La cible du laboratoire est de 2,5 mUI/L. (a) H0 : μ = 2,5 ; H1 : μ ≠ 2,5. (b) ES $= \frac{1,2}{\sqrt{16}} = 0,30$. (c) $t = \frac{3,1-2,5}{0,30} = 2,00$, ddl = 15. (d) t critique (15 ; 0,05) = 2,131 ; 2,00 < 2,131, donc p > 0,05 (valeur exacte, T.DIST.2T(2;15) = 0,064). (e) IC à 95% de la moyenne : $3,1 \pm 2,131 \times 0,30 = 2,46 - \frac{3,74 m_{\text{UI}}}{L}$, un intervalle qui contient 2,5. (f) On ne rejette pas H0. (g) Les données ne permettent pas de conclure que la TSH moyenne diffère de la cible ; l'intervalle est cependant large, le résultat est donc indéterminé, et avec un échantillon plus grand, la même différence moyenne de 0,6 mUI/L pourrait devenir significative.
5. Avec données dans l'énoncé (avec solution complète). Une étude en projet veut détecter une différence de glycémie à jeun de 8 mg/dL entre deux groupes, avec $\sigma = 16$ mg/dL, α = 0,05 en bilatéral et une puissance de 80%. $d = \frac{8}{16} = 0,5$; $n = \frac{2 \times (1,96+0,84)^2 \times 16^2}{8^2} = \frac{2 \times 7,84 \times 256}{64} = 62,7$, soit 63 patients par groupe (64 avec le calcul exact). Si l'on prévoit 15% de perdus de vue, on recrute 64 / 0,85 = 75,3, soit 76 patients par groupe.

Ce qu'il faut rendre et barème indicatif

Il faut rendre, sur Moodle, un seul fichier (PDF ou document) contenant : le tableau des hypothèses pour les 8 questions de L5.1 (1 point) ; le résultat du test t pour tas0 par rapport à 125 et par rapport à 130 dans JASP, avec la capture du tableau et la décision correctement formulée (2 points) ; la feuille Google Sheets de L5.4, avec les formules visibles et le tableau complété pour

les quatre valeurs de μ0 (2 points) ; le résultat de la simulation de L5.6 (les 5 proportions, l'histogramme des valeurs p, la puissance estimée pour la différence de 7 mmHg) avec une interprétation de 3-4 phrases (2 points) ; le tableau de puissance de L5.7 complété (1 point) ; un devoir au format a-g, sur l'un des problèmes supplémentaires 1-3 (2 points). Total : 10 points. Sont pénalisés : des hypothèses écrites avec la moyenne de l'échantillon, l'expression "on accepte H0", un p rapporté sous la forme 0,000, l'absence d'intervalle de confiance et l'absence de conclusion de recherche formulée en mots.

Variante hors ligne

Si le réseau ou la plateforme ne fonctionnent pas, la séance peut être réalisée entièrement avec une calculatrice de poche et la table des valeurs critiques de t du chapitre 5, ainsi qu'avec JASP installé en local (le logiciel n'a pas besoin d'internet après l'installation). Le fichier mg2026_curat.csv se copie à partir de la clé USB de l'enseignant. À la place de Google Sheets, on utilise n'importe quel tableur installé localement (LibreOffice Calc ou Excel), avec les mêmes fonctions : T.DIST.2T, T.INV.2T, NORM.INV, RAND, COUNTIF (dans LibreOffice en roumain, les noms des fonctions peuvent apparaître traduits, et le séparateur d'arguments reste ";"). La simulation de L5.6 peut aussi se faire "à la main", en groupe : chaque étudiant lance 20 fois deux dés, note la somme et compare sa moyenne à celle de son voisin de table par un test t calculé à la main ; à l'échelle du groupe, la proportion de comparaisons "significatives" sera faible, et la discussion sur l'erreur de type I reste la même.

Travaux pratiques 6. Tests paramétriques : le test t et l'ANOVA

Objectif

Dans cette séance, vous appliquez au jeu de données mg2026 les tests paramétriques du chapitre 6 : le test t pour échantillons appariés, le test t pour échantillons indépendants (Student et Welch), l'ANOVA à un facteur avec les tests post hoc et l'ANOVA à mesures répétées. Nous travaillons en parallèle dans JASP, MedCalc et Google Sheets. Le but n'est pas d'appuyer sur des boutons, mais de reconnaître le schéma d'étude, de vérifier les conditions d'application avant de lire la valeur p et de pouvoir refaire dans Google Sheets, avec des formules simples, un résultat obtenu dans un logiciel spécialisé. Chaque étape indique les chiffres que vous devez obtenir, afin que vous puissiez vous vérifier vous-mêmes.

À la fin, l'étudiant doit être capable :

- de choisir entre le test t apparié, le test t pour échantillons indépendants et l'ANOVA en fonction du schéma d'étude et de formuler les hypothèses en mots et mathématiquement ;
- d'exécuter le test t apparié et le test t pour échantillons indépendants dans JASP et MedCalc, avec la vérification de la normalité et de l'égalité des variances, et d'interpréter la différence entre les variantes de Student et de Welch ;
- d'exécuter l'ANOVA à un facteur dans JASP, avec le test de Levene, le graphique Q-Q des résidus, les tests post hoc de Tukey et de Games-Howell et le graphique des moyennes ;
- de construire dans Google Sheets le tableau d'ANOVA avec les fonctions AVERAGEIF, DEVSQ, FILTER et F.DIST.RT et d'utiliser T.TEST avec les types 1, 2 et 3 ;
- d'exécuter et d'interpréter une ANOVA à mesures répétées, y compris le test de Mauchly et la correction de Greenhouse-Geisser ;
- de rédiger les résultats au format a-g, avec la taille de l'effet et l'intervalle de confiance.

Outils

Logiciels : JASP 0.98.1 [47] ; MedCalc 24.0 [50] ; Google Sheets avec les paramètres régionaux France ou Roumanie (séparateur d'arguments ";", virgule décimale) [43]. Fichiers sur Moodle [52] : mg2026_curat.csv ; pour les exercices supplémentaires, BD_lp3 (glycémie à 0, 3 et 6 mois, statut d'assuré), BPOC (fréquence cardiaque, VEMS, indice de perméabilité bronchique avant et après traitement, médicament reçu) et BD_Studenti (taille, taille idéale, sexe, groupe). Durée indicative : 100 minutes (10 minutes pour la préparation des données, 25 minutes pour le test apparié et le test pour échantillons indépendants, 35 minutes pour l'ANOVA, 15 minutes pour les mesures répétées, 15 minutes pour Google Sheets et le devoir).

Les colonnes de mg2026_curat.csv, dans l'ordre du fichier : A id, B brat, C sex, D varsta, E mediu, F educatie, G fumat, H activitate, I inaltime, J greutate0, K imc0, L talie, M tas0, N tad0, O fc0, P colesterol, Q hdl, R trigliceride, S glicemie0, T glicemie3, U glicemie6, V hba1c0, W findrisc, X diabet_ogtt, Y greutate6, Z tas6, AA scadere5, AB satisfactie.

Déroulement

L6.1 Préparation des données : les variables de variation

La question principale de l'étude mg2026 porte sur la baisse de la glycémie à 6 mois selon les bras. Pour l'ANOVA, nous avons besoin d'une nouvelle variable, calculée pour chaque patient.

1. Dans JASP, dans la vue des données, cliquez sur le signe "+" situé après la dernière colonne (Add computed column). Nom : dglic ; type : Scale ; méthode : formule R ou constructeur visuel. Formule : glicemie0 - glicemie6. Confirmez avec Compute column. La moyenne de la nouvelle colonne doit être de 10,19 mg/dL.
2. Créez de la même manière dgr = greutate0 - greutate6 (perte de poids, kg). Moyenne : 2,31 kg.
3. Dans Google Sheets, dans la feuille des données, écrivez en AC1 "dglic", en AC2 =S2-U2 et faites glisser jusqu'en AC161 ; en AD1 "dgr", en AD2 =J2-Y2 et faites glisser. Vérifiez : =AVERAGE(AC2:AC161) donne 10,19 ; =AVERAGE(AD2:AD161) donne 2,31.
4. Notez la convention de signe : les valeurs positives signifient une baisse. Une erreur fréquente à l'examen consiste à interpréter à l'envers le signe de la différence.

L6.2 Test t pour échantillons appariés dans JASP

La question : la glycémie à jeun change-t-elle entre l'inclusion et 6 mois chez les mêmes patients ? H0 : $\mu D = 0$; H1 : $\mu D \neq 0$, où D = glicemie0 - glicemie6.

1. T-Tests > Paired Samples T-Test. Placez la paire glicemie0 et glicemie6 dans la zone Variable pairs (la première variable à gauche, la seconde à droite).
2. Sous Tests, cochez Student et Wilcoxon signed-rank ; hypothèse alternative "Measure 1 ≠ Measure 2".
3. Sous Additional Statistics, cochez Location parameter, Confidence interval, Effect size, Descriptives et Descriptives plots (ou le graphique "raincloud", s'il est disponible dans votre version).
4. Sous Assumption Checks, cochez Normality (Shapiro-Wilk). JASP teste la normalité des différences, c'est-à-dire exactement la bonne condition.

Ce que vous devez obtenir : glycémie moyenne de 105,4 mg/dL à l'inclusion et de 95,2 mg/dL à 6 mois ; t = 11,14, ddl = 159, p < 0,001 ; différence moyenne de 10,19 mg/dL, IC à 95% 8,38-11,99 ; d de Cohen (pour données appariées, dz) = 0,88 ; Shapiro-Wilk sur les différences p = 0,725 ; test de Wilcoxon p < 0,001.

5. Ajoutez les paires glicemie0-glicemie3 et glicemie3-glicemie6 dans la même analyse. Valeurs attendues : 0-3 mois, différence de 6,08 mg/dL (IC à 95% 4,61-7,54), t = 8,22, dz = 0,65, Shapiro-Wilk p = 0,133 ; 3-6 mois, différence de 4,11 mg/dL (IC à 95% 2,10-6,13), t = 4,03, p < 0,001, dz = 0,32, Shapiro-Wilk p = 0,868. Attention : trois tests appariés sur les mêmes données constituent des comparaisons multiples ; l'analyse correcte pour trois temps de mesure est l'ANOVA à mesures répétées (L6.5).
6. La paire tas0-tas6 : différence de 4,78 mmHg (IC à 95% 3,36-6,19), t = 6,68, ddl = 159, p < 0,001, dz = 0,53, Shapiro-Wilk p = 0,129.
7. Exercice de contrôle : exécutez, volontairement à tort, le test t pour échantillons indépendants sur les mêmes données (la glycémie à l'inclusion et à 6 mois traitées comme deux groupes). Dans Google Sheets, =T.TEST(S2:S161;U2:U161;2;2). Comparez la valeur p à celle du test apparié et expliquez la différence (chapitre 6, section 6.3.3).

L6.3 Test t pour échantillons indépendants dans JASP

La question : la PAS diffère-t-elle entre les femmes et les hommes ? H0 : $\mu F = \mu M$; H1 : $\mu F \neq \mu M$.

1. T-Tests > Independent Samples T-Test. Placez tas0 dans Variables et sex dans Grouping Variable.
2. Sous Tests, cochez Student, Welch et Mann-Whitney.
3. Sous Additional Statistics, cochez Location parameter, Confidence interval, Effect size (Cohen's d), Descriptives et Descriptives plots.
4. Sous Assumption Checks, cochez Normality et Equality of variances (test de Levene).

Ce que vous devez obtenir : femmes $131{,}37 \pm 14{,}47$ mmHg$(n = 86)$, hommes $130{,}85 \pm 13{,}25$ mmHg$(n = 74)$; Shapiro-Wilk p = 0,426 (F) et 0,648 (M) ; Levene p ≈ 0,52 (les variances peuvent être considérées comme égales) ; test de Student t = 0,24, ddl = 158, p = 0,81 ; test de Welch t = 0,24, ddl = 157,4, p = 0,81 ; différence F - M de 0,52 mmHg (IC à 95% d'environ -3,8 à 4,9) ; d = 0,04 ; Mann-Whitney p = 0,80. JASP calcule la différence dans l'ordre alphabétique des groupes (F moins M), c'est pourquoi le signe est positif ; dans le chapitre, nous avons rapporté M moins F, soit -0,5 mmHg. Le signe ne change pas la conclusion, mais il doit être précisé dans le texte.

5. Refaites l'analyse pour glicemie0 selon le sexe. Valeurs attendues : F 104,81 ± 16,30, M 106,16 ± 15,19 ; Levene p ≈ 0,43 ; Student t = -0,54, ddl = 158, p = 0,591 ; Welch p = 0,589 ; différence F - M de -1,35 mg/dL (IC à 95% -6,27 à 3,57) ; d = -0,09. Attention à la vérification de la normalité : dans le groupe des femmes, Shapiro-Wilk p = 0,022. Avec 86 patientes et un écart modéré, le test t reste acceptable ; le test de Mann-Whitney (p = 0,44) confirme la conclusion. Écrivez les deux résultats dans votre devoir.
6. Refaites l'analyse pour inaltime selon le sexe : différence F - M de -13,37 cm (IC à 95% -15,28 à -11,45), t = -13,79, ddl = 158, p < 0,001, d ≈ -2,19 ; Levene p > 0,7. Un effet énorme, également évident sur la boîte à moustaches.
7. Refaites l'analyse pour varsta selon le sexe : F 55,42 ± 8,95, M 57,31 ± 8,50 ; Welch t = -1,37, p = 0,17 ; d = -0,22. Discutez des raisons pour lesquelles la vérification de l'âge selon le sexe importe avant l'interprétation d'autres différences entre les sexes (l'âge est un facteur de confusion possible).

L6.4 ANOVA à un facteur dans JASP

La question principale de l'étude : la baisse de la glycémie à 6 mois diffère-t-elle entre les bras A, B et C ? H0 : μA = μB = μC ; H1 : au moins une moyenne diffère.

1. ANOVA > ANOVA. Placez dglic dans Dependent Variable et brat dans Fixed Factors.
2. Sous Display, cochez Descriptive statistics et Estimates of effect size (η² et ω²).
3. Sous Assumption Checks, cochez Homogeneity tests (Levene) et Q-Q plot of residuals. Sous Homogeneity corrections, laissez None pour l'analyse principale ; cochez aussi Welch pour comparaison.
4. Sous Post Hoc Tests, placez brat dans la zone de droite et cochez Tukey ; ajoutez Games-Howell et Bonferroni pour comparaison ; cochez Confidence intervals et Effect size.
5. Sous Descriptives Plots, placez brat sur l'axe horizontal et cochez Display error bars, Confidence interval 95%.
6. Sous Nonparametrics, placez brat pour le test de Kruskal-Wallis.
7. Pour la normalité par groupe, utilisez Descriptives > Descriptive Statistics, avec dglic dans Variables et brat dans Split, et cochez Shapiro-Wilk test.

Ce que vous devez obtenir : A 3,0 ± 8,8 mg/dL$(n = 53)$, B $11{,}1 \pm 11{,}2(n = 53)$, C $16{,}3 \pm 10{,}6(n = 54)$. Tableau d'ANOVA : brat SS = 4806,9, ddl = 2, MS = 2403,4 ; Residuals SS = 16459,5, ddl = 157, MS = 104,8 ; F = 22,93, p < 0,001 ; η² = 0,226 ; ω² = 0,215. Test de Levene : p = 0,089 ou p = 0,163, selon la variante du test utilisée par le logiciel (centrée sur la moyenne ou sur la médiane) ; dans les deux cas, supérieur à 0,05. ANOVA de Welch : F = 26,05, p < 0,001. Shapiro-Wilk par bras : A p = 0,575, B p = 0,052, C p = 0,903. Tukey : B - A 8,1 mg/dL (IC à 95% 3,4-12,8 ; p < 0,001), C - A 13,3 (8,6-18,0 ; p < 0,001), C - B 5,2 (0,6-9,9 ; p = 0,024). Games-Howell : p = 0,0002, < 0,001 et 0,038. Kruskal-Wallis : H = 36,63, ddl = 2, p < 0,001.

8. Refaites l'analyse pour dgr (perte de poids). Valeurs attendues : A 0,71 ± 2,48 kg, B 2,84 ± 2,36, C 3,37 ± 2,39 ; F(2 ; 157) = 18,25, p < 0,001 ; η² = 0,189 ; Levene p = 0,79 ; Tukey : B - A 2,13 kg

(p < 0,001), C - A 2,66 kg (p < 0,001), C - B 0,53 kg (p = 0,49) ; Kruskal-Wallis H = 27,50, p < 0,001.

9. Refaites l'analyse pour tas0 avec fumat comme facteur. Valeurs attendues : anciens fumeurs $127{,}55 \pm 14{,}39 (n = 33)$, fumeurs $137{,}04 \pm 14{,}63 (n = 53)$, non-fumeurs $128{,}50 \pm 11{,}76 (n = 74)$; F(2 ; 157) = 7,85, p < 0,001 (0,0006) ; η^2 = 0,091 ; Levene p ≈ 0,11 ; Tukey : fumeurs - anciens fumeurs 9,49 mmHg (p = 0,004), fumeurs - non-fumeurs 8,54 mmHg (p = 0,001), non-fumeurs - anciens fumeurs 0,95 mmHg (p = 0,94).
10. Interprétez sur papier, en 4-5 phrases, le résultat du point 7 : quels bras diffèrent, quelle est l'ampleur de l'effet, que signifie l'intervalle de confiance de la différence C - B qui commence à 0,6 mg/dL.

L6.5 ANOVA à mesures répétées dans JASP

La question : la glycémie change-t-elle entre 0, 3 et 6 mois ? H0 : μ0 = μ3 = μ6.

1. ANOVA > Repeated Measures ANOVA. Sous Repeated Measures Factors, renommez le facteur "temps" et les niveaux "0 mois", "3 mois", "6 mois".
2. Placez glicemie0, glicemie3 et glicemie6 dans Repeated Measures Cells, dans le bon ordre.
3. Sous Assumption Checks, cochez Sphericity tests, et sous Sphericity corrections, cochez None et Greenhouse-Geisser.
4. Sous Post Hoc Tests, placez temps et cochez Bonferroni. Sous Descriptives Plots, placez temps sur l'axe horizontal.

Ce que vous devez obtenir : moyennes de 105,4 / 99,4 / 95,2 mg/dL ; F(2 ; 318) = 64,97, p < 0,001 ; test de Mauchly W = 0,873, p < 0,001 (la sphéricité n'est pas respectée) ; ε de Greenhouse-Geisser = 0,887, degrés de liberté corrigés 1,77 et 282,2, p < 0,001 ; comparaisons de Bonferroni : p < 0,001 pour les trois paires. Test de Friedman (ANOVA > Repeated Measures ANOVA > Nonparametrics, ou chapitre 7) : χ^2 = 74,67, p < 0,001.

5. Modèle mixte (en option, pour les étudiants qui terminent plus vite) : ajoutez brat dans Between Subject Factors. Valeurs attendues : effet du temps F(2 ; 314) = 73,56, p < 0,001 ; effet du bras F(2 ; 157) = 8,50, p < 0,001 ; interaction temps × bras F(4 ; 314) = 11,51, p < 0,001. Moyennes à 0 / 3 / 6 mois : A 107,8 / 105,7 / 104,8 ; B 106,3 / 101,3 / 95,2 ; C 102,3 / 91,2 / 86,0 mg/dL. Une interaction significative signifie que l'évolution dans le temps diffère entre les bras : c'est là, en réalité, la question de l'efficacité des interventions. Tracez le graphique des moyennes avec une ligne par bras (Descriptives Plots, brat dans Separate lines).

L6.6 Les mêmes analyses dans MedCalc

1. Importez mg2026_curat.csv (File > Import ou File > Open, type CSV) et vérifiez que les variables numériques sont reconnues comme des nombres.
2. Le test apparié : Statistics > T-tests > Paired samples t-test ; Variable 1 : glicemie0, Variable 2 : glicemie6. Le résultat doit coïncider avec celui de L6.2 (différence moyenne, IC à 95%, t = 11,14, p < 0,0001).
3. Le test pour échantillons indépendants : Statistics > T-tests > Independent samples t-test. Pour Sample 1, choisissez la variable tas0 avec le critère de sélection sex="F", et pour Sample 2, la même variable avec sex="M". MedCalc rapporte aussi le test F d'égalité des variances, le test t classique et la variante corrigée pour variances inégales.
4. ANOVA : Statistics > ANOVA > One-way analysis of variance. Pour Data, vous pouvez écrire directement l'expression glicemie0-glicemie6 (ou une variable calculée au préalable) ; pour Factor, choisissez brat. Cochez le test de Levene et un test post hoc (Tukey-Kramer ou Bonferroni). Comparez le tableau avec celui de JASP.
5. À partir de données résumées, sans fichier : le menu Tests > Comparison of means (t-test) permet de saisir la moyenne, l'ÉT et n pour deux groupes. Saisissez les valeurs d'IMC d'un article publié,

$28{,}15 \pm \frac{6{,}49\ \mathrm{kg}}{m^2} (n = 52)$ et $24{,}47 \pm \frac{4{,}60\ \mathrm{kg}}{m^2} (n = 84)$ [71], et comparez le résultat à la valeur p rapportée par les auteurs (p = 0,001). Vous devriez obtenir un t d'environ 3,6-3,9 (selon la variante du test) et p < 0,001.

L6.7 Tests t et tableau d'ANOVA dans Google Sheets

La fonction T.TEST(plage1; plage2; queues; type) donne directement la valeur p. L'argument "queues" vaut 1 ou 2 (utilisez 2, bilatéral). L'argument "type" vaut 1 pour des échantillons appariés, 2 pour des échantillons indépendants de variances égales (Student) et 3 pour des échantillons indépendants de variances inégales (Welch).

1. Le test apparié : =T.TEST(S2:S161;U2:U161;2;1). Résultat : environ 1,1 × 10^-21, soit p < 0,001.
2. Le test pour échantillons indépendants selon le sexe, avec FILTER pour séparer les groupes : =T.TEST(FILTER(M2:M161;C2:C161="F");FILTER(M2:M161;C2:C161="M");2;2) donne 0,814 (Student), et avec le type 3, 0,813 (Welch).
3. La statistique t du test apparié, construite à la main : dans des cellules séparées, calculez la moyenne des différences =AVERAGE(AC2:AC161) (10,19), l'écart-type =STDEV(AC2:AC161) (11,57), l'erreur standard =STDEV(AC2:AC161)/SQRT(160) (0,915), t = moyenne / ES (11,14), p =T.DIST.2T(t;159) et les bornes de l'IC à 95% = moyenne ± T.INV.2T(0,05;159) × ES (8,38 et 11,99).
4. Le test F du rapport des variances : =F.TEST(FILTER(M2:M161;C2:C161="F");FILTER(M2:M161;C2:C161="M")) donne environ 0,44. Le test F est sensible aux écarts à la normalité ; le test de Levene, dans JASP, est préférable, mais tous deux conduisent ici à la même conclusion.
5. Le tableau d'ANOVA pour dglic selon les bras, construit avec des formules. Dans une nouvelle feuille, écrivez en A2:A4 les lettres A, B, C. En B2 =COUNTIF(donnees!B$2:B$161;A2) (n par bras : 53, 53, 54). En C2 =AVERAGEIF(donnees!B$2:B$161;A2;donnees!AC$2:AC$161) (moyennes 3,02 ; 11,09 ; 16,33). En D2 =DEVSQ(FILTER(donnees!AC$2:AC$161;donnees!B$2:B$161=A2)) (somme des carrés des écarts dans chaque bras : 3987,0 ; 6532,5 ; 5940,0). Faites glisser les formules jusqu'à la ligne 4.
6. Moyenne générale : =AVERAGE(donnees!AC2:AC161) (10,19). SS totale : =DEVSQ(donnees!AC2:AC161) (21266,4). SS intra-groupes : =SUM(D2:D4) (16459,5). SS inter-groupes : =SUMPRODUCT(B2:B4;(C2:C4-moyenne générale)^2) ou, plus simplement, SS totale - SS intra-groupes (4806,9).
7. MS inter = SS inter / 2 = 2403,4 ; MS intra = SS intra / 157 = 104,8 ; F = MS inter / MS intra = 22,93 ; p =F.DIST.RT(F;2;157), environ 1,8 × 10^-9 ; valeur critique =F.INV.RT(0,05;2;157) = 3,05 ; êta² = SS inter / SS totale = 0,226. Disposez les résultats dans un tableau avec les colonnes Source, SS, ddl, MS, F, p, exactement comme le tableau 6.5 du chapitre.
8. Variante avec module complémentaire : depuis Extensions > Add-ons, on peut installer un module complémentaire d'analyse statistique (par exemple, XLMiner Analysis ToolPak), qui propose "Anova: Single Factor". Ce module exige les données de chaque groupe dans une colonne distincte ; vous pouvez les obtenir avec =FILTER(donnees!AC2:AC161;donnees!B2:B161="A") et les formules analogues pour B et C. Le résultat doit coïncider avec le tableau construit à la main.

L6.8 Le devoir au format a-g : le modèle pour l'ANOVA

(a) Hypothèse de recherche : la baisse de la glycémie à jeun à 6 mois diffère entre les patients ayant reçu un conseil standard, un conseil associé à une application et un conseil associé à une application et à une téléconsultation.

(b) H0 : la baisse moyenne de la glycémie est la même dans les trois populations, μA = μB = μC. H1 : au moins un bras présente une baisse moyenne différente des autres. α = 0,05.

(c) Méthode : variable numérique (dglic) comparée entre trois groupes indépendants ; ANOVA à un facteur, suivie du test de Tukey ; conditions vérifiées : indépendance (randomisation individuelle), normalité par bras (Shapiro-Wilk p compris entre 0,052 et 0,903), homogénéité des variances (Levene $p > 0,05$, rapport des ÉT 1,27). Alternative en cas d'écart : ANOVA de Welch et Kruskal-Wallis.

(d) Le logiciel : JASP, ANOVA > ANOVA, avec Post Hoc Tests Tukey et Descriptives Plots ; vérification dans Google Sheets avec le tableau construit en L6.7.

(e) Résultat : A 3,0 ± 8,8, B 11,1 ± 11,2, C 16,3 ± 10,6 mg/dL ; $F(2 ; 157) = 22,93$, $p < 0,001$, $\eta^2 = 0,226$; Tukey B - A 8,1 (IC à 95% 3,4-12,8), C - A 13,3 (8,6-18,0), C - B 5,2 (0,6-9,9), tous $p < 0,05$. Le graphique : boîtes à moustaches par bras ou graphique des moyennes avec IC à 95%.

(f) Conclusion statistique : $p < 0,001$, on rejette H0 ; les différences entre les bras sont extrêmement significatives, et les trois paires diffèrent toutes significativement.

(g) Conclusion de recherche : l'ajout de l'application d'automesure au conseil a augmenté la baisse de la glycémie à jeun à 6 mois d'environ 8 mg/dL, et la téléconsultation mensuelle a apporté un bénéfice supplémentaire d'environ 5 mg/dL par rapport à l'application, avec une estimation moins précise (IC à 95% 0,6-9,9 mg/dL). Le bras de traitement explique près d'un quart de la variabilité de la baisse de la glycémie. L'étude est randomisée, les différences peuvent donc être attribuées aux interventions ; les données sont toutefois simulées à des fins pédagogiques.

Exercices supplémentaires

Les exercices 1-6 se travaillent sur les fichiers classiques de la discipline disponibles sur Moodle ; les résultats ne sont pas fournis, et chacun se résout au format a-g, avec la vérification des conditions d'application et le graphique approprié. Les exercices 7 et 8 comportent les données dans l'énoncé et une solution complète.

1. Le fichier BD_lp3. Existe-t-il des différences significatives entre la glycémie au temps zéro et la glycémie à 3 mois ? Et entre 3 et 6 mois ? Traitez ensuite la question sur les trois temps de mesure avec une ANOVA à mesures répétées et comparez les conclusions.
2. Le fichier BD_lp3. La glycémie à 6 mois diffère-t-elle entre les patients assurés et non assurés ? L'âge diffère-t-il entre les deux groupes ? Si l'âge diffère, comment pourrait-il influencer l'interprétation de la différence de glycémie ?
3. Le fichier BD_lp3. L'évolution de la glycémie à 0, 3 et 6 mois diffère-t-elle entre assurés et non-assurés ? Utilisez le modèle mixte de L6.5 (le statut d'assuré comme facteur inter-sujets) et interprétez le terme d'interaction.
4. Le fichier BPOC. La fréquence cardiaque a-t-elle baissé significativement après le traitement (FC1 par rapport à FC2) ? Le VEMS en pourcentage a-t-il augmenté (VEMSP1 par rapport à VEMSP2) ? Vérifiez la normalité des différences et rapportez, le cas échéant, le test de Wilcoxon.
5. Le fichier BPOC. L'indice de perméabilité bronchique initial (IPB1) diffère-t-il entre les patients ayant reçu les médicaments A, B et C ? Appliquez l'ANOVA avec le test de Levene et le test post hoc approprié ; si les variances diffèrent, utilisez l'ANOVA de Welch et le test de Games-Howell.
6. Le fichier BD_Studenti. (a) La taille diffère-t-elle entre les sexes ? (b) La taille réelle diffère-t-elle de la taille idéale déclarée (test apparié) ? (c) La taille diffère-t-elle entre les groupes d'étudiants ? Pour (c), discutez des raisons pour lesquelles un résultat significatif pourrait ne refléter que la composition des groupes selon le sexe.

7. Données dans l'énoncé, avec solution. Fréquence cardiaque (battements/min) chez 6 patients atteints de BPCO avant et après un bêtabloquant cardiosélectif : avant 96, 88, 102, 90, 94, 100 ; après 84, 82, 90, 86, 85, 91. Les différences sont 12, 6, 12, 4, 9, 9 ; $m_D = \frac{52}{6} = 8{,}67$; écarts à la moyenne : 3,33 ; -2,67 ; 3,33 ; -4,67 ; 0,33 ; 0,33 ; somme des carrés 11,11 + 7,11 + 11,11 + 21,78 + 0,11 + 0,11 = 51,33 ; $s_D = \sqrt{\frac{51{,}33}{5}} = 3{,}20$; ESD = $\frac{3{,}20}{\sqrt{6}} = 1{,}31$; $t = \frac{8{,}67}{1{,}31} = 6{,}63$, ddl = 5 ; t critique (5 ; 0,05) = 2,571, donc p < 0,05 (valeur exacte p = 0,0012). IC à 95% : 8,67 ± 2,571 × 1,31 = 5,30-12,03 battements/min. Conclusion : la fréquence cardiaque a baissé significativement, en moyenne d'environ 9 battements/min ; avec 6 patients et sans groupe témoin, le résultat n'est qu'indicatif.
8. Données dans l'énoncé, avec solution. L'hémoglobine (g/dL) dans trois groupes de 4 femmes enceintes chacun, aux premier, deuxième et troisième trimestres (groupes indépendants) : T1 12,5 ; 13,0 ; 12,8 ; 12,9 ; T2 11,8 ; 12,0 ; 11,6 ; 12,2 ; T3 11,5 ; 11,9 ; 11,2 ; 11,8. Moyennes : 12,80 ; 11,90 ; 11,60 ; moyenne générale 12,10. SS inter = 4 × [(0,70)² + (-0,20)² + (-0,50)²] = 4 × 0,78 = 3,12, ddl = 2, MS = 1,56. SS intra : T1 0,0900 + 0,0400 + 0,0000 + 0,0100 = 0,14 ; T2 0,0100 + 0,0100 + 0,0900 + 0,0900 = 0,20 ; T3 0,0100 + 0,0900 + 0,1600 + 0,0400 = 0,30 ; total 0,64, ddl = 9, MS = 0,0711. $F = \frac{1{,}56}{0{,}0711} = 21{,}9$; F critique (2 ; 9 ; 0,05) = 4,26, donc p < 0,05 (valeur exacte p < 0,001). $\eta^2 = \frac{3{,}12}{3{,}76} = 0{,}83$. Test de Tukey : q(3 ; 9 ; 0,05) = 3,95 ; $\text{HSD} = 3{,}95 \times \sqrt{\frac{0{,}0711}{4}} = 3{,}95 \times 0{,}133 = 0{,}53$ g/dL ; T1 - T2 = 0,90 et T1 - T3 = 1,20 sont significatives, T2 - T3 = 0,30 ne l'est pas. Conclusion : l'hémoglobine moyenne est plus élevée au premier trimestre qu'aux deuxième et troisième trimestres, ce qui concorde avec l'hémodilution physiologique de la grossesse ; la différence entre les deuxième et troisième trimestres n'a pas pu être mise en évidence sur cet échantillon très petit.

Ce qu'il faut rendre et barème indicatif

Il faut rendre sur Moodle un seul fichier contenant : les résultats du test apparié pour glicemie0-glicemie6 (le tableau de JASP, le test de normalité sur les différences, le graphique) avec leur interprétation (2 points) ; les résultats du test pour échantillons indépendants pour tas0 et glicemie0 selon le sexe, avec le test de Levene et la justification du choix entre Student, Welch et Mann-Whitney (2 points) ; l'ANOVA pour dglic selon les bras, avec la vérification des conditions, le test de Tukey et le graphique des moyennes (2 points) ; le tableau d'ANOVA construit dans Google Sheets, avec les formules visibles (1,5 point) ; l'ANOVA à mesures répétées avec le test de Mauchly et la correction de Greenhouse-Geisser (1 point) ; un devoir au format a-g sur l'un des exercices supplémentaires 1-6 (1,5 point). Total : 10 points. Sont pénalisés : un test pour échantillons indépendants appliqué à des données appariées, une normalité vérifiée sur chaque série au lieu des différences pour le test apparié, des tests t multiples à la place de l'ANOVA, l'absence de test post hoc après un F significatif, l'absence de taille de l'effet et d'intervalle de confiance, la formulation "on accepte H0".

Variante hors ligne

Sans réseau, JASP et MedCalc fonctionnent en local, après installation ; le fichier mg2026_curat.csv se copie à partir de la clé USB de l'enseignant. La partie Google Sheets est remplacée par LibreOffice Calc ou Excel : les fonctions T.TEST, T.DIST.2T, T.INV.2T, DEVSQ, AVERAGEIF, COUNTIF, F.DIST.RT et F.INV.RT y existent aussi (dans Excel, FILTER est disponible dans les versions récentes ; sinon, triez les données par bras et sélectionnez

manuellement les plages de chaque groupe). Dans Excel, l'ANOVA s'obtient aussi par Data > Data Analysis > Anova: Single Factor, après activation du complément Analysis ToolPak. Si aucun ordinateur n'est disponible, les exercices supplémentaires 7 et 8 se résolvent entièrement à la main, avec les tables des valeurs critiques de t et de F et avec la table q pour le test de Tukey.

Travaux pratiques 7. Tests non paramétriques : Mann-Whitney, Wilcoxon, Kruskal-Wallis, Friedman et khi-deux

Objectif

Dans cette séance, vous appliquez sur ordinateur les tests du chapitre 7, sur le jeu de données mg2026, dans trois logiciels différents, et vous vérifiez que vous obtenez les mêmes résultats. L'accent ne porte pas sur l'action d'appuyer sur des boutons, mais sur trois points qui compteront à l'examen pratique : le choix du test adapté au type de variable et de schéma d'étude, la lecture correcte des tableaux de sortie (quelle statistique, quelle valeur p, quelle taille de l'effet) et la formulation de la conclusion. Chaque étape vous donne les chiffres que vous devez obtenir, afin que vous puissiez vous autovérifier ; si votre résultat diffère, cherchez l'erreur avant d'aller plus loin.

À la fin, l'étudiant doit être capable :

- d'appliquer dans JASP les tests de Mann-Whitney, de Wilcoxon, de Kruskal-Wallis (avec post hoc de Dunn) et de Friedman et d'en extraire la statistique, la valeur p et la taille de l'effet ;
- de construire dans JASP un tableau de contingence avec les effectifs attendus, le test du khi-deux, la correction de Yates, le test de Fisher et le V de Cramér ;
- de saisir un tableau 2 x 2 et un tableau 2 x n dans Epi Info StatCalc et d'interpréter la sortie (khi-deux non corrigé, Yates, Fisher, OR, RR, test de tendance) ;
- de reproduire dans Google Sheets les effectifs observés, les effectifs attendus et la valeur p du test du khi-deux ;
- de vérifier les conditions d'application (effectifs attendus, données appariées ou indépendantes) et de choisir la variante correcte du test ;
- de formuler la conclusion statistique et la conclusion clinique pour chaque test.

Outils

- JASP 0.98.1 [47], modules par défaut T-Tests, ANOVA et Frequencies.
- MedCalc 24.0 [50] (sur les ordinateurs du laboratoire).
- Epi Info 7.2.7 [51], composant StatCalc et, en option, Classic Analysis.
- Google Sheets, avec les paramètres régionaux pour la Roumanie (le séparateur d'arguments est ";" et le séparateur décimal est la virgule).
- Le fichier mg2026_curat.csv (160 patients, 28 colonnes), disponible sur la plateforme Moodle de la discipline.
- Durée indicative : 100 minutes (JASP 50 minutes, Epi Info 15 minutes, Google Sheets 20 minutes, conclusions et dépôt 15 minutes).

Les colonnes du fichier sont, dans l'ordre : A id, B brat, C sex, D varsta, E mediu, F educatie, G fumat, H activitate, I inaltime, J greutate0, K imc0, L talie, M tas0, N tad0, O fc0, P colesterol, Q hdl, R trigliceride, S glicemie0, T glicemie3, U glicemie6, V hba1c0, W findrisc, X diabet_ogtt, Y greutate6, Z tas6, AA scadere5, AB satisfactie. Les données occupent les lignes 2-161.

Déroulement

L7.1 Préparation des données dans JASP

1. Ouvrez JASP et chargez le fichier : menu principal (les trois traits en haut à gauche) > Open > Computer > Browse, puis sélectionnez mg2026_curat.csv.
2. Vérifiez le type de chaque variable, indiqué par l'icône à gauche du nom de la colonne (une règle pour une variable continue, des barres en escalier pour une variable ordinale, trois cercles pour une variable nominale). Cliquez sur l'icône pour le modifier. Définissez brat, sex, mediu, fumat, diabet_ogtt et scadere5 comme **nominales** ; educatie, activitate et satisfactie comme **ordinales** ; les autres comme **continues**.
3. Créez la variable "dglic" (baisse de la glycémie à 6 mois) : cliquez sur le signe "+" à côté de la dernière colonne, nommez-la dglic, choisissez le type continu et la méthode de calcul par formule R, puis écrivez glicemie0 - glicemie6 et appuyez sur Compute column. Vérification : la moyenne de la colonne dglic doit être de 10,19.
4. De la même manière, créez la variable "obez", nominale, avec la formule R ifelse(imc0 >= 30, 1, 0). Vous devez obtenir 54 valeurs 1 et 106 valeurs 0.
5. Créez la variable "glic100_0" avec la formule ifelse(glicemie0 >= 100, 1, 0) et "glic100_6" avec la formule ifelse(glicemie6 >= 100, 1, 0). Vous devez obtenir respectivement 97 et 62 valeurs 1.

Pourquoi le réglage du type de variable est-il important ? JASP ne vous laisse pas placer une variable continue dans la case "Rows" d'un tableau de contingence et, inversement, ne propose que les tests adaptés au type déclaré. Un type erroné est la première raison pour laquelle "le test n'apparaît pas" en séance de travaux pratiques.

L7.2 Le test U de Mann-Whitney dans JASP : la satisfaction dans le bras C par rapport au bras A

Le test de Mann-Whitney compare deux groupes ; vous devez donc exclure temporairement le bras B.

1. Cliquez sur le nom de la colonne brat ; dans la fenêtre des étiquettes qui s'ouvre, décochez "Filter" pour la valeur B (ou utilisez le bouton de filtre, le symbole en forme d'entonnoir). Les lignes des patients du bras B deviennent grises : elles sont exclues des analyses. Vérifiez dans la barre du bas : 107 lignes actives.
2. T-Tests > Classical > Independent Samples T-Test.
3. Dependent Variables : satisfactie (si JASP n'accepte pas la variable ordinale, définissez-la temporairement comme "Scale"). Grouping Variable : brat.
4. Dans Tests, décochez Student et cochez **Mann-Whitney**. Dans Additional Statistics, cochez **Effect size** (pour le test de Mann-Whitney, JASP rapporte la corrélation bisériale des rangs) et **Descriptives**.
5. Pour la description, utilisez Descriptives > Descriptive Statistics, avec satisfactie dans Variables et brat dans Split, en cochant Median et Quartiles (ou Interquartile range).

Ce que vous devez obtenir : médiane 3 (Q1 2, Q3 3) dans le bras A et 4 (Q1 3, Q3 5) dans le bras C ; W (la statistique de Mann-Whitney rapportée par JASP) = 596,5 ou 2265,5, selon le groupe pris comme référence (l'ordre A, C donne 596,5) ; $p < 0{,}001$; corrélation bisériale des rangs de 0,58 en valeur absolue (le signe dépend de l'ordre des groupes). Notez sur la fiche de travail : "$U = \frac{2265{,}5}{596{,}5}$ sont la même statistique, vue depuis chacun des deux groupes ; U1 + U2 = 54 x 53 = 2862".

6. Recommencez pour les bras B et A (filtrez maintenant le bras C). Vous devez obtenir $p < 0{,}001$ et une corrélation bisériale des rangs de 0,42. Puis pour C et B : $p = 0{,}079$, corrélation de 0,19.

7. À la fin, réactivez tous les bras (cochez de nouveau "Filter" pour B et C). Oublier un filtre actif est une source classique de résultats erronés dans les exercices suivants.

Interprétation écrite (sur la fiche de travail) : "La satisfaction a été plus élevée dans le bras C que dans le bras A (médiane 4 contre 3 ; Mann-Whitney p < 0,001 ; corrélation bisériale des rangs 0,58, effet important)."

L7.3 Le test de Wilcoxon pour données appariées dans JASP

Exercice A : la glycémie à 0 et à 6 mois, tous les patients.

1. T-Tests > Classical > Paired Samples T-Test.
2. Placez glicemie0 et glicemie6 dans la case Variable Pairs (elles forment une paire).
3. Dans Tests, cochez à la fois Student et **Wilcoxon signed-rank**. Dans Additional Statistics, cochez Effect size et Descriptives, et dans Assumption Checks, Normality (Shapiro-Wilk sur les différences).

Ce que vous devez obtenir : test t apparié t = 11,14, ddl 159, p < 0,001 ; Shapiro-Wilk sur les différences p = 0,725 (les différences suivent une loi normale, le test t est donc l'analyse principale) ; Wilcoxon p < 0,001. Lorsque les deux tests concordent, rapportez le test choisi à l'avance ; ici, le test t apparié, avec une différence moyenne de 10,19 mg/dL (IC à 95% 8,38-11,99).

Exercice B : le poids à 0 et à 6 mois, bras A uniquement.

1. Filtrez les bras B et C (il reste 53 lignes actives).
2. Paired Samples T-Test avec la paire greutate0 - greutate6, en cochant Student, Wilcoxon signed-rank et Normality.

Ce que vous devez obtenir : Wilcoxon W = 870 (somme des rangs positifs pour la différence greutate0 - greutate6 ; si vous inversez l'ordre des variables, vous obtenez 456), p = 0,052 ; test t apparié t = 2,07, ddl 52, p = 0,043 ; Shapiro-Wilk sur les différences p = 0,696. La médiane des différences est de 1,1 kg, la moyenne de 0,71 kg.

Discutez sur la fiche de travail, en 4-5 lignes : quel test est le bon ici et pourquoi (la distribution des différences est normale, le test t apparié est donc justifié) ; pourquoi le test de Wilcoxon donne une valeur p un peu plus grande (il n'utilise que les rangs, son efficacité relative est d'environ 95%) ; pourquoi nous n'avons pas le droit de choisir le test après avoir vu quelle valeur p est inférieure à 0,05.

3. En option, le test des signes : dans la colonne des différences, comptez combien sont positives (32), négatives (19) et nulles (2). Dans Frequencies > Binomial Test, sur une variable qui code le signe de la différence, avec Test value 0,5, vous devez obtenir p = 0,092.

L7.4 Le test de Kruskal-Wallis avec post hoc de Dunn dans JASP

1. Réactivez toutes les lignes. ANOVA > Classical > ANOVA.
2. Dependent Variable : dglic ; Fixed Factors : brat.
3. Ouvrez la section **Nonparametrics** et placez brat dans la case du test de Kruskal-Wallis. Cochez **Dunn's post hoc test** (dans certaines versions, Dunn apparaît dans la section Post Hoc Tests, sous Type : Dunn).
4. À titre de comparaison, demandez aussi l'ANOVA classique, avec Assumption Checks > Homogeneity tests et Post Hoc Tests > Tukey.

Ce que vous devez obtenir : Kruskal-Wallis H = 36,63, ddl 2, p < 0,001 ; Dunn avec correction de Bonferroni : A contre B p = 0,0003, A contre C p < 0,001, B contre C p = 0,12 (non ajusté 0,041). ANOVA : F(2 ; 157) = 22,93, p < 0,001 ; Levene p = 0,163 ; Tukey C contre B p = 0,024.

5. Recommencez pour satisfactie (Dependent Variable : satisfactie). Vous devez obtenir H = 32,16, ddl 2, p < 0,001 ; Dunn Bonferroni : A contre B p = 0,0003, A contre C p < 0,001, B contre C p = 0,29.
6. Calculez à la main la taille d'effet $\varepsilon^2 = \frac{H}{N-1}$: 36,63/159 = 0,23 pour dglic et 32,16/159 = 0,20 pour la satisfaction.

Question de contrôle (répondez sur la fiche) : pourquoi la comparaison B-C est-elle significative avec Tukey (p = 0,024) mais pas avec Dunn et la correction de Bonferroni (p = 0,12) ? Quelle est l'analyse principale pour dglic, et pourquoi ?

L7.5 Le test de Friedman dans JASP

1. ANOVA > Classical > Repeated Measures ANOVA.
2. Dans la case Repeated Measures Factors, renommez le facteur "RM Factor 1" en "temps" et les niveaux en "0 mois", "3 mois", "6 mois".
3. Placez glicemie0, glicemie3 et glicemie6, dans cet ordre, dans la case Repeated Measures Cells.
4. Ouvrez la section **Nonparametrics**, placez le facteur temps dans la case du test de Friedman et cochez Conover's post hoc tests.
5. Pour la description : Descriptives > Descriptive Statistics sur les trois variables, avec Median et Quartiles.

Ce que vous devez obtenir : médianes de 104,5, 99,0 et 92,5 mg/dL ; Friedman χ^2 = 74,67, ddl 2, p < 0,001 ; Kendall's W = 0,23 (si JASP ne l'affiche pas, calculez-le : 74,67/(160 x 2)) ; ANOVA à mesures répétées F = 64,97, p < 0,001. Les comparaisons post hoc de Conover doivent montrer des différences significatives entre les trois temps de mesure ; à titre de vérification indépendante, les tests de Wilcoxon avec correction de Bonferroni du tableau 7.7 du chapitre 7 donnent des p ajustés < 0,001 (0-3 mois), 0,0007 (3-6 mois) et < 0,001 (0-6 mois).

6. Filtrez uniquement le bras A et recommencez. Vous devez obtenir un χ^2 de Friedman = 6,04, p = 0,049, avec des médianes de 111, 104 et 105 mg/dL. Discutez : le résultat global est à la limite ; reste-t-il une comparaison significative après correction ? (Réponse à vérifier : non, la plus petite valeur ajustée par Bonferroni est de 0,075.)

L7.6 Tableaux de contingence et test du khi-deux dans JASP

Exercice A : le bras et la perte de poids ≥ 5% (tableau 3 x 2).

1. Réactivez toutes les lignes. Frequencies > Classical > Contingency Tables.
2. Rows : brat ; Columns : scadere5.
3. Dans Statistics, cochez χ^2 et, sous Nominal, **Phi and Cramér's V**. Dans Cells, cochez Counts **Expected** et Percentages **Row**.

Ce que vous devez obtenir : effectifs observés 6/47, 16/37, 16/38 ; effectifs attendus 12,59/40,41 (A), 12,59/40,41 (B), 12,82/41,18 (C) ; χ^2 = 6,77, ddl 2, p = 0,034 ; V de Cramér = 0,21 ; pourcentages en ligne 11,3%, 30,2% et 29,6%. Vérifiez la condition d'application : le plus petit effectif attendu est de 12,6, le test est donc valide.

4. Le graphique : Descriptives > Descriptive Statistics, variable scadere5 avec Split sur brat, Plots > Distribution plots ; ou, plus simplement, construisez le diagramme en barres des pourcentages dans Google Sheets (étape L7.8). Retenez l'observation faite en séance : pour le khi-deux, la représentation graphique ne se fait pas depuis le menu Contingency Tables, mais depuis Descriptives.

Exercice B : l'obésité et le diabète (tableau 2 x 2).

1. Contingency Tables avec Rows : obez, Columns : diabet_ogtt.

2. Cochez χ^2, **χ^2 continuity correction** et, dans la section consacrée aux tableaux 2 x 2, **Log odds ratio** et **Fisher's exact test**. Dans Cells : Expected, Row percentages.

Ce que vous devez obtenir : observés 95/11 (non obèses) et 36/18 (obèses) ; attendus 86,79/19,21 et 44,21/9,79 ; χ^2 = 12,70, ddl 1, p < 0,001 ; avec correction de continuité 11,20, p = 0,0008 ; Fisher p = 0,0008 ; φ = 0,28. Log odds ratio = 1,46, donc $\mathrm{OR} = e^{1,46} = 4,32$(; IC 95% 1,86 − 10,03), indépendamment du fait que vous le calculiez comme obèses par rapport aux non obèses.

3. Écrivez sur la fiche pourquoi, pour ce tableau, les trois variantes (Pearson, Yates, Fisher) mènent à la même conclusion (tous les effectifs attendus dépassent 5, le plus petit étant de 9,79), et dans quelle situation elles auraient donné des conclusions différentes.

Exercice C : une situation dans laquelle le khi-deux n'est pas valide.

Filtrez uniquement les patients de moins de 50 ans (ajoutez une variable calculée ifelse(varsta < 50, 1, 0) et filtrez selon celle-ci), puis construisez le tableau obez x diabet_ogtt pour ce sous-groupe. Notez combien de cellules ont un effectif attendu inférieur à 5 et quel test vous choisissez. (Le sous-groupe compte 34 patients et seulement 2 cas de diabète ; le test approprié est celui de Fisher.)

L7.7 Les mêmes tests dans MedCalc

Ouvrez mg2026_curat.csv dans MedCalc (File > Open, type de fichier texte/CSV ; vérifiez que la virgule décimale est correctement reconnue). Les intitulés des menus peuvent différer légèrement d'une version à l'autre ; cherchez les tests par leur nom.

1. Mann-Whitney : menu Statistics, parmi les tests sur les rangs, Mann-Whitney test (independent samples). Variable : satisfactie ; Grouping/Filter : brat = "A" pour le premier échantillon et brat = "C" pour le second. Ce que vous devez obtenir : la même valeur p (< 0,001) que dans JASP.
2. Wilcoxon : Wilcoxon test (paired samples), avec greutate0 et greutate6, Filter : brat = "A". Ce que vous devez obtenir : p = 0,052.
3. Kruskal-Wallis : dans l'analyse de la variance, Kruskal-Wallis test, Variable : dglic (créez-la auparavant avec Variable editor ou écrivez directement l'expression glicemie0-glicemie6), Factor : brat. Ce que vous devez obtenir : H = 36,63, p < 0,001, ainsi que les comparaisons post hoc proposées par le logiciel.
4. Friedman : Friedman test, avec les trois colonnes de glycémie. Ce que vous devez obtenir : χ^2 = 74,67 (ou une valeur très proche, selon la correction pour les ex aequo), p < 0,001.
5. Khi-deux : dans les tableaux de contingence, Chi-squared test, avec brat et scadere5. Ce que vous devez obtenir : χ^2 = 6,77, ddl 2, p = 0,034. MedCalc propose aussi le test de tendance (Chi-squared test for trend), utile ici parce que les bras ont une intensité croissante : vous devez obtenir une valeur du χ^2 de tendance d'environ 4,9, avec p = 0,027.
6. McNemar : McNemar test, avec glic100_0 et glic100_6 (créées comme variables 0/1). Ce que vous devez obtenir : paires discordantes 36 (≥ 100 uniquement à l'inclusion) et 1 (≥ 100 uniquement à 6 mois), p < 0,001. La valeur de la statistique dépend de la variante affichée : 33,11 sans correction ou 31,24 avec correction.

L7.8 Tableaux 2 x 2 et 2 x n dans Epi Info StatCalc

Epi Info StatCalc travaille sur des tableaux déjà résumés : il n'a pas besoin du fichier de données, seulement des effectifs.

Tableau 2 x 2 : obésité x diabète.

1. Ouvrez Epi Info 7 > StatCalc > Tables (2 x 2, 2 x n).
2. Saisissez : exposés (obèses) avec la maladie (diabète) 18, exposés sans la maladie 36, non exposés avec la maladie 11, non exposés sans la maladie 95. Attention à l'orientation du tableau : dans

StatCalc, les lignes correspondent à l'exposition, les colonnes à la maladie, et la cellule en haut à gauche est "exposé et malade".

3. Lisez la sortie. Ce que vous devez obtenir : Chi-square uncorrected 12,70 (p = 0,0004) ; Chi-square Yates corrected 11,20 (p = 0,0008) ; Fisher exact, bilatéral, environ 0,0008 ; Odds ratio 4,32 (IC à 95% 1,86-10,03) ; Risk ratio 3,21 (IC à 95% 1,64-6,31). StatCalc affiche aussi une variante de Mantel-Haenszel du khi-deux, légèrement inférieure à la valeur non corrigée (12,62), ainsi que des intervalles mid-p ; ne les confondez pas avec les valeurs de JASP.
4. Interprétez, sur la fiche : le diabète est 3,2 fois plus fréquent chez les obèses (rapport de prévalences, étude transversale), et l'association est statistiquement significative.

Tableau 2 x n : le bras et la perte de poids ≥ 5%.

5. Dans le même module, ajoutez des colonnes afin d'obtenir trois niveaux d'exposition (A, B, C) et saisissez les cas (perte ≥ 5% : 6, 16, 16) et les non-cas (47, 37, 38). Ce que vous devez obtenir : un khi-deux d'environ 6,77, avec 2 degrés de liberté, p = 0,034.
6. StatCalc > Chi square for trend : saisissez les scores d'exposition 0, 1, 2 pour les bras A, B, C, avec les cas et les non-cas ci-dessus. Ce que vous devez obtenir : un khi-deux de tendance d'environ 4,9, p = 0,027 (Epi Info utilise la variante de Mantel du test, qui peut différer à la deuxième décimale). Expliquez sur la fiche pourquoi le test de tendance a 1 degré de liberté, et le test global 2.

En option, dans Classic Analysis : READ (le fichier CSV), puis TABLES brat scadere5 pour un tableau 3 x 2 avec le khi-deux affiché automatiquement, et MEANS dglic brat, qui affiche l'ANOVA, le test de Bartlett et le test de Kruskal-Wallis dans la même sortie.

L7.9 Le khi-deux pas à pas dans Google Sheets

Importez mg2026_curat.csv dans une nouvelle feuille (File > Import > Upload, Separator type : comma). Travaillez dans une deuxième feuille, nommée "chi".

1. Le tableau observé (le bras en lignes, scadere5 en colonnes). Dans les cellules A2:A4, écrivez A, B, C ; en B1 écrivez "oui", en C1 "non". En B2, écrivez la formule :

   =COUNTIFS(mg2026!$B$2:$B$161;$A2;mg2026!$AA$2:$AA$161;1)

   et en C2 :

   =COUNTIFS(mg2026!$B$2:$B$161;$A2;mg2026!$AA$2:$AA$161;0)

   Recopiez les formules vers le bas jusqu'à la ligne 4. Vous devez obtenir 6/47, 16/37, 16/38.
2. Les totaux : en D2 =SUM(B2:C2) (recopiez vers le bas), en B5 =SUM(B2:B4) (recopiez vers la droite), en D5 =SUM(D2:D4). Vous devez obtenir 53, 53, 54 en lignes, 38 et 122 en colonnes, 160 au total.
3. Le tableau attendu, en F2:G4. En F2, écrivez :

   =$D2*B$5/$D$5

   et recopiez la formule dans tout le bloc F2:G4. Observez les symboles $: la ligne du total de colonne et la colonne du total de ligne restent fixes. Vous devez obtenir 12,59/40,41, 12,59/40,41, 12,82/41,18.
4. La valeur p directement : =CHISQ.TEST(B2:C4;F2:G4). Vous devez obtenir 0,034. Attention : la fonction reçoit le tableau observé et le tableau attendu, non les données brutes, et n'applique aucune correction de continuité.
5. La statistique χ^2 : =SUMPRODUCT((B2:C4-F2:G4)^2/F2:G4). Vous devez obtenir 6,77. Vérifiez p : =CHISQ.DIST.RT(6,77;2) donne 0,034, et la valeur critique =CHISQ.INV.RT(0,05;2) donne 5,99.
6. V de Cramér : =SQRT(6,77/(160*1)), soit 0,21.

7. Le graphique : sélectionnez la colonne des pourcentages en ligne (calculez en H2 =B2/D2, au format pourcentage) et insérez un diagramme en barres (Insert > Chart, Column chart). Il doit ressembler à la figure 7.4 du chapitre 7.
8. Les rangs pour un test de Mann-Whitney à la main (en option) : pour l'exemple du chapitre 7 (les scores 2, 3, 3, 5, 6 et 4, 6, 7, 7, 8, 9 en A2:A12), écrivez en B2 =RANK.AVG(A2;$A$2:$A$12;1) et recopiez vers le bas. La somme des rangs des cinq premières valeurs, =SUM(B2:B6), doit être de 17,5, et U1 = 17,5 - 5 x 6/2 = 2,5.

Reprenez les étapes 1-6 pour le tableau obésité x diabète (en utilisant une colonne auxiliaire avec =IF(K2>=30;1;0) dans la feuille de données). Vous devez obtenir χ^2 = 12,70 et p = 0,0004 (sans correction de Yates).

Exercices supplémentaires

1. Dans le fichier BPOC (données de BPCO) sur Moodle, choisissez une variable ordinale ou numérique asymétrique et comparez-la entre les sexes avec le test de Mann-Whitney ; rapportez les médianes, U, p et la corrélation bisériale des rangs. Justifiez pourquoi vous n'avez pas utilisé le test t.
2. Dans le même fichier, comparez une variable mesurée au début et à la fin du traitement (par exemple une valeur de la fonction respiratoire au temps 1 et au temps 2) avec le test de Wilcoxon et avec le test t apparié. Vérifiez la normalité des différences et décidez quel test est le bon.
3. Dans le fichier BD_Studenti, étudiez l'association entre le sexe et l'opinion sur la biostatistique (tableau de contingence). Vérifiez les effectifs attendus ; si l'une des catégories compte trop peu d'étudiants, proposez un regroupement qui ait un sens et refaites le test.
4. Dans le fichier Date_HbA1c, comparez la proportion de patients qui atteignent la cible thérapeutique (HbA1c < 7,5%) entre la première et la dernière année. Attention : s'il s'agit des mêmes patients, le test approprié est celui de McNemar, et non le khi-deux habituel.
5. Problème avec données dans l'énoncé : dans une étude de cohorte menée chez des retraités, 30 des 80 personnes exposées aux poussières et 30 des 240 personnes non exposées ont présenté une baisse de la capacité pulmonaire. Construisez le tableau, calculez à la main les effectifs attendus et le χ^2, puis vérifiez dans Epi Info StatCalc.

   Solution : le total des personnes présentant une baisse est de 60 sur 320 (18,75%). Effectifs attendus : exposés $\frac{80\times60}{320} = 15$ avec baisse et 65 sans ; non exposés $\frac{240\times60}{320} = 45$ avec baisse et 195 sans. Observés : 30/50 et 30/210. |O - E| = 15 dans chaque cellule. $\chi^2 = \frac{225}{15} + \frac{225}{65} + \frac{225}{45} + \frac{225}{195} = 15 + 3{,}46 + 5 + 1{,}15 = 24{,}62$, ddl 1, p < 0,001. Le risque chez les exposés est de 30/80 = 37,5%, chez les non exposés de 30/240 = 12,5%, RR = 3,0. L'exposition aux poussières est un facteur de risque significatif de baisse de la capacité pulmonaire dans cet échantillon.

Ce qu'il faut rendre et barème indicatif

Vous déposez sur Moodle un seul fichier PDF (ou le document Google exporté en PDF) qui contient :

- les captures d'écran des tableaux de sortie de JASP pour les étapes L7.2-L7.6 (2 points) ;
- le tableau de vérification complété, avec les valeurs que vous avez obtenues à côté de celles de ce texte, et l'explication de toute différence (2 points) ;

- la sortie d'Epi Info pour le tableau 2 x 2 et pour le test de tendance, avec l'interprétation de l'OR et du RR (1,5 point) ;
- la feuille Google Sheets (lien ou capture) avec les tableaux observé et attendu et les formules visibles (1,5 point) ;
- les réponses aux questions de contrôle de L7.3, L7.4 et L7.6 (2 points) ;
- la conclusion statistique et clinique, en 2-3 phrases, pour chacun des cinq tests (1 point).

Total : 10 points. Points retirés pour : test appliqué sur des données mal filtrées (filtre resté actif par oubli), moyenne et écart-type rapportés pour la satisfaction, khi-deux rapporté sans vérification des effectifs attendus, conclusion "on accepte H0".

Variante hors ligne

Si le réseau ou la plateforme Moodle ne sont pas disponibles, le fichier mg2026_curat.csv est copié depuis la clé USB de l'enseignant ou depuis le dossier partagé du laboratoire. JASP, MedCalc et Epi Info fonctionnent en local, sans connexion. L'étape L7.9 (Google Sheets) est remplacée par le même calcul dans un tableur installé localement (LibreOffice Calc ou Excel), où les fonctions portent les mêmes noms (CHISQ.TEST, CHISQ.DIST.RT, RANK.AVG, COUNTIFS) et où le séparateur d'arguments reste ";" dans la configuration roumaine. Si aucun logiciel n'est disponible, on travaille à la main sur le tableau obésité x diabète (effectifs attendus, χ^2, ddl, comparaison avec 3,84) et sur l'exemple de Mann-Whitney de la figure 7.1, avec les tables de valeurs critiques de l'annexe de l'ouvrage. Le fichier des résultats est remis à la séance suivante.

Travaux pratiques 8. Choix du test statistique et présentation standardisée des résultats

Objectif

Les séances de travaux pratiques précédentes vous ont appris les tests un par un. Dans cette séance, on ne vous dit plus quel test appliquer. Vous recevez 15 questions de recherche formulées sur le jeu de données mg2026 et, pour chacune, vous devez choisir vous-mêmes le test, le justifier, l'appliquer dans le logiciel et rapporter le résultat dans le format standardisé de la discipline (a-g). Dans la dernière partie de la séance, vous évaluez le rapport d'un collègue selon une grille commune. C'est, en pratique, une répétition générale de l'examen pratique, et les erreurs commises ici sont les erreurs les moins coûteuses que vous ferez jamais.

À la fin, l'étudiant doit être capable :

- d'identifier, pour une question de recherche donnée, l'objectif de l'analyse, le type de la variable de résultat et le schéma d'étude (indépendant ou apparié, nombre de groupes) ;
- de choisir le test approprié et son alternative, à l'aide de l'algorithme du chapitre 8, et de vérifier les conditions d'application dans le logiciel ;
- d'appliquer le test dans JASP (ou MedCalc, Epi Info) et d'en extraire la statistique, les degrés de liberté, la valeur p et une taille d'effet ;
- de rédiger la solution complète au format a-g, avec un graphique approprié ;
- d'évaluer de manière critique, selon une grille, le rapport d'un collègue et de formuler des observations constructives ;
- de reconnaître les questions pour lesquelles le test statistique n'est pas la réponse appropriée (par exemple les comparaisons à l'inclusion dans un essai randomisé).

Outils

- JASP 0.98.1 [47] (le logiciel principal) ; MedCalc 24.0 [50] et Epi Info 7.2.7 [51] pour les vérifications ; Google Sheets pour les graphiques simples et pour le document du rapport.
- Le fichier mg2026_curat.csv sur Moodle, avec les variables calculées lors des travaux pratiques 7 (dglic, obez, glic100_0, glic100_6). Pour cette séance, créez en plus : dgr (greutate0 - greutate6), dtas (tas0 - tas6) et tas140_0, tas140_6 (ifelse(tas0 >= 140, 1, 0), et de même pour tas6).
- Le modèle de rapport (section L8.2), disponible sur Moodle sous forme de document Google.
- La figure 8.1 (l'algorithme de choix du test) et le tableau 8.1 (la synthèse des tests) du chapitre 8, ouverts à côté.
- Durée indicative : 100 minutes (révision 10 minutes, résolution des problèmes 60 minutes, évaluation par les pairs 20 minutes, discussion finale et dépôt 10 minutes).

Déroulement

L8.1 Révision de l'algorithme (10 minutes)

Avant l'ordinateur, travaillez sur papier. Pour chaque question de la liste de L8.3, remplissez seulement les quatre premières colonnes du tableau ci-dessous, sans ouvrir le logiciel. Ce n'est qu'après avoir décidé du test que vous passez à JASP. L'exercice vous oblige à choisir le test à partir du schéma d'étude, et non à partir de ce qui "sort".

Tableau L8.1. Fiche de décision (à remplir avant l'analyse).

N°	Objectif (différence / association / accord / prédiction)	Variable de résultat et son type	Schéma d'étude (groupes, indépendant ou apparié)	Test choisi	Alternative (si les conditions ne sont pas remplies)
1					
...					
15					

L'enseignant vérifie la fiche de décision de chaque étudiant avant le début des calculs (pour les grands groupes, par sondage). Un choix erroné se corrige maintenant, et non après 40 minutes de travail.

L8.2 Le modèle de rapport standardisé

Chaque problème est rédigé selon le modèle ci-dessous. Le texte total d'un problème ne doit pas dépasser une demi-page, plus le graphique. La concision fait partie de la consigne : un bon rapport clinique dit tout ce qu'il faut, rien de plus.

Tableau L8.2. Le modèle de rapport a-g.

Composante	Ce qu'on écrit	Exemple de formulation
(a) Hypothèse de recherche	une phrase, en langage clinique	"Les femmes ont un HDL-cholestérol plus élevé que les hommes."
(b) Hypothèses statistiques	H0 et H1 en mots et sous forme mathématique ; test bilatéral	"H0 : $\mu F = \mu M$, il n'y a pas de différence entre les moyennes de HDL ; H1 : $\mu F \neq \mu M$."
(c) Méthode et justification	type des variables ; schéma d'étude ; vérification des conditions avec les valeurs p des tests de vérification ; test choisi et alternative	"HDL numérique, sexe nominal à 2 catégories, groupes indépendants ; SW p = 0,41 et 0,48 ; Levene p = 0,034, donc test t de Welch."
(d) Calcul et logiciel	logiciel, menu, options cochées	"JASP, T-Tests > Independent Samples T-Test : Welch, Descriptives, Effect size, Assumption checks."
(e) Résultat, avec graphique	statistiques descriptives par groupe ; statistique du test, ddl, p ; taille d'effet avec IC à 95% ; graphique	"F 53,92 ± 11,70 ; M 44,41 ± 9,66 ; t = 5,63, ddl 157,7, p < 0,001 ; différence 9,51 (IC à 95% 6,18-12,85) ; d = 0,88 ; boîte à moustaches."
(f) Conclusion statistique	décision par rapport à H0	"p < 0,001, nous rejetons H0 : la différence est statistiquement significative."
(g) Conclusion de recherche	ce que cela signifie cliniquement, avec la taille d'effet et les limites	"Les femmes ont un HDL plus élevé de 9,5 mg/dL en moyenne ; effet important ; analyse transversale sur un échantillon de patients à risque cardiométabolique."

Règles de rédaction, vérifiées lors de l'évaluation :

- les valeurs p s'écrivent avec trois décimales (p = 0,034), et en dessous de 0,001 sous la forme "p < 0,001" ; jamais "p = 0,000" ;
- les variables numériques normales se décrivent par moyenne ± ÉT, les variables asymétriques et ordinales par médiane (EIQ), les variables nominales par n (%) ;

- la décision se formule "nous rejetons H0" ou "nous ne rejetons pas H0", jamais "nous acceptons H0" ;
- le graphique doit correspondre au type de données : boîte à moustaches ou graphique des moyennes avec IC pour les variables numériques, diagramme en barres (pourcentages) pour les variables nominales et ordinales, nuage de points pour l'association de deux variables numériques ;
- les décimales s'écrivent avec une virgule.

Un exemple de rapport complet, sur une question qui ne fait pas partie de la liste ci-dessous, montre quelle doit être, en réalité, la longueur d'une bonne réponse.

La question : la glycémie à jeun à l'inclusion diffère-t-elle entre les patients avec et sans diabète confirmé par HGPO ?

(a) Les patients chez lesquels l'HGPO a confirmé le diabète ont une glycémie à jeun plus élevée que ceux sans diabète. (b) H0 : la moyenne de la glycémie à jeun est la même dans les deux groupes (μdiabète = μsans diabète) ; H1 : les moyennes diffèrent (μdiabète ≠ μsans diabète), test bilatéral. (c) La glycémie est numérique, le diabète est nominal à deux catégories, les groupes sont indépendants et d'effectifs inégaux (29 et 131). Shapiro-Wilk par groupe : p = 0,198 (diabète) et p = 0,061 (sans diabète) ; Levene p = 0,457. Test choisi : test t pour échantillons indépendants ; alternative : Mann-Whitney. (d) JASP, T-Tests > Independent Samples T-Test ; Dependent Variables : glicemie0 ; Grouping Variable : diabet_ogtt ; options cochées Student, Welch, Mann-Whitney, Descriptives, Effect size, Assumption checks ; graphique depuis Descriptives (boîtes à moustaches par groupe). (e) Sans diabète $101{,}85 \pm 13{,}97$ mg/dL$(n\,131)$, avec diabète $121{,}62 \pm 13{,}19$ mg/dL$(n\,29)$; t = -6,96, ddl 158, p < 0,001 (Welch t = -7,22, ddl 43,1, p < 0,001) ; différence de 19,77 mg/dL (IC à 95% 14,25-25,28) ; d de Cohen = 1,43 ; Mann-Whitney p < 0,001 ; boîtes à moustaches jointes. (f) p < 0,001 < 0,05, donc nous rejetons H0 : la différence est statistiquement significative. (g) Les patients avec un diabète confirmé avaient une glycémie à jeun plus élevée d'environ 20 mg/dL en moyenne, un effet très important. Ce résultat était attendu, car la glycémie à jeun fait partie des critères diagnostiques ; si nous voulions savoir dans quelle mesure la glycémie à jeun identifie le diabète confirmé par HGPO, la question serait une question de prédiction, et la méthode appropriée serait la courbe ROC (AUC 0,844 dans mg2026, chapitre 12).

L8.3 Les 15 problèmes

Résolvez les problèmes dans l'ordre. Les problèmes 1-10 sont obligatoires pour tous les étudiants ; les problèmes 11-15 sont répartis entre les groupes (chaque étudiant en résout au moins deux, selon les indications de l'enseignant). Le seuil de signification est α = 0,05, les tests sont bilatéraux. Les réponses attendues, destinées à l'enseignant, se trouvent dans le fichier de réponses du chapitre 8.

Problème L1. Le tour de taille diffère-t-il entre les femmes et les hommes ? Faites attention au résultat du test de normalité dans chaque groupe et à la taille des groupes : décidez et argumentez si vous utilisez le test t ou le test de Mann-Whitney.

Problème L2. Les triglycérides diffèrent-ils entre les patients avec et sans diabète confirmé par HGPO ? Tenez compte de la distribution des triglycérides (chapitre 3) et de la taille très inégale des groupes.

Problème L3. L'HbA1c à l'inclusion diffère-t-elle entre les trois bras ? Effectuez le calcul, puis discutez en (g) si cette question a un sens dans un essai randomisé.

Problème L4. Le cholestérol total à l'inclusion diffère-t-il entre les trois bras ? (Même discussion que pour L3 ; comparez les deux résultats.)

Problème L5. La perte de poids à 6 mois (dgr) diffère-t-elle entre les trois bras ? Si le test global est significatif, précisez quels bras diffèrent entre eux.

Problème L6. L'IMC diffère-t-il entre les non-fumeurs, les anciens fumeurs et les fumeurs ? Vérifiez la distribution de l'IMC dans chaque groupe et l'homogénéité des variances avant de choisir le test.

Problème L7. La satisfaction à 6 mois diffère-t-elle entre les femmes et les hommes ?

Problème L8. L'âge diffère-t-il entre les patients vivant en milieu urbain et ceux vivant en milieu rural ?

Problème L9. La proportion de patients ayant perdu au moins 5% de leur poids diffère-t-elle entre les femmes et les hommes ? Calculez et rapportez les effectifs attendus.

Problème L10. Le statut tabagique est-il associé au milieu de résidence ?

Problème L11. Le statut tabagique est-il distribué de la même manière dans les trois bras ? Calculez le test, puis expliquez en (g) comment vous interprétez une différence à l'inclusion dans un essai randomisé.

Problème L12. La proportion de patients ayant une glycémie à jeun ≥ 100 mg/dL a-t-elle changé entre l'inclusion et 6 mois, séparément dans le bras A et dans le bras B ? Choisissez le test approprié pour données appariées et commentez la différence entre les deux bras.

Problème L13. La glycémie à jeun à l'inclusion est-elle associée à l'HbA1c à l'inclusion ? Construisez le nuage de points et choisissez le coefficient de corrélation approprié.

Problème L14. La glycémie à jeun à 0, 3 et 6 mois a-t-elle évolué dans le bras A ? Vérifiez les conditions de l'ANOVA à mesures répétées (normalité, sphéricité) et décidez quel test vous retenez comme test principal. Comparez ensuite le résultat avec celui de l'alternative et discutez.

Problème L15. L'âge est-il associé au score FINDRISC ? Choisissez le coefficient de corrélation et discutez en (g) si l'association peut être en partie "construite" par la façon dont le score FINDRISC est calculé.

L8.4 Évaluation par les pairs (20 minutes)

Une fois les problèmes 1-10 terminés, échangez votre rapport avec votre voisin (en partageant le document Google avec droit de commentaire). Chacun évalue le rapport de l'autre pour trois problèmes choisis par l'enseignant, selon la grille du tableau L8.3. Les commentaires s'écrivent directement dans le document, en marge du texte. Le but n'est pas de "prendre en faute", mais de trouver les erreurs maintenant, à la place de l'examinateur.

Tableau L8.3. Grille d'évaluation par les pairs (pour un problème ; 0 = absent ou faux, 1 = partiel, 2 = complet et correct).

Critère	Ce que vous vérifiez	Note
1. Les hypothèses	H0 et H1 sont écrites en mots et sous forme mathématique, avec le bon paramètre (μ, Me, p)	0-2
2. Le choix du test	le test correspond au type de la variable et au schéma d'étude ; l'alternative est mentionnée	0-2
3. La vérification des conditions	normalité par groupe ou sur les différences, Levene, effectifs attendus, selon le cas, avec les valeurs	0-2
4. Le résultat	description appropriée (moyenne ± ÉT, ou médiane et EIQ, ou n et %) ; statistique, ddl, p ; taille d'effet	0-2
5. Le graphique	le type de graphique convient aux données ; les axes ont un titre et une unité	0-2

Critère	Ce que vous vérifiez	Note
6. Les conclusions	décision statistique correcte ("nous ne rejetons pas", et non "nous acceptons") ; la conclusion clinique tient compte de la taille d'effet et du schéma d'étude	0-2

En plus de la note, l'évaluateur écrit une observation positive ("ce que tu as très bien fait") et une suggestion concrète d'amélioration. Après l'évaluation, chaque étudiant dispose de 5 minutes pour corriger son rapport. Le rapport final conserve les commentaires du collègue et signale ce qui a été corrigé.

Questions qu'il vaut la peine de poser lors de l'évaluation, d'après l'expérience des séances de travaux pratiques :

- Le collègue a-t-il vérifié la normalité sur l'ensemble de l'échantillon ou dans chaque groupe ?
- Pour un problème avec données appariées (L12, L14), a-t-il utilisé un test pour données appariées ?
- A-t-il rapporté la moyenne pour la satisfaction (L7) ? Si oui, la description ne correspond pas au test.
- A-t-il écrit "il n'y a pas de différence" après un résultat non significatif, au lieu de "nous n'avons pas trouvé de différence significative" ?
- Pour L3, L4 et L11, a-t-il remarqué que tester les différences à l'inclusion dans un essai randomisé est discutable ?
- Pour L14, a-t-il choisi le test avant de voir la valeur p, ou a-t-il choisi celui qui "sort" ?

L8.5 Discussion finale et dépôt

Au cours des 10 dernières minutes, l'enseignant discute avec tout le groupe de trois problèmes pour lesquels le choix a été le plus souvent erroné et d'un problème dans lequel les tests ont donné des résultats discordants (en règle générale L1, L12 ou L14). Chaque étudiant dépose sur Moodle son rapport final (PDF), avec la fiche de décision, les 10-12 problèmes résolus et les commentaires du collègue.

Exercices supplémentaires

1. Sur le fichier BPOC disponible sur Moodle, formulez vous-mêmes trois questions de recherche : une qui demande un test pour deux groupes indépendants, une pour données appariées et une pour deux variables qualitatives. Résolvez-les au format a-g, sans regarder les solutions des autres séances.
2. Sur le fichier Date_HbA1c, résolvez le problème "La moyenne de l'HbA1c varie-t-elle significativement entre les années étudiées ?" en deux variantes : en supposant des mesures indépendantes (des patients différents chaque année) et en supposant des mesures répétées sur les mêmes patients. Expliquez pourquoi le test change.
3. Problème avec données dans l'énoncé, avec solution complète. Une étude a comparé la durée de la douleur postopératoire (jours) après deux techniques de hernioplastie : technique 1 (n = 6) : 2, 3, 3, 4, 5, 9 ; technique 2 (n = 6) : 4, 5, 6, 6, 8, 10. Choisissez et appliquez le test.

   Solution. (a) La technique 1 entraîne une durée plus courte de la douleur. (b) H0 : les distributions de la durée sont les mêmes (Me1 = Me2) ; H1 : les distributions diffèrent. (c) La variable est numérique, mais il s'agit d'une durée en jours, sur des groupes de 6, avec une valeur isolée (9 et 10 jours) : test U de Mann-Whitney ; groupes indépendants. (d) Calcul à la main. Série commune ordonnée : 2, 3, 3, 4, 4, 5, 5, 6, 6, 8, 9, 10. Les rangs : 2

→ 1 ; 3 et 3 → 2,5 ; 4 et 4 → 4,5 ; 5 et 5 → 6,5 ; 6 et 6 → 8,5 ; 8 → 10 ; 9 → 11 ; 10 → 12. R1 = 1 + 2,5 + 2,5 + 4,5 + 6,5 + 11 = 28 ; R2 = 4,5 + 6,5 + 8,5 + 8,5 + 10 + 12 = 50 ; vérification $28 + 50 = 78 = \frac{12\times13}{2}$. U1 = 28 - 21 = 7 ; U2 = 50 - 21 = 29 ; U = 7. (e) La valeur critique pour n1 = n2 = 6 (tableau 7.15) est 5 ; médianes de 3,5 et 6 jours ; $r_{\mathrm{rb}} = 1 - \frac{2\times7}{36} = 0{,}61$. (f) U = 7 > 5, nous ne rejetons pas H0 (p exact d'environ 0,09). (g) La douleur a duré moins longtemps après la technique 1 (médiane de 3,5 contre 6 jours), avec un effet estimé important, mais la différence n'est pas statistiquement significative sur 12 patients ; une étude de plus grande taille est nécessaire.

Ce qu'il faut rendre et barème indicatif

Vous déposez sur Moodle un seul fichier PDF contenant :

- la fiche de décision remplie avant l'analyse (tableau L8.1), pour l'ensemble des 15 problèmes (1 point) ;
- les problèmes L1-L10 résolus au format a-g, chacun avec son graphique (5 points, soit 0,5 point par problème) ;
- au moins deux problèmes parmi L11-L15, dans le même format (1,5 point) ;
- la grille d'évaluation remplie pour le rapport du collègue, avec l'observation positive et la suggestion d'amélioration (1,5 point) ;
- la liste des corrections faites après l'évaluation par les pairs (1 point).

Total : 10 points. Erreurs qui annulent la note d'un problème : test pour groupes indépendants appliqué à des données appariées ; khi-deux rapporté avec des effectifs attendus inférieurs à 5 sans aucune discussion ; test choisi d'après le résultat (par exemple rapporter l'alternative uniquement parce qu'elle donne $p < 0{,}05$) ; conclusion "nous acceptons H0".

Variante hors ligne

Sans accès à Moodle, le fichier de données est copié depuis le dossier partagé du laboratoire ou depuis la clé USB, et le rapport est rédigé dans un traitement de texte local et remis à la séance suivante. L'évaluation par les pairs se fait sur des feuilles imprimées : les étudiants échangent leurs rapports et remplissent à la main la grille du tableau L8.3. Si les ordinateurs ne sont pas non plus disponibles, les problèmes sont résolus sur papier à partir des statistiques descriptives fournies par l'enseignant (moyennes, écarts-types, médianes, tableaux de fréquences), et la consigne se limite aux composantes (a), (b), (c), (f) et (g) du format, plus le calcul à la main d'un test du khi-deux ou de Mann-Whitney choisi par l'enseignant.

Travaux pratiques 9. Analyse de corrélation et de régression : Pearson, Spearman et régression linéaire simple

Objectif

Cette séance transforme le chapitre 9 en étapes de travail : à partir de deux variables mesurées chez les mêmes patients, l'étudiant trace le diagramme de corrélation, choisit le coefficient approprié, le teste, estime la droite de régression, vérifie les résidus et formule la conclusion dans le format des devoirs de la discipline. Nous travaillons sur le jeu de données mg2026, avec quatre logiciels différents, pour que vous constatiez que les résultats coïncident quel que soit l'outil et pour que vous puissiez vous débrouiller avec n'importe lequel d'entre eux à l'examen pratique.

À la fin, l'étudiant doit être capable :

- de construire un diagramme de corrélation avec la droite de régression, l'équation et le R^2 affichés ;
- de calculer les coefficients de Pearson, de Spearman et de Kendall avec la valeur p et l'intervalle de confiance, dans JASP et MedCalc ;
- d'obtenir dans Google Sheets r, R^2, la pente, l'ordonnée à l'origine, l'erreur standard de l'estimation et une prédiction, à l'aide des fonctions CORREL, PEARSON, RSQ, SLOPE, INTERCEPT, FORECAST, STEYX et LINEST ;
- d'exécuter une régression linéaire simple dans JASP, MedCalc et Epi Info et de lire le tableau des coefficients ;
- de vérifier les conditions de la régression sur les graphiques des résidus ;
- de rédiger la solution d'un problème de corrélation et de régression au format a-g.

Outils

Logiciels : JASP 0.98.1 [47] ; MedCalc 24.0 [50] ; Epi Info 7.2.7 [51] ; Google Sheets, avec les paramètres régionaux pour la Roumanie (le séparateur décimal est la virgule, et le séparateur d'arguments dans les formules est ";") [43]. Fichiers : mg2026_curat.csv (160 lignes, sur Moodle, plateforme Emediqual [52]) ; pour les exercices supplémentaires, les fichiers classiques de la discipline BD_Studenti et Probleme BPOC (problèmes sur la BPCO). Durée indicative : 100 minutes.

Dans mg2026_curat.csv, les colonnes utilisées aujourd'hui sont (après l'importation dans Google Sheets) : D = varsta, H = activitate, I = inaltime, J = greutate0, K = imc0, M = tas0, Q = hdl, R = trigliceride, S = glicemie0, V = hba1c0. Les données occupent les lignes 2-161.

Déroulement

L9.1 Première étape : le diagramme de corrélation

1. Ouvrez le fichier dans JASP (File > Open > Computer > mg2026_curat.csv). Vérifiez dans la vue des données que imc0, tas0, glicemie0 et hba1c0 sont marquées comme variables continues (l'icône en forme de règle), et activitate et educatie comme ordinales. Si JASP a mal lu une variable, cliquez sur l'icône dans l'en-tête de la colonne et changez le type.
2. Dans JASP, menu Descriptives > Descriptive Statistics : placez tas0 et imc0 dans la case Variables ; dans Plots, cochez "Scatter plots" (dans certaines versions, l'option s'appelle

"Correlation plots"). Vous verrez le nuage de points avec l'IMC sur un axe et la PAS sur l'autre. Posez-vous les quatre questions du chapitre 9 : forme, sens, intensité, valeurs isolées. Réponse attendue : nuage allongé, croissant, assez dispersé, avec quelques patients ayant un IMC supérieur à 40 qui attirent le regard vers la droite.

3. Dans Google Sheets : importez le fichier (File > Import > Upload, avec la virgule comme séparateur ; vérifiez que les décimales ont été lues correctement). Sélectionnez les colonnes K et M (IMC et PAS), puis Insert > Chart. Dans Chart editor, sous Chart type, choisissez Scatter chart. Assurez-vous que l'IMC se trouve sur l'axe X (la première colonne sélectionnée). Dans l'onglet Customize > Series, cochez Trendline, sous Type choisissez Linear, sous Label choisissez "Use equation" et cochez "Show R²". Le graphique doit afficher une équation de la forme y = 1,43x + 89,5 et $R^2 = 0{,}164$.
4. Refaites le graphique pour glicemie0 (S) et hba1c0 (V). Le nuage est visiblement plus resserré ($R^2 = 0{,}545$). Comparez les deux graphiques : voilà à quoi ressemble la différence entre une corrélation faible (r = 0,405) et une corrélation modérée à forte (r = 0,738).

L9.2 Les coefficients de corrélation dans JASP

5. Menu Regression > Correlation (dans les versions récentes, le module s'appelle Regression et contient "Correlation"). Placez dans la case Variables : imc0, tas0, varsta, glicemie0, hba1c0, hdl, trigliceride.
6. Dans la section Sample Correlation Coefficient, cochez Pearson's r, Spearman's rho et Kendall's tau-b. Dans Additional Options, cochez "Report significance", "Flag significant correlations" et "Confidence intervals" (95%). Dans Plots, cochez "Scatter plots" et, pour une vue d'ensemble, "Heatmap".
7. JASP affiche une matrice. Pour les paires d'intérêt, vous devez retrouver les valeurs du tableau L9.1.

Tableau L9.1. Valeurs d'autocontrôle sur mg2026 (n = 160).

Paire	r de Pearson (IC à 95%)	p	rho de Spearman	tau-b de Kendall	Colton
IMC - PAS	0,405 (0,266-0,527)	< 0,001	0,372	0,258	faible
âge - PAS	0,256 (0,106-0,396)	0,001	0,174	0,122	faible
glycémie - HbA1c	0,738 (0,659-0,802)	< 0,001	0,728	0,557	modérée
HDL - triglycérides	0,076 (-0,080 à 0,229)	0,34	0,075	0,050	absence de corrélation
taille - poids	0,652 (0,552-0,733)	< 0,001	0,693	0,505	modérée

8. Pour la variable ordinale activitate, exécutez séparément un Spearman avec imc0 : rho = -0,012, p = 0,88. Pour educatie et satisfactie : rho = 0,125, p = 0,115. Notez que pour les variables ordinales on ne rapporte pas le coefficient de Pearson ; si vous le cochez, JASP le calcule quand même, mais son interprétation n'est pas justifiée.
9. Ce qu'il faut noter dans le cahier : pour IMC-PAS, pourquoi nous rapportons les deux coefficients (l'IMC ne suit pas une distribution normale : Shapiro-Wilk p < 0,001, asymétrie 1,42) et le fait que les deux valeurs (0,405 et 0,372) sont proches, de sorte que les valeurs extrêmes ne faussent pas beaucoup le résultat.

L9.3 La régression linéaire simple dans JASP

10. Menu Regression > Linear Regression. Placez tas0 dans Dependent Variable et imc0 dans Covariates. Méthode : Enter.
11. Dans Statistics, cochez : Estimates, Confidence intervals (95%), Model fit, Descriptives ; dans Residuals, cochez Durbin-Watson et Casewise diagnostics (avec le seuil par défaut de 3 écarts-types).
12. Dans Plots, cochez : "Residuals vs. predicted", "Residuals histogram" (avec "Standardized residuals") et "Q-Q plot standardized residuals".
13. Le tableau Model Summary doit indiquer R = 0,405, $R^2 = 0{,}164$, R^2 ajusté $= 0{,}158$ et RMSE = 12,73. Dans JASP, la RMSE (angl. root mean square error) correspond à l'erreur standard de l'estimation (SEE) du cours.
14. Le tableau ANOVA : somme des carrés de la régression 5015,2, des résidus 25 617,0, totale 30 632,2 ; F(1 ; 158) = 30,93, p < 0,001.
15. Le tableau Coefficients : ordonnée à l'origine 89,47 (ES 7,56 ; t = 11,84 ; IC 74,54-104,40), pente pour imc0 1,435 (ES 0,258 ; t = 5,56 ; p < 0,001 ; IC 0,925-1,944). Vérifiez que le t de la pente est le même que celui obtenu à la main pour r au chapitre 9 et que $F = t^2 (5{,}56^2 = 30{,}93)$.
16. Les graphiques : les résidus en fonction des valeurs prédites doivent former une bande horizontale sans courbure et sans forme d'entonnoir (comparez avec la figure 9.5 du chapitre) ; l'histogramme et le graphique Q-Q doivent montrer une distribution approximativement normale. Si vous voulez un test formel, enregistrez les résidus (option d'ajout des résidus comme nouvelle variable dans les données, lorsque la version le permet) et appliquez le test de Shapiro-Wilk depuis Descriptives : p = 0,878.
17. Refaites la régression avec varsta comme prédicteur : PAS = 108,28 + 0,406 x âge, IC de la pente 0,166-0,646, $R^2 = 0{,}066$. Puis avec glicemie0 comme prédicteur de hba1c0 : HbA1c = 3,379 + 0,0209 x glycémie, $R^2 = 0{,}545$, RMSE = 0,301.

L9.4 Google Sheets : les mêmes résultats au moyen des fonctions

18. Dans une nouvelle feuille (ou dans les colonnes libres à droite des données), écrivez les étiquettes et les formules ci-dessous. Toutes utilisent les plages des lignes 2-161. Attention à l'ordre des arguments : pour SLOPE, INTERCEPT, RSQ, STEYX, FORECAST et LINEST, on donne d'abord Y (la variable dépendante), puis X ; pour CORREL et PEARSON, l'ordre n'a pas d'importance.

Tableau L9.2. Les fonctions de Google Sheets pour la corrélation et la régression (PAS en fonction de l'IMC).

Ce que nous calculons	Formule	Résultat attendu
coefficient de Pearson	=CORREL(K2:K161;M2:M161)	0,405
le même, autre fonction	=PEARSON(K2:K161;M2:M161)	0,405
coefficient de détermination	=RSQ(M2:M161;K2:K161)	0,164
pente b	=SLOPE(M2:M161;K2:K161)	1,435
ordonnée à l'origine a	=INTERCEPT(M2:M161;K2:K161)	89,47
erreur standard de l'estimation	=STEYX(M2:M161;K2:K161)	12,73
prédiction pour un IMC de 30	=FORECAST(30;M2:M161;K2:K161)	132,5
prédiction pour un IMC de 35	=FORECAST(35;M2:M161;K2:K161)	139,7
statistique t pour r (r dans la cellule X2)	=X2*SQRT(158)/SQRT(1-X2^2)	5,56

Ce que nous calculons	Formule	Résultat attendu
valeur p bilatérale (t dans la cellule X3)	=T.DIST.2T(X3;158)	environ 0,0000001

19. La fonction LINEST renvoie toutes les statistiques de la régression dans un bloc de 5 lignes et 2 colonnes. Sélectionnez une cellule vide qui a de l'espace libre en dessous et à droite, et écrivez =LINEST(M2:M161;K2:K161;TRUE;TRUE). Le bloc obtenu se lit ainsi : ligne 1, la pente 1,435 et l'ordonnée à l'origine 89,47 ; ligne 2, leurs erreurs standard 0,258 et 7,56 ; ligne 3, R² = 0,164 et SEE = 12,73 ; ligne 4, F = 30,93 et les degrés de liberté résiduels 158 ; ligne 5, la somme des carrés de la régression 5015,2 et la somme des carrés résiduelle 25 617,0. LINEST n'affiche pas d'étiquettes ; écrivez-les donc vous-mêmes à côté.
20. L'intervalle de confiance de la pente : si la pente se trouve dans la cellule Y1 et son erreur standard dans Y2, écrivez =Y1-T.INV.2T(0,05;158)*Y2 et =Y1+T.INV.2T(0,05;158)*Y2. Vous devez obtenir 0,925 et 1,944.
21. Sheets ne dispose pas d'une fonction dédiée au coefficient de Spearman. On procède comme dans la définition : dans une nouvelle colonne, calculez le rang de chaque IMC avec =RANK.AVG(K2;K$2:K$161;1) et étirez la formule jusqu'à la ligne 161 ; faites de même pour la PAS. Ensuite, =CORREL sur les deux colonnes de rangs donne rho = 0,372. RANK.AVG attribue aux valeurs égales la moyenne des rangs, exactement comme au chapitre 9 ; la fonction RANK simple ne le fait pas et donne un résultat légèrement différent.
22. Pour activitate et l'IMC, appliquez la même procédure et vérifiez que rho = -0,012. Avec de nombreuses valeurs égales (l'activité n'a que 5 catégories), il est important d'utiliser RANK.AVG.

L9.5 MedCalc

23. Importez le fichier (File > Import, ou ouvrez directement le fichier CSV, avec la virgule comme séparateur de champs et le point comme séparateur décimal ; vérifiez dans la feuille de données que les valeurs sont numériques).
24. Statistics > Correlation > Correlation coefficient : Variable Y = tas0, Variable X = imc0. MedCalc rapporte la taille de l'échantillon, le coefficient r = 0,405, le niveau de signification P < 0,0001 et l'IC à 95% de r 0,266-0,527.
25. Statistics > Correlation > Rank correlation : les mêmes variables. Cochez à la fois Spearman's rho et Kendall's tau, si la version propose les deux options. Valeurs attendues : rho = 0,372 (IC à 95% environ 0,23-0,50) et tau = 0,258.
26. Statistics > Regression > Regression : Variable Y = tas0, Variable X = imc0, modèle "y = a + b x". Dans les options, cochez la représentation des résidus. Le résultat comprend le coefficient de détermination $R^2 = 0{,}164$, le R² ajusté 0,158, l'écart-type résiduel (Residual standard deviation) 12,73, l'équation y = 89,47 + 1,435 x, le tableau des coefficients avec l'erreur standard, l'IC à 95%, t et P, ainsi que le tableau ANOVA. MedCalc affiche aussi automatiquement le nuage de points avec la droite de régression ; sinon, utilisez Graphs > Scatter diagram, avec l'option d'affichage de la droite de régression et de la bande de confiance.
27. Statistics > Regression > Regression, avec Y = hba1c0 et X = glicemie0 : vérifiez HbA1c = 3,379 + 0,0209 x glycémie.

L9.6 Epi Info 7

28. Dans le menu principal, Analyze Data (Classic Analysis). Commande Read (Import) : choisissez le type de fichier texte/CSV et chargez mg2026_curat.csv. Vérifiez avec List ou avec Frequencies que les variables ont été lues.
29. Statistics > Linear Regression : Outcome Variable = tas0, Other Variables = imc0. Le résultat affiche le coefficient de corrélation sous la forme "r^2" = 0,16, le tableau des coefficients (pour imc0 : 1,435, erreur standard 0,258 ; pour la constante : 89,47, erreur standard 7,56) et un tableau ANOVA. Au lieu de la statistique t, Epi Info affiche pour chaque coefficient un test F, qui est égal à t² : pour imc0, F = 30,93, et pour la constante F = 140,1. Les valeurs p sont les mêmes.
30. Autre possibilité, dans Visual Dashboard : Add Analysis Gadget > Advanced Statistics > Linear Regression, avec les mêmes variables. Les résultats sont identiques.
31. Epi Info n'a pas de menu séparé pour le coefficient de Spearman dans Classic Analysis ; pour les données ordinales, utilisez JASP, MedCalc ou la procédure avec les rangs dans Sheets.

L9.7 Rédaction du résultat au format a-g

32. Rédigez la solution du problème "Existe-t-il une association entre l'IMC et la PAS à l'inclusion ?" sur le modèle ci-dessous et comparez avec votre propre version.

(a) Hypothèse de recherche : les patients ayant un IMC plus élevé ont, en moyenne, une pression systolique plus élevée.

(b) Hypothèses statistiques : H0 : ρ = 0 (l'IMC et la PAS ne sont pas corrélés linéairement dans la population) ; H1 : ρ ≠ 0.

(c) Méthode : deux variables numériques chez les mêmes patients, donc corrélation et régression linéaire simple ; X = IMC (prédicteur), Y = PAS (réponse). La PAS suit une loi normale (Shapiro-Wilk p = 0,303), l'IMC est asymétrique (p < 0,001) ; c'est pourquoi nous rapportons les coefficients de Pearson et de Spearman.

(d) Logiciel : JASP, Regression > Correlation et Regression > Linear Regression ; vérification dans Google Sheets.

(e) Résultat : r = 0,405 (IC à 95% 0,266-0,527), p < 0,001 ; rho = 0,372, p < 0,001 ; PAS = 89,47 + 1,435 x IMC (IC à 95% de la pente 0,925-1,944) ; $R^2 = 0{,}164$; SEE = 12,73 mmHg ; nuage de points avec la droite de régression (figure 9.4 du chapitre).

(f) Conclusion statistique : p < 0,001, on rejette H0 ; la corrélation est positive et faible selon la règle de Colton.

(g) Conclusion de recherche : chez les patients à risque cardiométabolique de l'étude, chaque kg/m² supplémentaire est associé à une PAS plus élevée d'environ 1,4 mmHg en moyenne. L'IMC n'explique que 16,4% de la variation de la PAS ; 83,6% relèvent d'autres facteurs. La prédiction pour un IMC de 30 est de 132,5 mmHg, pour un IMC de 35 de 139,7 mmHg ; le modèle ne doit pas être utilisé pour des IMC en dehors de l'intervalle 23,5-48,3 kg/m².

33. Reprenez la rédaction pour la paire taille-poids et répondez aux deux questions classiques de la séance : quelle proportion de la variation du poids relève de facteurs autres que la taille ($1 - R^2 = 0{,}576$, soit 57,6%) et quel poids est prédit pour 1,78 m (93,1 kg) et pour 2,35 m (la formule donne 155,0 kg, mais il s'agit d'une extrapolation en dehors du domaine 145-192 cm ; la prédiction n'est donc pas crédible).

Exercices supplémentaires

Les exercices 1-4 se résolvent sur les fichiers classiques de la discipline disponibles sur Moodle. Nous ne donnons pas les résultats numériques : vous les comparerez avec vos collègues et avec l'enseignant pendant la séance.

1. Le fichier BD_Studenti. Étudiez l'association entre le poids et la taille : (1) de quel type d'investigation statistique s'agit-il et quelles sont ses caractéristiques ; (2) les hypothèses statistiques en mots et sous forme mathématique ; (3) la représentation graphique appropriée ; (4) le traitement statistique ; (5) quelle proportion de la variation du poids est due à des facteurs autres que la variation de la taille ; (6) l'interprétation par rapport au graphique et aux hypothèses ; (7) quel poids devraient avoir un étudiant de 1,78 m et un étudiant de 2,35 m, avec un commentaire sur l'extrapolation.
2. Le fichier BD_Studenti. Calculez l'IMC = poids (kg) / taille² (m²) pour chaque étudiant, ainsi que l'IMC souhaité (à partir du poids souhaité et de la même taille). Étudiez l'association entre l'IMC actuel et l'IMC souhaité. Choisissez le coefficient après avoir vérifié la normalité des deux variables. Discutez séparément comment vous montreriez, sur les mêmes données, si les étudiants souhaitent en moyenne un IMC différent de leur IMC actuel (indice : il s'agit d'une question de comparaison appariée, chap. 6, et non de corrélation ; une corrélation élevée ne signifie pas que les valeurs sont égales).
3. Le fichier Probleme BPOC. Dans quelle mesure l'âge du patient (années) permet-il de prédire le débit expiratoire maximal initial (DEM1) ? Écrivez l'équation de la droite de régression, interprétez la pente et le R², vérifiez les résidus et précisez l'intervalle d'âge dans lequel l'équation peut être utilisée.
4. Le fichier Probleme BPOC. Existe-t-il une corrélation entre la pression artérielle systolique initiale (TMAX1) et la fréquence cardiaque initiale (FC1) ? Choisissez entre Pearson et Spearman après avoir vérifié les conditions et justifiez votre choix.
5. Problème avec données dans l'énoncé, avec solution complète. Chez cinq patients, on a mesuré la glycémie à jeun (mg/dL) et l'HbA1c (%) : (90 ; 5,3), (100 ; 5,2), (110 ; 5,8), (125 ; 5,7), (140 ; 6,4). Calculez r, testez-le, écrivez la droite de régression de l'HbA1c en fonction de la glycémie et prédisez l'HbA1c pour une glycémie de 120 mg/dL.

Solution. Les moyennes : $m_x = \frac{565}{5} = 113$ mg/dL, $m_y = \frac{28,4}{5} = 5,68\%$. Les écarts de X : -23, -13, -3, 12, 27 ; ceux de Y : -0,38 ; -0,48 ; 0,12 ; 0,02 ; 0,72. Les produits : 8,74 ; 6,24 ; -0,36 ; 0,24 ; 19,44, donc Sxy = 34,30. Sxx = 529 + 169 + 9 + 144 + 729 = 1580. Syy = 0,1444 + 0,2304 + 0,0144 + 0,0004 + 0,5184 = 0,908. $r = \frac{34,30}{\sqrt{1580 \times 0,908}} = \frac{34,30}{37,88} = 0,906$, corrélation forte selon Colton. Test : $t = \frac{0,906 \times \sqrt{3}}{\sqrt{1-0,820}} = \frac{1,569}{0,424} = 3,70$; la valeur critique pour 3 degrés de liberté est 3,182 ; comme 3,70 > 3,182, on rejette H0 (p = 0,034). Pente $b = \frac{34,30}{1580} = 0,0217$ point de pourcentage par mg/dL ; ordonnée à l'origine a = 5,68 - 0,0217 x 113 = 3,23. Équation : HbA1c = 3,23 + 0,0217 x glycémie. Pour 120 mg/dL : 3,23 + 0,0217 x 120 = 5,83%. $R^2 = 0,82$. Commentaire : avec seulement cinq patients, la corrélation n'est significative que de justesse, bien que r soit de 0,9 ; sur mg2026 (n = 160), la corrélation est plus faible (0,738), mais estimée de façon beaucoup plus précise. Vérifiez le résultat dans Sheets avec =CORREL, =SLOPE, =INTERCEPT et =FORECAST.

Ce qu'il faut rendre et barème indicatif

Vous déposez sur Moodle, dans l'activité de la séance, un seul document (PDF ou lien Google Docs) qui contient : le nuage de points IMC-PAS avec la droite de régression, l'équation et le R²

(depuis Sheets ou JASP) ; le tableau L9.1 complété avec les valeurs que vous avez obtenues ; la capture du tableau Coefficients de JASP et le graphique des résidus ; le bloc LINEST de Sheets, avec les étiquettes ; la rédaction a-g pour IMC-PAS et pour taille-poids ; la solution de l'exercice supplémentaire 1 ou 3.

Barème indicatif (10 points) : graphique correct, avec les axes appropriés (X prédicteur, Y réponse) et la droite affichée, 1 point ; coefficients de Pearson et de Spearman avec p et IC, choix justifié, 2 points ; droite de régression avec interprétation de la pente et de l'ordonnée à l'origine, 2 points ; vérification des résidus, 1 point ; R^2 et 1 - R^2 interprétés correctement, 1 point ; prédiction et commentaire sur l'extrapolation, 1 point ; rédaction a-g complète et correcte, 2 points. Des points sont retirés pour la formulation "on accepte H0", l'interprétation de la corrélation comme une causalité et l'absence d'unités de mesure pour la pente.

Variante hors ligne

Si vous n'avez pas accès à Internet ou aux logiciels, la séance peut se faire entièrement avec une calculatrice de poche et un extrait imprimé de mg2026 (les 20 premiers patients, colonnes imc0 et tas0, fourni par l'enseignant). Les étapes sont celles du problème résolu 9.1 : le tableau des écarts, des produits et des carrés, Sxy, Sxx, Syy, puis r, t (comparé à la valeur critique de la table de la loi t distribuée en séance), b, a, une prédiction et R^2. Le diagramme de corrélation se dessine sur du papier millimétré, avec la droite tracée par le point moyen (mx, my) et par un second point calculé à partir de l'équation. Le coefficient de Spearman se calcule avec les rangs, selon le tableau 9.4. Les résultats sur 20 patients différeront de ceux sur 160 : comparez-les et expliquez pourquoi l'intervalle de confiance de r est plus large.

Travaux pratiques 10. Analyses épidémiologiques : le risque relatif, le rapport de cotes et l'interprétation des études observationnelles

Objectif

Cette séance applique les méthodes du chapitre 10 aux problèmes classiques de la discipline et au jeu de données mg2026. L'accent porte sur trois compétences vérifiées à l'examen pratique : transformer un énoncé en un tableau 2 x 2 correct, choisir la mesure d'association adaptée au schéma d'étude (RR, rapport de prévalences ou OR) et interpréter l'intervalle de confiance. Les calculs se font une fois à la main, puis sont vérifiés dans Epi Info, MedCalc, JASP et Google Sheets.

À la fin, l'étudiant doit être capable :

- de reconnaître le type d'étude à partir d'un énoncé et de construire le tableau 2 x 2 avec la notation a, b, c, d ;
- de calculer le RR, l'OR, leurs intervalles de confiance à 95% et le test du khi-deux ou de Fisher, à la main et dans Epi Info StatCalc ;
- d'utiliser les calculateurs de risque relatif et de rapport de cotes de MedCalc et les tableaux de contingence de JASP ;
- de construire dans Google Sheets une feuille de calcul réutilisable pour n'importe quel tableau 2 x 2 ;
- d'estimer l'impact au niveau de la population : le risque attribuable, la fraction attribuable et le NST ou le NNH ;
- de formuler la conclusion statistique et la conclusion clinique pour une étude observationnelle, avec ses limites.

Outils

Logiciels : Epi Info 7.2.7 [51], modules StatCalc et Classic Analysis ; MedCalc 24.0 [50] ; JASP 0.98.1 [47] ; Google Sheets avec les paramètres régionaux pour la Roumanie (virgule décimale, ";" comme séparateur d'arguments) [43]. Fichiers : mg2026_curat.csv sur Moodle [52] ; pour les exercices supplémentaires, le fichier Probleme BPOC. Durée indicative : 100 minutes.

Déroulement

L10.1 De l'énoncé au tableau 2 x 2

1. Pour chaque problème, avant tout calcul, répondez par écrit à quatre questions : Quelle est l'exposition (le facteur de risque) ? Quel est le critère de jugement (la maladie) ? D'où part l'étude : de l'exposition, de la maladie ou d'une "photographie" de la population ? Quels totaux sont fixés par le schéma d'étude ?
2. Construisez le tableau avec l'exposition en lignes (les exposés en haut) et la maladie en colonnes (les malades à gauche). Remplissez d'abord les totaux donnés dans l'énoncé, puis les cellules données directement, puis les cellules obtenues par soustraction. Vérifiez à la fin que a + b + c + d = N.
3. Entraînez-vous sur les quatre énoncés classiques. Tabac-BPCO : 1200 personnes en bonne santé suivies pendant 5 ans, 300 fumeurs, 50 fumeurs et 70 non-fumeurs atteints de BPCO ; cohorte, donc a = 50, b = 250, c = 70, d = 830. Écran-acuité visuelle : 1000 personnes,

200 exposées au moins 8 heures par jour, dont 90 avec une acuité diminuée, et 150 avec une acuité diminuée parmi les 800 non exposées ; étude transversale, a = 90, b = 110, c = 150, d = 650. Contraceptifs oraux-polyarthrite rhumatoïde : 100 cas et 200 témoins, dont 40 cas et 120 témoins exposés ; cas-témoins, a = 40, b = 120, c = 60, d = 80. Poussières-capacité pulmonaire : 320 retraités, 80 exposés, 30 avec une capacité diminuée dans chaque groupe ; a = 30, b = 50, c = 30, d = 210.

4. Les erreurs à vérifier : dans le problème des poussières, "30 dans les deux groupes" ne signifie pas un risque égal (30 sur 80 contre 30 sur 240) ; dans le problème tabac-BPCO, le nombre de non-fumeurs est de 900, et non de 1200 ; dans le problème cas-témoins, les totaux fixés sont ceux des colonnes (100 cas et 200 témoins), et non ceux des lignes.

L10.2 Epi Info StatCalc : Tables (2 x 2 x n)

5. Dans le menu principal d'Epi Info, choisissez StatCalc, puis Tables (2 x 2 x n). Un tableau apparaît, avec les lignes Exposed + / Exposed - et les colonnes Outcome + / Outcome - (les intitulés peuvent légèrement différer selon la version). Pour tabac-BPCO, saisissez : 50, 250, 70, 830, dans l'ordre a, b, c, d.
6. Les résultats apparaissent immédiatement, en deux blocs. Le bloc des mesures d'association contient le rapport de cotes (Odds ratio) 2,37 avec un IC à 95% de 1,61-3,50 (méthode des séries de Taylor, équivalente à la méthode logarithmique du cours), une estimation du maximum de vraisemblance de l'OR (MLE Odds Ratio, environ 2,37, avec un intervalle exact un peu plus large), le risque relatif (Risk Ratio) 2,14 avec un IC à 95% de 1,53-3,01, la différence de risques (Risk Difference) 8,89% avec un IC à 95% d'environ 4,3-13,5% et les fractions étiologiques (chez les exposés 53,3%, dans la population 22,2%). Le bloc des tests contient le khi-deux non corrigé 19,75, le khi-deux de Mantel-Haenszel 19,74 et le khi-deux corrigé de Yates 18,78, chacun avec sa valeur p (toutes inférieures à 0,0001), ainsi que les tests exacts (mid-p et Fisher).
7. Le khi-deux de Mantel-Haenszel calculé sur un seul tableau est le khi-deux habituel multiplié par (N - 1) / N : $\frac{19{,}753 \times 1199}{1200} = 19{,}737$. La différence ne compte que pour les petits échantillons.
8. Saisissez la variante réduite (5, 25, 7, 83). Notez : RR 2,14 (IC à 95% 0,73-6,25), OR 2,37 (IC à 95% 0,69-8,13), khi-deux non corrigé 1,98 (p = 0,16), test exact de Fisher bilatéral p = 0,17. StatCalc signale qu'un effectif attendu est inférieur à 5 (l'effectif attendu pour la cellule a est de 3) et recommande le test exact. Conclusion : les mêmes mesures d'association, mais un résultat non significatif.
9. Saisissez tour à tour les autres problèmes classiques et complétez le tableau L10.1 de la section L10.8. Pour le problème ECG-CHD (8, 39, 10, 134), observez le résultat limite : khi-deux non corrigé p = 0,040, Yates p = 0,077, Fisher p = 0,0485, IC à 95% du RR 1,03-5,85.
10. L'analyse stratifiée. StatCalc permet de saisir plusieurs strates (d'où le "x n" dans le nom) : après le premier tableau, passez à la strate suivante. Saisissez l'association obésité-diabète de mg2026 par sexe : femmes 10, 19, 5, 52 ; hommes 8, 17, 6, 43. Le résultat récapitulatif affiche l'OR brut (Crude) 4,32 et l'OR ajusté de Mantel-Haenszel 4,31, avec un IC à 95% d'environ 1,86-10,0. Comme l'OR ajusté est pratiquement égal à l'OR brut, le sexe n'est pas un facteur de confusion. Les OR par strate (5,47 chez les femmes, 3,37 chez les hommes) diffèrent, mais le test d'homogénéité (Breslow-Day, lorsqu'il est affiché) ne montre pas de différence significative.

L10.3 Epi Info Classic Analysis sur les données brutes de mg2026

11. Analyze Data > Classic. Commande Read : importez mg2026_curat.csv.
12. Créez la variable d'exposition. Commande Define : obez (type numérique). Puis Assign (ou Recode) : obez = 1 si imc0 ≥ 30, sinon 0. Dans Classic Analysis, on peut écrire par exemple IF imc0 >= 30 THEN obez = 1 ELSE obez = 0 END. Vérifiez avec Frequencies obez : 54 valeurs 1 et 106 valeurs 0.
13. Statistics > Tables : Exposure Variable = obez, Outcome Variable = diabet_ogtt. Epi Info affiche le tableau 2 x 2 et les mêmes statistiques que StatCalc. Vérifiez l'ordre des lignes et des colonnes : la cellule en haut à gauche doit être "obez = 1 et diabète = 1", soit 18. Si le logiciel a placé les catégories 0 en premier, l'OR affiché sera l'inverse (1 / 4,32 = 0,23) ; dans ce cas, recodez les variables en Yes/No ou vérifiez les options d'affichage.
14. Valeurs attendues : OR 4,32 (IC à 95% 1,86-10,03), RR (ici le rapport de prévalences, puisqu'il s'agit d'une analyse transversale) 3,21 (IC à 95% 1,64-6,31), khi-deux non corrigé 12,70 (p = 0,0004), Yates 11,20 (p = 0,0008), test exact de Fisher p = 0,0008.

L10.4 MedCalc : les calculateurs de risque relatif et de rapport de cotes

15. Dans MedCalc, le menu Tests contient des calculateurs qui travaillent directement sur les effectifs, sans fichier de données. Choisissez Relative risk (dans certaines versions "Relative risk & Number needed to treat"). Saisissez pour le groupe exposé le nombre d'événements (50) et l'effectif total du groupe (300), et pour le groupe de comparaison 70 et 900.
16. MedCalc rapporte un risque relatif de 2,14, un IC à 95% de 1,53-3,01, la statistique z = 4,41 et le niveau de signification P < 0,0001. La statistique z est $\frac{\ln(\text{RR})}{\text{ES}[\ln(\text{RR})]} = \frac{0{,}762}{0{,}1727} = 4{,}41$, exactement le test construit sur l'échelle logarithmique du cours. Si l'on demande aussi le NST, MedCalc l'affiche comme nombre nécessaire pour produire un effet défavorable (NNT harm ou NNH) : 11,25, avec un IC à 95% d'environ 7,4-23,1.
17. Tests > Odds ratio : saisissez pour tabac-BPCO le nombre d'exposés et de non-exposés parmi les cas et parmi les non-cas (a = 50, c = 70 parmi les malades ; b = 250, d = 830 parmi les sujets sains, dans les champs correspondants du calculateur). Résultat : OR 2,37, IC à 95% 1,61-3,50, z = 4,35, P < 0,0001. Lisez attentivement les étiquettes des champs : le calculateur d'OR de MedCalc est conçu pour les études cas-témoins et demande les effectifs par colonne (cas, témoins), et non par ligne.
18. Recommencez pour contraceptifs-polyarthrite (cas : 40 exposées, 60 non exposées ; témoins : 120 exposées, 80 non exposées) : OR 0,44, IC à 95% 0,27-0,73, P = 0,001. N'utilisez pas le calculateur de RR pour ce problème : le schéma cas-témoins ne permet pas de calculer un RR.
19. Pour l'essai randomisé mg2026, le calculateur de risque relatif, avec 16 événements sur 54 (bras C) et 6 sur 53 (bras A), donne un RR de 2,62 (IC à 95% 1,11-6,17), P = 0,028, et un NST (NNT benefit) de 5,5, avec un IC à 95% d'environ 3,0-29,1. Comparez avec la valeur du cours. Notez que le P du calculateur MedCalc (test z sur l'échelle logarithmique) diffère légèrement du p du test du khi-deux (0,019) ; tous deux sont inférieurs à 0,05.

L10.5 JASP : tableaux de contingence avec le rapport de cotes

20. Ouvrez mg2026_curat.csv dans JASP. Créez une variable calculée : clic sur le signe "+" à côté de la dernière colonne, nom obez, type nominal, définie par la formule ifelse(imc0 >= 30, 1, 0) (éditeur R ou constructeur par glisser-déposer). Vérifiez les effectifs : 54 et 106.

21. Menu Frequencies > Contingency Tables. Rows : obez ; Columns : diabet_ogtt. Dans Statistics, cochez χ^2, "χ^2 continuity correction", "Log odds ratio (2 x 2 only)" avec l'intervalle de confiance à 95% et, s'il est disponible, le test exact de Fisher. Dans Cells, cochez Expected counts et Row percentages.
22. Lisez les résultats : effectifs attendus 9,8 ; 44,2 ; 19,2 ; 86,8 (tous supérieurs à 5, le khi-deux est donc valide) ; $\chi^2 = 12{,}70$, $p < 0{,}001$; avec correction 11,20, p = 0,0008. JASP rapporte le logarithme du rapport de cotes : ln(OR) = 1,463, avec un IC à 95% de 0,620-2,305. La transformation en OR se fait par exponentiation : $e^{1{,}463} = 4{,}32$; $e^{0{,}620} = 1{,}86$; $e^{2{,}305} = 10{,}03$.
23. Attention à l'ordre des catégories. JASP ordonne les catégories par ordre alphabétique ou numérique (0 avant 1) ; dans le tableau affiché, la cellule "obez = 0, diabète = 0" peut donc apparaître en haut à gauche. Le rapport de cotes reste le même si les lignes et les colonnes sont toutes deux inversées, mais il devient l'inverse (0,23, avec ln = -1,463) si une seule d'entre elles est inversée. Vérifiez toujours le signe du logarithme à l'aide du tableau observé. Dans JASP, vous pouvez changer l'ordre des niveaux depuis la fenêtre de la variable (clic sur l'en-tête de la colonne).
24. JASP travaille sur des données individuelles. Pour un problème donné uniquement sous forme d'effectifs (par exemple écran-acuité), créez un petit fichier de quatre lignes et trois colonnes : exposition, critère de jugement et effectif (90, 110, 150, 650), puis, dans Contingency Tables, placez l'effectif dans la case Counts.

L10.6 Google Sheets : une feuille de calcul pour n'importe quel tableau 2 x 2

25. Dans une nouvelle feuille, placez a en B2, b en C2, c en B3, d en C3 (par exemple 50, 250, 70, 830). Écrivez ensuite les formules du tableau L10.2. Après avoir vérifié les résultats sur tabac-BPCO, changez seulement les quatre cellules et la feuille recalcule tout pour n'importe quel problème.

Tableau L10.2. Les formules de Google Sheets pour l'analyse d'un tableau 2 x 2 (a en B2, b en C2, c en B3, d en C3).

Cellule	Ce qu'elle calcule	Formule	Tabac-BPCO
F2	risque chez les exposés R1	=B2/(B2+C2)	0,1667
F3	risque chez les non-exposés R0	=B3/(B3+C3)	0,0778
F4	RR	=F2/F3	2,14
F5	ES[ln(RR)]	=SQRT(1/B2-1/(B2+C2)+1/B3-1/(B3+C3))	0,1727
F6	borne inférieure de l'IC du RR	=EXP(LN(F4)-1,96*F5)	1,53
F7	borne supérieure de l'IC du RR	=EXP(LN(F4)+1,96*F5)	3,01
F8	OR	=(B2*C3)/(C2*B3)	2,37
F9	ES[ln(OR)]	=SQRT(1/B2+1/C2+1/B3+1/C3)	0,1987
F10	borne inférieure de l'IC de l'OR	=EXP(LN(F8)-1,96*F9)	1,61
F11	borne supérieure de l'IC de l'OR	=EXP(LN(F8)+1,96*F9)	3,50
F12	différence de risques (RA)	=F2-F3	0,0889

Cellule	Ce qu'elle calcule	Formule	Tabac-BPCO
F13	NST ou NNH	=ROUNDUP(1/ABS(F12);0)	12
F14	fraction attribuable chez les exposés	=(F4-1)/F4	0,533
F15	proportion d'exposés pe	=(B2+C2)/SUM(B2:C3)	0,25
F16	fraction attribuable dans la population	=F15*(F4-1)/(1+F15*(F4-1))	0,222

26. Le test du khi-deux dans Sheets nécessite le tableau des effectifs attendus. En H2:I3, écrivez les effectifs attendus : en H2 =(B2+C2)*(B2+B3)/SUM($B$2:$C$3), puis les formules analogues pour les trois autres cellules (total de ligne x total de colonne / N). Vérifiez : 30, 270, 90, 810. Ensuite, =CHISQ.TEST(B2:C3;H2:I3) renvoie directement la valeur p (environ 0,0000088 pour tabac-BPCO). La valeur de la statistique s'obtient avec =CHISQ.INV.RT(p;1) ou avec la formule Σ (O - E)² / E : 19,75. CHISQ.TEST n'applique pas la correction de Yates.
27. Pour une étude cas-témoins, la feuille calcule aussi le RR, mais le résultat n'a pas de sens. Ajoutez une cellule d'avertissement (par exemple une liste déroulante "schéma" avec les valeurs cohorte, transversale, cas-témoins et une formule =IF(K1="cas-témoins";"RR non interprétable";F4)).

L10.7 Estimation du risque au niveau de la population

28. Sur tabac-BPCO, complétez les phrases d'interprétation avec les valeurs de la feuille de calcul. Le risque attribuable : "le tabagisme ajoute environ 8,9 cas de BPCO pour 100 fumeurs en 5 ans". Le NNH : "un cas supplémentaire apparaît pour 12 fumeurs". La fraction attribuable chez les exposés : "environ 53% des cas de BPCO chez les fumeurs sont attribuables au tabagisme". La fraction attribuable dans la population : "environ 22% de tous les cas de cette population seraient évitables en l'absence de tabagisme". Toutes ces phrases supposent que la relation est causale et non confondue.
29. Sur mg2026, l'association obésité-diabète : proportion d'obèses pe = 54 / 160 = 0,3375 ; RP = 3,21 ; fraction attribuable dans la population = 0,43. La prévalence du diabète dans l'ensemble de l'échantillon (18,1%) moins la prévalence chez les non-obèses (10,4%) donne le risque attribuable dans la population, 7,7 points de pourcentage ; divisé par 18,1%, il donne de nouveau 0,43. Formulez la phrase au conditionnel ("si..."), car les données sont transversales.
30. Sur l'essai randomisé mg2026 (bras C contre A, perte de poids ≥ 5%) : différence absolue de 18,3 points de pourcentage (IC à 95% 3,4-33,2), NST 5,5, arrondi à 6 (IC à 95% 3,0-29,1). Interprétation : "en traitant 6 patients avec le programme complet au lieu du conseil standard, nous obtenons en moyenne un patient de plus avec une perte de poids cliniquement pertinente". Ici, l'interprétation causale est justifiée par la randomisation.

L10.8 Les problèmes classiques : tableau d'autocontrôle

31. Remplissez vous-mêmes les colonnes, puis comparez avec le tableau L10.1. Tous les intervalles sont calculés par la méthode logarithmique.

Tableau L10.1. Résultats d'autocontrôle pour les problèmes classiques et pour mg2026.

Problème (a, b, c, d)	Schéma d'étude	Mesure principale (IC à 95%)	OR (IC à 95%)	Test	Conclusion
Tabac-BPCO (50, 250, 70,	cohorte	RR 2,14 (1,53-3,01)	2,37 (1,61-3,50)	χ^2 19,75, p < 0,001	facteur de risque

Problème (a, b, c, d)	Schéma d'étude	Mesure principale (IC à 95%)	OR (IC à 95%)	Test	Conclusion
830)					
Tabac-BPCO réduit (5, 25, 7, 83)	cohorte	RR 2,14 (0,73-6,25)	2,37 (0,69-8,13)	Fisher p = 0,17	non significatif
Écran-acuité (90, 110, 150, 650)	transversale	RP 2,40 (1,94-2,96)	3,55 (2,55-4,93)	χ^2 60,44, p < 0,001	facteur de risque
Contraceptifs-PR (40, 120, 60, 80)	cas-témoins	OR uniquement	0,44 (0,27-0,73)	χ^2 10,71, p = 0,001	facteur protecteur
ECG-décès CHD (8, 39, 10, 134)	cohorte	RR 2,45 (1,03-5,85)	2,75 (1,02-7,44)	Fisher p = 0,0485	à la limite
Poussières (30, 50, 30, 210)	transversale	RP 3,00 (1,94-4,65)	4,20 (2,32-7,60)	χ^2 24,62, p < 0,001	facteur de risque
Âge, non-exposés (30, 210, 12, 188)	transversale	RP 2,08 (1,10-3,96)	2,24 (1,11-4,50)	χ^2 5,34, p = 0,021	facteur de risque
Obésité-diabète mg2026 (18, 36, 11, 95)	transversale	RP 3,21 (1,64-6,31)	4,32 (1,86-10,03)	χ^2 12,70, p < 0,001	facteur de risque
Bras C contre A mg2026 (16, 38, 6, 47)	essai randomisé	RR 2,62 (1,11-6,17)	3,30	χ^2 5,49, p = 0,019	intervention efficace

32. Pour le problème écran-acuité, répondez aussi à la question classique "quelle est la cote (odds) d'une acuité diminuée chez ceux qui travaillent moins de 8 heures par jour" : 150 / 650 = 0,231. Pour le problème des contraceptifs, expliquez en deux phrases pourquoi on ne calcule pas de RR.

L10.9 Interprétation d'une étude observationnelle

33. Choisissez l'un des articles de collègues discutés au chapitre 10 (par exemple la cohorte de 110 patientes avec RR = 2,75, IC à 95% 1,34-5,64 [114] ou l'étude cas-témoins appariée 1:2 portant sur 435 patients [116]) et remplissez une fiche de lecture comportant les rubriques suivantes : le schéma d'étude et l'argument du texte qui permet de l'identifier ; l'exposition et le critère de jugement ; la mesure d'association et son adéquation au schéma d'étude ; l'intervalle de confiance et son interprétation ; les facteurs de confusion pris en compte (si le résumé les mentionne) ; les critères de Bradford Hill qui sont étayés ; ce qu'il faudrait encore rapporter selon STROBE [25].

34. Discutez en groupe : un OR ajusté de 7,20 (IC à 95% 4,26-12,17), obtenu dans une étude transversale où le critère de jugement a une prévalence de 24,1% [4], peut-il se traduire par "un risque 7 fois plus élevé" ? La réponse se trouve dans le tableau 10.4 du chapitre.

Exercices supplémentaires

1. Le fichier Probleme BPOC (sans résultats numériques fournis). Les patients âgés de plus de 50 ans ont-ils un risque plus élevé de présenter une dyspnée au début de l'étude (DISPN = 1) que ceux de moins de 50 ans ? Créez la variable de classe d'âge, construisez le tableau 2 x 2, précisez le schéma d'étude (les données à l'inclusion sont transversales), calculez le RP et l'OR avec leur IC à 95% dans Epi Info et dans Sheets, appliquez le test approprié et formulez les conclusions au format a-g.
2. Le fichier Probleme BPOC (sans résultats numériques fournis). La présence de râles est-elle associée à la classe d'âge ? Calculez l'OR et vérifiez si l'intervalle contient 1. Comparez le résultat du test du khi-deux avec celui du test de Fisher et expliquez les éventuelles différences.

3. Problème avec données dans l'énoncé, avec solution complète. Au cours d'une saison grippale, un cabinet de médecine générale suit 1000 patients atteints de maladies chroniques : 400 vaccinés contre la grippe, dont 20 font une grippe confirmée, et 600 non vaccinés, dont 72 font une grippe. Analysez les données.

Solution. Schéma d'étude : cohorte (observationnelle, les patients ont choisi de se faire vacciner ou non). Le tableau : a = 20 (vaccinés malades), b = 380, c = 72 (non vaccinés malades), d = 528. Les risques : $R_1 = \frac{20}{400} = 5{,}0\%$; $R_0 = \frac{72}{600} = 12{,}0\%$. $\mathrm{RR} = \frac{5{,}0}{12{,}0} = 0{,}42$; $\mathrm{ES}[\ln(\mathrm{RR})] = \sqrt{\frac{1}{20} - \frac{1}{400} + \frac{1}{72} - \frac{1}{600}} = \sqrt{0{,}0597} = 0{,}244$; $\mathrm{IC}\,95\% = \exp(-0{,}875 \pm 0{,}479) = 0{,}26 - 0{,}67$. $\mathrm{OR} = \frac{20\times528}{380\times72} = \frac{10\,560}{27\,360} = 0{,}39 (; \mathrm{IC}\,95\%\, 0{,}23 - 0{,}64)$. χ^2 = 14,08, p < 0,001. La vaccination apparaît comme un facteur protecteur : le risque de grippe est inférieur de 58% chez les vaccinés (efficacité vaccinale estimée = 1 - RR = 58%). La différence absolue de risque est de 7,0 points de pourcentage (IC à 95% 3,6-10,4), et NNT $= \frac{1}{0{,}07} = 14{,}3$, arrondi à 15 (IC à 95% environ 10-28) : il faut vacciner 15 patients pour prévenir un cas de grippe au cours d'une saison. Principale limite : l'étude est observationnelle, et les patients qui choisissent de se faire vacciner peuvent avoir d'autres comportements protecteurs (l'effet du "vacciné en bonne santé"), de sorte que l'estimation peut être optimiste.

Ce qu'il faut rendre et barème indicatif

Vous déposez sur Moodle un document qui contient : le tableau L10.1 complété avec vos propres valeurs (d'Epi Info ou de Sheets) pour au moins cinq problèmes ; la capture d'écran de StatCalc pour tabac-BPCO et pour la variante réduite ; la feuille Google Sheets de L10.6 (lien avec droit de lecture) ; le résultat JASP pour obésité-diabète, avec la transformation du log OR en OR ; les phrases d'interprétation de L10.7 ; la fiche de lecture de L10.9 pour un article ; la solution de l'exercice supplémentaire 1.

Barème indicatif (10 points) : tableaux 2 x 2 correctement construits et schéma d'étude identifié, 2 points ; RR et OR calculés et choisis correctement en fonction du schéma d'étude, 2 points ; intervalles de confiance et test approprié (khi-deux ou Fisher), 2 points ; mesures d'impact (RA, NST/NNH, fraction attribuable) avec interprétation, 1 point ; feuille de calcul fonctionnelle, 1 point ; fiche de lecture critique, 1 point ; conclusion clinique formulée avec prudence, avec les limites du schéma d'étude, 1 point. Des points sont retirés pour le calcul d'un RR dans une étude cas-témoins, l'utilisation du total général comme dénominateur du risque chez les non-exposés et la conclusion "le facteur n'a pas d'effet" tirée d'un résultat non significatif.

Variante hors ligne

Tous les problèmes classiques peuvent être résolus avec une calculatrice de poche, si elle dispose des fonctions ln et e^x. Les étapes sont : le tableau 2 x 2 et la vérification des totaux ; R1, R0, RR et OR ; l'ES sur l'échelle logarithmique (formules du chapitre 10, section 10.7) ; l'intervalle sur l'échelle logarithmique et l'exponentiation des bornes ; les effectifs attendus et $\chi^2 = \frac{\sum(O-E)^2}{E}$, comparé à la valeur critique 3,84 pour 1 degré de liberté et α = 0,05 ; le RA, le NST ou le NNH et les fractions attribuables. Lorsqu'un effectif attendu est inférieur à 5, le test exact de Fisher ne peut pratiquement pas être fait à la main ; notez que ce serait le test approprié et fondez la conclusion sur l'intervalle de confiance, en précisant qu'il est approximatif. Comparez à la séance suivante les résultats obtenus à la main avec ceux d'Epi Info.

Travaux pratiques 11. Analyse bayésienne : le raisonnement probabiliste dans le diagnostic et le dépistage

Objectif

Cette séance transforme la théorie du chapitre 11 en outils que vous pourrez utiliser aussi après l'examen : un calculateur bayésien réutilisable dans Google Sheets, un arbre de fréquences naturelles, un tableau de diagnostic différentiel et les premières analyses bayésiennes dans JASP. Tous les calculs se font d'abord à la main, sur papier, puis dans le logiciel, afin que vous puissiez vérifier vos résultats.

À la fin, l'étudiant doit être capable :

- de calculer la probabilité post-test à partir de la probabilité pré-test, de la sensibilité et de la spécificité, aussi bien par le théorème de Bayes que par la forme en cotes ;
- de construire dans Google Sheets un calculateur bayésien et un arbre de fréquences naturelles qui se mettent à jour automatiquement lorsque la prévalence change ;
- d'utiliser le nomogramme de Fagan pour une estimation graphique rapide ;
- de résoudre un tableau de diagnostic différentiel avec trois maladies et trois signes cliniques ;
- d'obtenir à partir du jeu de données mg2026 le tableau 2 x 2 du score FINDRISC et de calculer les probabilités post-test pour différents contextes de dépistage ;
- d'exécuter dans JASP un test t bayésien et un tableau de contingence bayésien et d'interpréter le facteur BF10.

Outils

Google Sheets (paramètres régionaux pour la Roumanie : virgule décimale et séparateur d'arguments ";"), JASP 0.98.1 [47], en option MedCalc 24.0 [50]. Le fichier mg2026_curat.csv sur Moodle (160 lignes). Le nomogramme de Fagan imprimé (figure 11.3 du chapitre) et une règle. Durée indicative : 100 minutes.

Dans le fichier mg2026_curat.csv, importé dans Google Sheets avec la ligne 1 comme en-tête, les colonnes utilisées dans cette séance sont : C (sex), K (imc0), M (tas0), S (glicemie0), W (findrisc) et X (diabet_ogtt, 1 = diabète confirmé par HGPO, 0 = pas de diabète). Les données occupent les lignes 2-161.

Déroulement

L11.1 Le calculateur bayésien dans Google Sheets (20 minutes)

Ouvrez une nouvelle feuille et nommez-la "Bayes". Construisez le calculateur dans les colonnes A (étiquettes) et B (valeurs), exactement comme dans le tableau ci-dessous. Les cellules B1-B3 sont des données d'entrée ; le reste, ce sont des formules.

Tableau L11.1. Structure du calculateur bayésien.

Cellule	Étiquette (colonne A)	Contenu de la colonne B
B1	Probabilité pré-test	0,181
B2	Sensibilité	0,759
B3	Spécificité	0,748
B4	RV+	=B2/(1-B3)

Cellule	Étiquette (colonne A)	Contenu de la colonne B
B5	RV-	=(1-B2)/B3
B6	Cote pré-test	=B1/(1-B1)
B7	Cote post-test (test positif)	=B6*B4
B8	Probabilité post-test (test positif)	=B7/(1+B7)
B9	Cote post-test (test négatif)	=B6*B5
B10	Probabilité post-test (test négatif)	=B9/(1+B9)
B11	VPP par la formule de Bayes (vérification)	=B2*B1/(B2*B1+(1-B3)*(1-B1))
B12	VPN par la formule de Bayes (vérification)	=B3*(1-B1)/(B3*(1-B1)+(1-B2)*B1)

Mettez B8, B10, B11 et B12 au format pourcentage avec une décimale (Format > Number > Percent). Avec les valeurs pour FINDRISC ≥ 12 du jeu de données mg2026, vous devez obtenir : RV+ 3,01 ; RV- 0,32 ; cote pré-test 0,221 ; probabilité post-test de 40,0% après un score positif et de 6,7% après un score négatif. B11 doit être égal à B8 (40,0%), et B12 à 1 - B10 (93,3%). Si les deux voies ne donnent pas le même résultat, une formule contient une erreur.

Utilisez maintenant le calculateur comme un outil. Changez seulement B1 et complétez le tableau L11.2 (les résultats corrects sont ceux du tableau 11.7 du chapitre).

Tableau L11.2. Probabilités post-test pour FINDRISC ≥ 12, à compléter.

Probabilité pré-test (B1)	Post-test positif (B8)	Post-test négatif (B10)
0,02	5,8%	0,7%
0,05		
0,10		
0,30		
0,50		

Pour une tabulation automatique, écrivez "pré-test" en D1, "post +" en E1, "post -" en F1. En D2:D7, saisissez 0,02 ; 0,05 ; 0,10 ; 0,181 ; 0,30 ; 0,50. En E2, écrivez =D2/(1-D2)*$B$4/(1+D2/(1-D2)*$B$4), et en F2 =D2/(1-D2)*$B$5/(1+D2/(1-D2)*$B$5), puis étirez les formules vers le bas jusqu'à la ligne 7. Le signe $ fixe la référence au RV, pour qu'elle ne se déplace pas lors de la copie. Sélectionnez D1:F7 et insérez un graphique en courbes (Insert > Chart, Line chart) : vous avez reconstruit une partie de la figure 11.2.

Vérification à la main (obligatoire, dans le cahier) : pour une probabilité pré-test de 0,10, la cote est de 0,111 ; 0,111 x 3,01 = 0,334 ; 0,334/1,334 = 25,1%.

L11.2 L'arbre des fréquences naturelles (15 minutes)

Dans la même feuille, à partir de la cellule A15, construisez un arbre de fréquences naturelles pour une population de 1000 personnes, qui utilise B1-B3.

Tableau L11.3. L'arbre des fréquences naturelles dans Google Sheets.

Cellule	Étiquette	Formule
B15	Population	1000
B16	Malades	=B15*B1
B17	Non-malades	=B15-B16
B18	Malades avec test positif (VP)	=B16*B2
B19	Malades avec test négatif (FN)	=B16-B18
B20	Non-malades avec test positif (FP)	=B17*(1-B3)
B21	Non-malades avec test négatif (VN)	=B17-B20
B22	Total des tests positifs	=B18+B20
B23	VPP à partir de l'arbre	=B18/B22
B24	VPN à partir de l'arbre	=B21/(B19+B21)

Étape 1. Fixez B1 = 0,10, B2 = 0,80, B3 = 0,90. Vous devez obtenir 100 malades, 80 VP, 20 FN, 900 non-malades, 90 FP, 810 VN, 170 tests positifs, une VPP de 47,1% et une VPN de 97,6%, c'est-à-dire exactement la figure 11.1 du chapitre.

Étape 2. Gardez Se et Sp et changez la prévalence à 0,01, puis à 0,50. Notez la VPP et la VPN. À 1%, vous devriez obtenir 8 VP et 99 FP (VPP 7,5%) ; à 50%, 400 VP et 50 FP (VPP 88,9%).

Étape 3. Écrivez dans le cahier, en deux phrases, comment vous expliqueriez à un patient le résultat positif dans chacune des trois situations, en utilisant des nombres entiers ("sur 170 personnes ayant un test positif, 80 ont la maladie"). L'exercice n'a pas une seule bonne réponse, mais il sera discuté à la fin.

L11.3 Le nomogramme de Fagan (10 minutes)

Utilisez le nomogramme imprimé (figure 11.3 du chapitre) et une règle. Pour chaque situation, tracez une ligne depuis le point pré-test, en passant par la valeur du RV, lisez la probabilité post-test, puis vérifiez-la avec le calculateur de L11.1 (saisissez directement la valeur du RV en B4 ou B5, à la place de la formule, ou faites le calcul à part).

Tableau L11.4. Exercices avec le nomogramme de Fagan.

Situation	Pré-test	RV	Post-test (nomogramme)	Post-test (calcul)
FINDRISC ≥ 12, échantillon mg2026	18,1%	3,01	environ 40%	40,0%
FINDRISC < 12, échantillon mg2026	18,1%	0,32		6,7%
Test de confirmation très spécifique	10%	10		52,6%
Test d'exclusion, forte suspicion	50%	0,1		9,1%
Élastographie au-dessus du seuil de confirmation de la cirrhose [3]	28,1%	19,7		88,5%

Les différences entre la lecture et le calcul devraient être de quelques pour cent. Observez sur le nomogramme pourquoi un même RV déplace beaucoup la probabilité dans la zone médiane et peu aux extrêmes.

L11.4 Le tableau de diagnostic différentiel à trois maladies (20 minutes)

Scénario (données fictives, plausibles). Une patiente de 34 ans présente à l'hémogramme une anémie microcytaire légère. Le médecin généraliste envisage trois diagnostics : l'anémie ferriprive (probabilité a priori 0,60), la thalassémie mineure (0,15) et l'anémie des maladies chroniques (0,25). Trois résultats sont disponibles : une ferritine inférieure à 30 ng/mL (F), un indice de distribution des globules rouges (RDW) augmenté (R) et un indice de Mentzer (VGM/nombre d'hématies) inférieur à 13 (M). La fréquence de chaque résultat dans chaque maladie est donnée ci-dessous.

Tableau L11.5. Les probabilités a priori et les fréquences des résultats (données fictives).

Diagnostic	p(D)	$p(F/D)$	$p(R/D)$	$p(M/D)$
Anémie ferriprive	0,60	0,90	0,80	0,15
Thalassémie mineure	0,15	0,05	0,20	0,80
Anémie des maladies chroniques	0,25	0,10	0,40	0,20

La patiente a une ferritine normale (F-), un RDW normal (R-) et un indice de Mentzer inférieur à 13 (M+).

Étape 1. Dans le cahier, calculez la vraisemblance de la combinaison observée dans chaque maladie (pour un résultat absent, utilisez 1 - p) : ferriprive 0,10 x 0,20 x 0,15 = 0,003 ; thalassémie 0,95 x 0,80 x 0,80 = 0,608 ; maladies chroniques 0,90 x 0,60 x 0,20 = 0,108.

Étape 2. Dans Google Sheets, dans une nouvelle feuille "Differentiel", saisissez le tableau L11.5 en A1:E4 (en-tête sur la ligne 1, maladies sur les lignes 2-4). En G1, H1, I1, écrivez les résultats observés codés 1 (présent) ou 0 (absent) : G1 = 0 (F), H1 = 0 (R), I1 = 1 (M).

Étape 3. En F2, écrivez la formule de la vraisemblance combinée :

=IF($G$1=1;C2;1-C2)*IF($H$1=1;D2;1-D2)*IF($I$1=1;E2;1-E2)

En G2, le produit par la probabilité a priori : =B2*F2. En H2, la probabilité a posteriori : =G2/SUM($G$2:$G$4). Recopiez F2:H2 vers le bas jusqu'à la ligne 4.

Résultats de vérification : produits 0,0018 ; 0,0912 ; 0,0270 ; somme 0,1200 ; probabilités a posteriori : anémie ferriprive 1,5%, thalassémie mineure 76,0%, anémie des maladies chroniques 22,5%.

Étape 4. Modifiez le codage en G1:I1 pour une patiente ayant une ferritine basse, un RDW augmenté et un indice de Mentzer supérieur à 13 (1 ; 1 ; 0). Quel diagnostic devient le plus probable ? Notez le résultat. (Vous devriez obtenir environ 97,8% pour l'anémie ferriprive.)

Étape 5. Discussion : que se passerait-il si la patiente avait en réalité une anémie sidéroblastique, qui ne figure pas dans le tableau ? Quelle hypothèse du modèle serait violée si la ferritine et le RDW étaient corrélés au sein d'une même maladie ?

Le tableau obtenu est un modèle "Bayes naïf" fonctionnel. Vous pouvez le réutiliser pour n'importe quel problème de diagnostic différentiel, en changeant le nom des maladies, les probabilités a priori et les fréquences des signes ; le problème résolu 11.5 (la douleur thoracique) et l'exercice 7 du chapitre 11 (la céphalée) peuvent être vérifiés immédiatement.

L11.5 Dépistage par FINDRISC sur le jeu de données mg2026 (15 minutes)

Importez mg2026_curat.csv dans une nouvelle feuille "Donnees" (File > Import > Upload, avec "Replace current sheet" ou "Insert new sheet"). Vérifiez que W contient le score FINDRISC et X le diagnostic par HGPO.

Étape 1. Obtenez le tableau 2 x 2 pour le seuil ≥ 12, avec la fonction COUNTIFS, dans une feuille "Screening" :

- VP : =COUNTIFS(Donnees!X2:X161;1;Donnees!W2:W161;">=12") doit donner 22
- FN : =COUNTIFS(Donnees!X2:X161;1;Donnees!W2:W161;"<12") doit donner 7
- FP : =COUNTIFS(Donnees!X2:X161;0;Donnees!W2:W161;">=12") doit donner 33
- VN : =COUNTIFS(Donnees!X2:X161;0;Donnees!W2:W161;"<12") doit donner 98

Étape 2. Calculez la prévalence (29/160 = 18,1%), la Se (75,9%), la Sp (74,8%), la VPP (40,0%) et la VPN (93,3%), puis copiez la Se et la Sp dans le calculateur de L11.1.

Étape 3. Rendez le seuil variable : écrivez le seuil dans une cellule (par exemple B1 = 12) et remplacez ">=12" par ">="&B1 et "<12" par "<"&B1. Essayez les seuils 10, 14 et 16. Vous devez retrouver les valeurs du tableau 12.5 du chapitre 12 (par exemple pour ≥ 14 : VP 17, FP 16, FN 12, VN 115).

Étape 4. Pour le seuil ≥ 12, calculez la probabilité post-test pour trois patients fictifs : une femme de 30 ans sans facteur de risque (pré-test 2%), un homme de 55 ans en surpoids (10%) et une patiente obèse avec un prédiabète connu (30%). Formulez la recommandation pour chacun : réassurance, glycémie à jeun ou HGPO.

Étape 5 (tests séquentiels). Pour un patient ayant un FINDRISC ≥ 12 et une glycémie à jeun ≥ 110 mg/dL, multipliez successivement les rapports de vraisemblance (FINDRISC RV+ 3,01 ; glycémie RV+ 2,78). Vous obtenez une probabilité post-test d'environ 64,9%. Comptez ensuite directement dans les données : =COUNTIFS(Donnees!X2:X161;1;Donnees!W2:W161;">=12";Donnees!S2:S161;">=110") donne 20, et la même formule avec X = 0 donne 19. La probabilité observée est de 20/39 = 51,3%. Expliquez la différence en deux phrases (indication : section 11.5.2).

L11.6 Analyses bayésiennes dans JASP (20 minutes)

Ouvrez mg2026_curat.csv dans JASP. Vérifiez le type des variables : sex nominale, tas0 et imc0 continues (scale), diabet_ogtt nominale.

Étape 1. Le test t bayésien pour échantillons indépendants. Dans le menu T-Tests, section Bayesian, choisissez Independent Samples T-Test. Placez tas0 dans Variables et sex dans Grouping Variable. Laissez l'hypothèse alternative bilatérale et la loi a priori par défaut (Cauchy, largeur 0,707). Cochez Prior and posterior et Bayes factor robustness check dans Plots. Résultat attendu : BF10 d'environ 0,18 (si vous choisissez d'afficher le BF01, environ 5,7). Interprétation : des preuves modérées en faveur de l'absence de différence de PAS entre les sexes. Comparez avec le test t fréquentiste de la séance de travaux pratiques consacrée aux tests t (p = 0,81) : la valeur p ne permettait pas d'affirmer l'absence de différence, le facteur de Bayes la soutient modérément.

Étape 2. Le graphique de robustesse montre comment le BF10 change lorsque la largeur de la loi a priori varie. Si la conclusion (preuves en faveur de H0) reste la même pour toutes les largeurs raisonnables, le résultat est robuste. Notez comment le BF10 se modifie pour les lois a priori "wide" et "ultrawide".

Étape 3. Tableau de contingence bayésien. Créez d'abord la variable obez : dans JASP, utilisez le bouton "+" à côté des noms de colonnes pour ajouter une colonne calculée (Computed column), avec la formule imc0 >= 30, de type nominal. (Autre possibilité : créez la colonne dans Google Sheets avec =IF(K2>=30;1;0) et réimportez le fichier.) Ensuite, dans le menu Frequencies, section Bayesian, choisissez Contingency Tables. Placez obez dans Rows et diabet_ogtt dans Columns. Sous Sampling, choisissez Joint multinomial, avec la concentration a priori par défaut (1). Résultat attendu : BF10 d'environ 85, c'est-à-dire des preuves très fortes en faveur d'une association entre l'obésité et le diabète. Si vous choisissez Independent multinomial, rows fixed, le BF10 descend à environ 64 ; la conclusion reste la même. Notez que le type d'échantillonnage est un choix lié au schéma d'étude, qui doit être justifié et rapporté.

Étape 4 (en option). Dans le menu Frequencies, section Bayesian, choisissez Binomial Test pour la variable diabet_ogtt, avec la loi a priori par défaut Beta(1; 1). Cochez le graphique Prior and posterior. L'intervalle de crédibilité central à 95% pour la proportion de diabète doit être d'environ 12,9-24,8%, comme dans la figure 11.4 du chapitre. Remplacez la loi a priori par Beta(10; 40) et observez à quel point l'intervalle change peu (environ 13,6-24,1%).

Étape 5. Rapportez les résultats dans le format des devoirs de la discipline : (a) l'hypothèse de recherche ; (b) les hypothèses statistiques ; (c) la méthode et sa justification (pourquoi un test bayésien) ; (d) le logiciel et les réglages (la loi a priori, le type d'échantillonnage) ; (e) le résultat avec le graphique ; (f) la conclusion statistique, en utilisant l'échelle du tableau 11.8 du chapitre ; (g) la conclusion de recherche en mots.

Exercices supplémentaires

1. Sur le fichier BPOC disponible sur Moodle, considérez la dyspnée (DISPN) comme la "maladie" et les râles (RAL) comme le "symptôme". Construisez le tableau maladie-symptôme avec COUNTIFS et calculez p(B+), p(S+ / B+), p(S+ / B-) et p(B+ / S+).

Vérifiez p(B+ / S+) avec le théorème de Bayes. Commentez si les râles seraient un "test" utile pour la dyspnée dans cet échantillon.

2. Sur le fichier Date_HbA1c disponible sur Moodle, choisissez un seuil clinique pour l'HbA1c (par exemple 6,5%) et une variable pouvant servir de référence, s'il en existe une dans le fichier. Construisez le tableau 2 x 2, calculez Se, Sp, RV+ et RV- et utilisez le calculateur de L11.1 pour trois probabilités pré-test différentes. Discutez les limites de la référence choisie.
3. Problème avec données (solution complète). Un test rapide de détection de l'antigène streptococcique a une Se de 86% et une Sp de 95%. Chez un enfant atteint de pharyngite, le médecin estime la probabilité pré-test à 40%. Calculez la probabilité post-test après un test positif et après un test négatif.

Solution. $RV^+ = \frac{0,86}{0,05} = 17,2$; $RV^- = \frac{0,14}{0,95} = 0,147$. Cote pré-test = 0,40/0,60 = 0,667. Test positif : 0,667 x 17,2 = 11,47, probabilité 11,47/12,47 = 92,0%. Test négatif : 0,667 x 0,147 = 0,098, probabilité 0,098/1,098 = 8,9%. Avec un seuil de traitement de 25% (problème résolu 11.6), le test positif justifie le traitement, et le test négatif fait descendre la probabilité en dessous du seuil.

4. Problème avec données (solution complète). En utilisant le tableau de diagnostic différentiel de L11.4, une patiente a une ferritine basse (F+), un RDW normal (R-) et un indice de Mentzer supérieur à 13 (M-). Calculez les probabilités a posteriori.

Solution. Vraisemblances : ferriprive 0,90 x 0,20 x 0,85 = 0,153 ; thalassémie 0,05 x 0,80 x 0,20 = 0,008 ; maladies chroniques 0,10 x 0,60 x 0,80 = 0,048. Produits par les probabilités a priori : 0,60 x 0,153 = 0,0918 ; 0,15 x 0,008 = 0,0012 ; 0,25 x 0,048 = 0,0120. Somme 0,1050. Probabilités a posteriori : ferriprive 87,4%, thalassémie 1,1%, maladies chroniques 11,4%.

Ce qu'il faut rendre et barème indicatif

Vous déposez sur Moodle un seul fichier Google Sheets (lien avec droit de lecture) ou exporté au format .xlsx, avec les feuilles "Bayes", "Differentiel" et "Screening", plus un court document (une page au maximum) avec les résultats de JASP (captures des tableaux avec le BF10) et l'interprétation au format a-g.

Barème indicatif (10 points) : calculateur bayésien correct, avec la vérification croisée B8 = B11 (2 points) ; arbre de fréquences naturelles fonctionnel pour trois prévalences (1,5 point) ; exercices avec le nomogramme, avec la lecture et le calcul (1 point) ; tableau de diagnostic différentiel avec les formules correctes et les deux scénarios (2 points) ; tableau 2 x 2 du FINDRISC obtenu avec COUNTIFS et un seuil variable, plus l'explication des tests séquentiels (1,5 point) ; analyses JASP correctement rapportées, avec l'interprétation du BF10 sur l'échelle du cours (2 points).

Variante hors ligne

Si vous n'avez pas accès à Google Sheets ou à JASP, tous les calculs de L11.1-L11.5 peuvent se faire avec une calculatrice de poche et sur papier, en utilisant les tableaux de cette séance comme modèles : transformez la probabilité en cote, multipliez par le RV, revenez à la probabilité. L'arbre de fréquences naturelles se dessine à la main pour 1000 personnes, et le nomogramme de Fagan imprimé remplace le calculateur. Pour L11.5, le tableau 2 x 2 du FINDRISC est donné au chapitre 12 (tableau 12.3), et pour L11.6, utilisez les résultats de référence donnés plus haut (BF10 0,18 pour la PAS et le sexe ; environ 85 pour l'obésité et le diabète) et écrivez seulement l'interprétation. La même logique peut s'appliquer dans LibreOffice Calc ou Microsoft Excel, avec les mêmes formules (dans la configuration avec point décimal, le séparateur d'arguments devient ",").

Travaux pratiques 12. Performance des tests diagnostiques : sensibilité, spécificité, valeurs prédictives et courbe ROC

Objectif

Dans cette séance, vous évaluez un test diagnostique de bout en bout : vous construisez le tableau 2 x 2, calculez tous les indicateurs (d'abord à la main, puis dans Google Sheets et MedCalc), choisissez un seuil, construisez la courbe ROC et comparez deux tests. Le test principal est le score FINDRISC du jeu de données mg2026, avec l'HGPO comme standard de référence ; le test de laboratoire analysé à la fin est la glycémie à jeun.

À la fin, l'étudiant doit être capable :

- de calculer à partir d'un tableau 2 x 2 la prévalence, la Se, la Sp, la VPP, la VPN, l'exactitude, le taux d'erreur (b + c)/N, l'indice de Youden, le RV+, le RV- et le DOR, avec leurs intervalles de confiance ;
- d'utiliser la procédure d'évaluation d'un test diagnostique de MedCalc et de recalculer les valeurs prédictives pour une autre prévalence ;
- de construire dans Google Sheets un tableau de seuils et la courbe ROC correspondante et d'estimer l'AUC par la méthode des trapèzes ;
- d'exécuter dans MedCalc l'analyse ROC (AUC avec IC à 95%, tableau des critères, indice de Youden) et la comparaison de deux courbes ROC ;
- d'obtenir les mêmes résultats dans JASP, lorsque la version le permet, et de lire un tableau 2 x 2 dans Epi Info ;
- de rédiger l'évaluation de la performance d'un test de laboratoire au format a-g des devoirs de la discipline.

Outils

Google Sheets (paramètres régionaux pour la Roumanie : virgule décimale, séparateur ";"), MedCalc 24.0 [50], JASP 0.98.1 [47], Epi Info 7.2.7 [51]. Le fichier mg2026_curat.csv sur Moodle. Durée indicative : 100 minutes.

Les colonnes de mg2026_curat.csv utilisées ici : S (glicemie0, mg/dL), V (hba1c0, %), W (findrisc, 0-26), X (diabet_ogtt : 1 diabète confirmé par HGPO, 0 pas de diabète) ; les données occupent les lignes 2-161. Le jeu de données mg2026 comprend 29 personnes diabétiques et 131 non diabétiques.

Déroulement

L12.1 Les indicateurs d'un tableau 2 x 2 dans Google Sheets (15 minutes)

Nous partons du problème classique du cours : un test appliqué à 440 patients, avec a = 150, b = 30, c = 4, d = 256 (les malades sur la première ligne, le test positif dans la première colonne).

Étape 1. Dans une nouvelle feuille "2x2", saisissez le tableau : B1 "T+", C1 "T-", A2 "B+", A3 "B-", puis B2 = 150, C2 = 30, B3 = 4, C3 = 256. En D2, écrivez =B2+C2, en D3 =B3+C3, en B4 =B2+B3, en C4 =C2+C3, en D4 =D2+D3.

Étape 2. À partir de A6, construisez la liste des indicateurs (l'étiquette dans la colonne A, la formule dans la colonne B) :

Tableau L12.1. Les formules des indicateurs dans Google Sheets.

Indicateur	Formule dans la colonne B	Résultat attendu
Prévalence	=D2/D4	40,9%
Sensibilité	=B2/D2	83,3%
Spécificité	=C3/D3	98,5%
VPP	=B2/B4	97,4%
VPN	=C3/C4	89,5%
Exactitude	=(B2+C3)/D4	92,3%
Taux d'erreur	=(C2+B3)/D4	7,7%
Indice de Youden	=B2/D2+C3/D3-1	0,818
RV+	=(B2/D2)/(1-C3/D3)	54,2
RV-	=(1-B2/D2)/(C3/D3)	0,17
DOR	=(B2*C3)/(C2*B3)	320

Attention au taux d'erreur : la formule contient les cellules C2 (faux négatifs, b) et B3 (faux positifs, c). Une formule qui utilise C3 (vrais négatifs) est fausse ; vérifiez que l'exactitude plus le taux d'erreur donnent 1.

Étape 3. L'intervalle de confiance de Wilson pour la sensibilité. Si le numérateur se trouve en B2 et le dénominateur en D2, la borne inférieure est :

=(B2/D2+1,96^2/(2*D2)-1,96*SQRT(B2/D2*(1-B2/D2)/D2+1,96^2/(4*D2^2)))/(1+1,96^2/D2)

et la borne supérieure s'obtient en remplaçant "-1,96*SQRT" par "+1,96*SQRT". Pour Se = 150/180, vous devez obtenir environ 77,2-88,1%. Adaptez la formule pour la spécificité (C3 et D3) : environ 96,1-99,4%.

Étape 4. Les valeurs prédictives pour une autre prévalence. Dans une cellule (par exemple F2), écrivez une nouvelle prévalence, 0,10, et calculez la VPP avec le théorème de Bayes : =(B2/D2)*F2/((B2/D2)*F2+(1-C3/D3)*(1-F2)), et la VPN avec =(C3/D3)*(1-F2)/((C3/D3)*(1-F2)+(1-B2/D2)*F2). Pour une prévalence de 10%, vous devez obtenir une VPP de 85,8% et une VPN de 98,2%. Comparez avec les valeurs obtenues pour une prévalence de 40,9%.

Étape 5. Refaites dans le cahier, sans logiciel, les deux problèmes de reconstitution du chapitre 12 (problèmes résolus 12.2 et 12.3), puis vérifiez-les en saisissant les tableaux obtenus dans la feuille "2x2". Toutes les formules se recalculent automatiquement.

L12.2 L'évaluation du test dans MedCalc (10 minutes)

MedCalc dispose d'une procédure dédiée au tableau 2 x 2 d'un test diagnostique, dans le menu Tests, option Diagnostic test (2 x 2 table). Si la version installée organise les menus autrement, cherchez la procédure par son nom ; la même procédure existe gratuitement en ligne, sur le site de MedCalc, sous le nom "Diagnostic test evaluation calculator".

Étape 1. Attention à l'orientation du tableau : dans MedCalc, la maladie est en colonnes (Disease present / Disease absent), et le test en lignes (Test positive / Test negative). Saisissez : Test positive et Disease present = 150 ; Test positive et Disease absent = 4 ; Test negative et Disease present = 30 ; Test negative et Disease absent = 256. Si vous saisissez 30 à la place de 4, tous les indicateurs seront faux, sans que le logiciel vous avertisse.

Étape 2. Lisez les résultats. Le logiciel affiche la sensibilité, la spécificité, le RV+, le RV-, la prévalence, la VPP, la VPN et l'exactitude, chacun avec son IC à 95%. Pour la Se et la Sp, MedCalc utilise des intervalles binomiaux exacts (Clopper-Pearson), un peu plus larges que ceux de Wilson : Se 83,3% (IC à 95% 77,1-88,5), Sp 98,5% (IC à 95% 96,1-99,6), exactitude 92,3% (IC à 95% 89,4-94,6), RV+ 54,2 (IC à 95% environ 20,4-143,6), RV- 0,17 (IC à 95% environ 0,12-0,23).

Étape 3. Dans le champ de la prévalence (Disease prevalence), saisissez 10%. La VPP et la VPN sont recalculées pour la nouvelle prévalence ; vous devez retrouver les valeurs de L12.1, étape 4. Notez à quel point l'intervalle de confiance du RV+ est large : avec seulement 4 faux positifs, l'estimation est très imprécise.

L12.3 Le tableau des seuils et la courbe ROC du FINDRISC dans Google Sheets (20 minutes)

Importez mg2026_curat.csv dans une feuille "Donnees". Dans une nouvelle feuille "ROC", construisez le tableau des seuils.

Étape 1. En A1:I1, écrivez l'en-tête : Seuil, VP, FP, FN, VN, Se, Sp, 1-Sp, J. En A2:A8, écrivez les seuils 16, 15, 14, 12, 11, 10, 8 (ordonnés du plus strict au plus permissif).

Étape 2. En B2 (VP) : =COUNTIFS(Donnees!$X$2:$X$161;1;Donnees!$W$2:$W$161;">="&A2). En C2 (FP) : =COUNTIFS(Donnees!$X$2:$X$161;0;Donnees!$W$2:$W$161;">="&A2). En D2 (FN) : =29-B2. En E2 (VN) : =131-C2. En F2 : =B2/29. En G2 : =E2/131. En H2 : =1-G2. En I2 : =F2+G2-1. Recopiez la ligne 2 jusqu'à la ligne 8.

Étape 3. Vérifiez les résultats avec le tableau 12.5 du chapitre. Pour le seuil ≥ 12, vous devez obtenir VP 22, FP 33, FN 7, VN 98, Se 0,759, Sp 0,748 ; pour ≥ 11, J maximal = 0,595.

Étape 4. Ajoutez au-dessus des seuils le point (0 ; 0) et en dessous le point (1 ; 1), dans une zone séparée à deux colonnes (1-Sp et Se), par exemple K1:L9, dans l'ordre croissant de 1-Sp. Sélectionnez la zone et insérez un graphique de type Scatter chart ; dans Chart editor > Customize > Series, cochez "Line" pour relier les points. Ajoutez la série de la diagonale, avec les points (0 ; 0) et (1 ; 1). Titres des axes : "1 - spécificité" et "sensibilité". Le graphique doit ressembler à la figure 12.2 du chapitre.

Étape 5. Calculez l'AUC par la méthode des trapèzes. Si 1-Sp se trouve en K2:K10 et la Se en L2:L10 (avec (0 ; 0) sur la première ligne et (1 ; 1) sur la dernière), écrivez en M3 =(K3-K2)*(L3+L2)/2, recopiez jusqu'à M10, puis =SUM(M3:M10). Avec les sept seuils ci-dessus, vous obtiendrez une aire d'environ 0,845 ; avec tous les seuils possibles du score (0-25), l'aire devient 0,852. Expliquez dans le cahier pourquoi l'aire calculée avec peu de seuils est plus petite.

Étape 6 (en option). Recommencez pour la glycémie à jeun (colonne S), avec les seuils 126, 120, 115, 110, 105, 100. Résultats de contrôle : pour ≥ 110, VP 24, FP 39 (Se 82,8%, Sp 70,2%) ; l'aire des trapèzes avec ces six seuils est de 0,836 (problème résolu 12.7).

L12.4 L'analyse ROC et la comparaison des courbes dans MedCalc (20 minutes)

Ouvrez mg2026_curat.csv dans MedCalc (File > Open, ou importez le fichier CSV et vérifiez que le séparateur décimal a été correctement reconnu).

Étape 1. Dans le menu Statistics, choisissez ROC curves > ROC curve analysis. Sous Variable, sélectionnez findrisc ; sous Classification variable, sélectionnez diabet_ogtt (codage 1 = maladie présente). Sous Methodology, vous pouvez choisir la méthode de DeLong ou celle de Hanley et McNeil pour l'erreur standard. Cochez l'option d'affichage des valeurs du critère (Criterion values and coordinates of the ROC curve) et le graphique ROC.

Étape 2. Résultats attendus : AUC 0,852 ; test de l'hypothèse AUC = 0,5 avec p < 0,001. L'intervalle de confiance dépend de la méthode choisie : 0,762-0,943 avec la formule de Hanley-McNeil, environ 0,782-0,923 avec l'erreur standard de DeLong ; MedCalc peut afficher par défaut un intervalle "binomial exact", légèrement différent. Notez dans le rapport quelle méthode le logiciel a utilisée.

Étape 3. L'indice de Youden. MedCalc affiche le critère associé à l'indice de Youden maximal : J = 0,595, avec le critère "> 10". Pour un score à valeurs entières, "> 10" signifie exactement "≥ 11", le seuil du chapitre (Se 93,1%, Sp 66,4%). Cette notation piège de nombreux étudiants à l'examen : lisez toujours le signe placé devant le critère.

Étape 4. Dans le tableau des critères, trouvez la ligne "> 11" (c'est-à-dire ≥ 12) et vérifiez Se 75,86% et Sp 74,81%, avec +LR 3,01 et -LR 0,32. Le tableau des critères est, en fait, le tableau des seuils que vous avez construit manuellement en L12.3.

Étape 5. La comparaison de deux courbes ROC. Dans le menu Statistics, choisissez ROC curves > Comparison of ROC curves. Sélectionnez les variables findrisc et glicemie0, avec la classification diabet_ogtt. Résultats attendus : AUC FINDRISC 0,852, AUC glycémie 0,844 ; différence 0,008 (IC à 95% environ -0,060 à 0,077), z d'environ 0,24, p d'environ 0,81 (méthode de DeLong pour courbes corrélées, car les deux tests ont été appliqués aux mêmes patients). Conclusion : aucune différence significative n'a été trouvée entre les capacités de discrimination des deux tests. Le graphique montre deux courbes qui se croisent.

Étape 6 (en option). Ajoutez hba1c0 à la comparaison. L'AUC de l'HbA1c est de 0,764 ; comparez visuellement et numériquement avec les deux autres tests.

L12.5 La courbe ROC dans JASP (10 minutes)

JASP n'a pas de menu séparé appelé "ROC" dans les modules de base, mais la courbe ROC peut être obtenue par la régression logistique et, dans les versions récentes, par le module Machine Learning. Les options disponibles diffèrent d'une version à l'autre ; la description ci-dessous correspond à l'organisation habituelle, et si vous ne trouvez pas une option, utilisez MedCalc (L12.4) ou le calcul dans Sheets (L12.3).

Étape 1. Ouvrez mg2026_curat.csv. Définissez diabet_ogtt comme variable nominale.

Étape 2. Dans le menu Regression, choisissez Logistic Regression. Placez diabet_ogtt dans Dependent Variable et findrisc dans Covariates.

Étape 3. Dans la section des statistiques ou de la performance du modèle (Statistics / Performance Diagnostics), cochez Confusion matrix et, sous Performance metrics, AUC, Sensitivity, Specificity, Precision et F-measure. Dans la section des graphiques de performance, cochez ROC plot.

Étape 4. Interprétation. L'AUC rapportée pour le modèle à un seul prédicteur est de 0,852, la même que l'AUC du score FINDRISC, car le modèle logistique transforme le score de façon monotone en une probabilité, sans changer l'ordre des patients. La matrice de confusion et la sensibilité affichées par JASP utilisent toutefois, par défaut, le seuil de 0,5 pour la probabilité prédite, et non un seuil du score ; avec une prévalence de 18,1%, très peu de patients dépassent une probabilité prédite de 0,5, de sorte que la sensibilité affichée sera faible et la spécificité élevée. C'est un bon exemple de "même courbe ROC, autre point de fonctionnement". Si la version permet de modifier le seuil (cutoff), essayez 0,18 et comparez avec le tableau des seuils.

Étape 5. L'alternative du module Machine Learning (s'il est activé, depuis le bouton "+" de la barre des modules) : une procédure de classification rapporte aussi la matrice de confusion, l'AUC et la courbe ROC, généralement sur un jeu de test séparé. Les résultats différeront de ceux ci-

dessus, car l'évaluation se fait sur une partie des données ; il s'agit, en fait, d'une validation interne simple.

L12.6 Le tableau 2 x 2 dans Epi Info (5 minutes)

Epi Info n'a pas de procédure dédiée aux tests diagnostiques sous la forme que nous utilisons, mais le tableau 2 x 2 s'obtient facilement, et les pourcentages en ligne et en colonne contiennent les indicateurs recherchés.

Étape 1. Dans Classic Analysis ou dans Visual Dashboard, importez le fichier et créez la variable findrisc12 (1 si findrisc ≥ 12, sinon 0), par exemple avec les commandes DEFINE et RECODE ou avec une variable calculée.

Étape 2. Exécutez un tableau (Tables) avec findrisc12 comme exposition (en lignes) et diabet_ogtt comme critère de jugement (en colonnes). Epi Info affiche les effectifs, les pourcentages en ligne et en colonne, le test du khi-deux et les mesures d'association (OR, RR).

Étape 3. Lisez : sur la ligne findrisc12 = 1, le pourcentage en ligne des diabétiques est VPP $\left(\frac{22}{55} = 40{,}0\%\right)$; sur la ligne findrisc12 = 0, le pourcentage en ligne des non-diabétiques est VPN $\left(\frac{98}{105} = 93{,}3\%\right)$. Dans la colonne diabet_ogtt = 1, le pourcentage en colonne des sujets ayant findrisc12 = 1 est la sensibilité (22/29 = 75,9%) ; dans la colonne diabet_ogtt = 0, le pourcentage des sujets ayant findrisc12 = 0 est la spécificité (98/131 = 74,8%). L'OR affiché (9,33) est le rapport de cotes diagnostique.

Étape 4. Notez que la significativité du test du khi-deux montre seulement que le test et la maladie sont associés, et non à quel point le test est bon. Un test avec une AUC de 0,6 peut avoir $p < 0{,}001$ dans un grand échantillon.

L12.7 Analyse de la performance d'un test de laboratoire : la glycémie à jeun pour le diabète confirmé par HGPO (15 minutes)

Utilisez les résultats obtenus jusqu'ici (ou exécutez les mêmes procédures pour glicemie0) et rédigez une analyse complète, dans le format des devoirs de la discipline.

(a) Hypothèse de recherche : la glycémie à jeun permet de distinguer les adultes avec et sans diabète confirmé par HGPO, dans la population à risque cardiométabolique suivie en médecine générale.

(b) Hypothèses statistiques : H0 : AUC = 0,5 (le test ne discrimine pas) ; H1 : AUC ≠ 0,5.

(c) Méthode : analyse ROC, avec l'HGPO comme standard de référence ; IC à 95% de l'AUC ; choix du seuil par l'indice de Youden et comparaison avec le seuil clinique de 126 mg/dL.

(d) Logiciel : MedCalc (ROC curve analysis), vérification dans Google Sheets.

(e) Résultats attendus : AUC 0,844 (IC à 95% 0,752-0,937, Hanley-McNeil), $p < 0{,}001$. Seuil de Youden ≥ 110 mg/dL (MedCalc affiche "> 109" ou la valeur immédiatement inférieure présente dans les données) : Se 82,8%, Sp 70,2%, VPP 38,1%, VPN 94,8%, J = 0,530. Au seuil clinique ≥ 126 mg/dL : Se 37,9%, Sp 94,7%, VPP 61,1%, VPN 87,3%, RV+ d'environ 7,1, RV- d'environ 0,66. Incluez le graphique ROC.

(f) Conclusion statistique : on rejette H0 ; la glycémie à jeun a une capacité de discrimination très bonne sur l'échelle du cours (0,81-0,90), mais la borne inférieure de l'intervalle de confiance (0,752) correspond à une performance "bonne".

(g) Conclusion de recherche : dans cette population, la glycémie à jeun est un test de triage utile au seuil d'environ 110 mg/dL (elle manque peu de diabétiques, mais adresse à

l'HGPO environ 4 patients non diabétiques sur 10). Le seuil de 126 mg/dL confirme le diabète avec une spécificité élevée, mais manque près des deux tiers des cas identifiés par l'HGPO ; un résultat inférieur à 126 mg/dL n'exclut donc pas le diabète. La performance de la glycémie ne diffère pas significativement de celle du score FINDRISC (p = 0,81).

Ajoutez un paragraphe sur les limites : petit nombre de malades (29), d'où des intervalles larges pour la sensibilité ; seuil choisi sur les mêmes données (surestimation optimiste) ; données simulées ; STARD 2015 demande aussi le diagramme de flux des participants [140].

Exercices supplémentaires

1. Sur le fichier Date_HbA1c disponible sur Moodle, choisissez une variable numérique de laboratoire et une variable dichotomique pouvant servir de référence (s'il en existe une). Construisez dans Google Sheets le tableau des seuils, la courbe ROC et l'AUC par les trapèzes ; vérifiez dans MedCalc. Rapportez au format a-g.
2. Sur le fichier BPOC disponible sur Moodle, évaluez dans quelle mesure un indice spirométrique au début de l'étude (par exemple IPB1) "prédit" la présence de râles (RAL). Exécutez l'analyse ROC dans MedCalc, rapportez l'AUC avec son IC à 95% et commentez si une telle analyse a un sens clinique.
3. Problème avec données (solution complète). Un test a une Sp = 0,90, une VPP = 0,60, VN = 270 et FN = 10. Reconstituez le tableau et calculez la Se, la VPN et le taux d'erreur.

Solution. d = 270 et Sp $= \frac{d}{c+d} = 0{,}90$, donc c + d = 300 et c = 30. À partir de VPP $= \frac{a}{a+30} = 0{,}60$: a = 0,60a + 18, donc 0,40a = 18 et a = 45. b = 10. Tableau : a = 45, b = 10, c = 30, d = 270 ; N = 355. Se $= \frac{45}{55} = 81{,}8\%$; VPN $= \frac{270}{280} = 96{,}4\%$; TE $= \frac{10+30}{355} = 11{,}3\%$; Ex $= \frac{315}{355} = 88{,}7\%$.

4. Problème avec données (solution complète). Pour l'IL-6 dans le sepsis néonatal, on a rapporté une Se de 80,1% et une Sp de 85,7% [126]. En utilisant le calculateur de L12.1 (étape 4), calculez la VPP et la VPN pour des prévalences de 5% et de 30%.

Solution. $RV^+ = \frac{0{,}801}{0{,}143} = 5{,}60$; $RV^- = \frac{0{,}199}{0{,}857} = 0{,}232$. À 5% : cote 0,0526 ; positif 0,0526 x 5,60 = 0,295, VPP 22,8% ; négatif 0,0526 x 0,232 = 0,0122, probabilité post-test 1,2%, VPN 98,8%. À 30% : cote 0,429 ; positif 0,429 x 5,60 = 2,40, VPP 70,6% ; négatif 0,429 x 0,232 = 0,0995, probabilité post-test 9,1%, VPN 90,9%.

Ce qu'il faut rendre et barème indicatif

Vous déposez sur Moodle le fichier Google Sheets (feuilles "2x2", "ROC", éventuellement "Glycemie"), les captures de MedCalc (analyse ROC du FINDRISC et comparaison des courbes) et le document d'une page contenant l'analyse de la glycémie à jeun au format a-g.

Barème indicatif (10 points) : feuille "2x2" avec tous les indicateurs corrects, y compris le taux d'erreur (b + c)/N et l'IC de Wilson (2 points) ; évaluation dans MedCalc avec l'orientation correcte du tableau et le recalcul pour une prévalence de 10% (1 point) ; tableau des seuils, courbe ROC et AUC par les trapèzes dans Sheets (2 points) ; analyse ROC et comparaison des courbes dans MedCalc, avec l'interprétation du critère "> 10" (2 points) ; JASP ou Epi Info, avec l'explication de la différence entre le point de fonctionnement par défaut et les seuils du score (1 point) ; analyse de la glycémie au format a-g, avec les limites (2 points).

Variante hors ligne

Sans accès aux logiciels, la séance peut se faire entièrement sur papier. Les indicateurs de L12.1 se calculent avec une calculatrice de poche ; les tableaux des seuils du FINDRISC et de la glycémie sont donnés au chapitre 12 (tableau 12.5 et problème résolu 12.7) ; la courbe ROC se

dessine sur du papier millimétré, et l'AUC s'estime par la méthode des trapèzes. Pour l'analyse ROC et la comparaison des courbes, utilisez les résultats de référence de L12.4 (AUC 0,852 et 0,844, différence non significative, p = 0,81) et concentrez-vous sur l'interprétation et sur la rédaction au format a-g. MedCalc propose aussi la version en ligne gratuite du calculateur pour le tableau 2 x 2, utilisable depuis un téléphone lorsque le réseau revient.

Travaux pratiques 13. Conception d'une mini-étude clinique et analyse intégrée d'un jeu de données complet

Objectif

Cette séance de travaux pratiques relie les deux extrémités d'une étude : celle où les données naissent et celle où elles se transforment en conclusions. Dans la première partie, vous concevrez une mini-étude observationnelle, vous construirez le formulaire de recueil des données dans Google Forms et vous simulerez le recueil et la centralisation des données dans Google Sheets. Dans la seconde partie, vous analyserez entièrement le jeu de données mg2026, de la description jusqu'à la courbe ROC et à l'analyse du risque, et vous rédigerez un rapport final au format académique. La séance prépare directement l'examen pratique des travaux pratiques 14 et applique les notions du chapitre 13.

À la fin, l'étudiant doit être capable :

- de transformer une question clinique en une fiche de projet (PICO, schéma d'étude, variables, taille de l'échantillon) ;
- de construire un formulaire Google Forms avec une page de consentement, des sections, des embranchements et des règles de validation, relié à une feuille Google Sheets ;
- de vérifier, de nettoyer et de documenter un jeu de données brut, avec un journal des modifications ;
- de choisir et d'appliquer les tests appropriés sur le jeu de données mg2026 dans JASP et MedCalc et de vérifier ses résultats à l'aide des valeurs de contrôle de la fiche ;
- de rédiger un rapport IMRaD de 2-3 pages, avec un tableau, un graphique, des résultats correctement rapportés et les limites.

Outils

- Google Forms et Google Sheets (compte institutionnel), avec les paramètres régionaux pour la Roumanie : le séparateur décimal est la virgule, et le séparateur d'arguments dans les formules est ";" [43,45,46,148].
- JASP 0.98.1 [47], MedCalc 24.0 [50], en option Epi Info 7.2.7 (StatCalc) [51] ; les guides de travail dans JASP et le recueil d'applications de la discipline [14,48,49].
- Les fichiers mg2026_curat.csv et mg2026_brut.csv, sur la plateforme Moodle de la discipline [52].
- Le chapitre 13 de cet ouvrage, pour le dictionnaire des données et pour le plan d'analyse.
- Durée indicative : 100 minutes (environ 40 minutes pour la partie conception et formulaire, 50 pour l'analyse, le reste pour le rapport, qui est finalisé à la maison).

Déroulement

L13.1 La fiche de projet de la mini-étude

Travaillez en équipes de 3-4 étudiants. Chaque équipe conçoit une mini-étude observationnelle transversale dans un cabinet de médecine générale. Le thème proposé, que vous pouvez adapter : "L'activité physique, l'indice de masse corporelle et la pression artérielle chez les adultes qui consultent au cabinet de médecine générale". Remplissez, dans un document Google Docs commun, une fiche de projet comportant les rubriques suivantes.

1. La question de recherche au format PICO [24]. Exemple : P, adultes de 18-80 ans ayant consulté au cabinet au cours d'un mois ; I (l'exposition), activité physique réduite (moins de 150 minutes par semaine) ; C, activité physique d'au moins 150 minutes par semaine ; O, l'hypertension artérielle (PAS ≥ 140 mmHg ou PAD ≥ 90 mmHg) et la valeur de la PAS.
2. Le schéma d'étude : transversal, un seul temps de mesure. Notez explicitement ce que ce schéma ne permet pas (la séquence temporelle, et donc la conclusion causale).
3. Les hypothèses statistiques pour le critère de jugement principal, en mots et sous forme mathématique. Exemple : H0, la moyenne de la PAS est la même chez les personnes peu actives et chez les personnes actives ($\mu 1 = \mu 2$) ; H1, les moyennes diffèrent ($\mu 1 \neq \mu 2$).
4. Les variables, avec le type de chacune (numérique, ordinale, nominale) et le test que vous utiliserez. Au minimum 10 variables : code, âge, sexe, milieu de résidence, niveau d'études, tabagisme (avec le nombre de cigarettes par jour pour les fumeurs), activité physique en minutes par semaine, taille, poids, PAS, PAD, un item de Likert ("Dans quelle mesure êtes-vous satisfait de votre état de santé ?", 1-5).
5. La taille de l'échantillon. Si vous voulez estimer la prévalence de l'HTA, supposée de 30%, avec une marge de ±10 points de pourcentage, $n = \frac{1{,}96^2 \times 0{,}30 \times 0{,}70}{0{,}10^2} = 80{,}7$, soit 81 patients. Faites aussi le calcul pour une marge de ±5 points de pourcentage et commentez la différence (le résultat correct est 323).
6. Les aspects éthiques : l'information et le consentement, l'anonymisation (code d'étude à la place du nom), la minimisation des données (l'âge en années, et non la date de naissance), l'avis du comité d'éthique pour une étude réelle [32,35].

Résultat attendu : la fiche de projet d'une page, avec le dictionnaire des données sous forme de tableau (nom de la variable, description, type, unité, valeurs autorisées), sur le modèle du tableau 13.2.

L13.2 Construction du formulaire dans Google Forms

Ouvrez Google Forms depuis le compte institutionnel et créez un nouveau formulaire, portant le titre de l'étude. Construisez-le par sections, dans l'ordre ci-dessous. Le bouton pour une nouvelle question est le signe "+" de la barre latérale ; le bouton pour une nouvelle section est le symbole à deux rectangles de la même barre.

Section 1, Information et consentement. Ajoutez un bloc de texte (titre et description) qui explique le but de l'étude, quelles données sont recueillies, que la participation est volontaire, que les données sont anonymes et que la personne peut se retirer à tout moment. Ajoutez ensuite une question de type "Choix multiples", marquée "Obligatoire" : "Acceptez-vous de participer à cette étude ?", avec les options "Oui, j'accepte" et "Je n'accepte pas". Dans le menu à trois points de la question, choisissez "Accéder à la section en fonction de la réponse" (dans l'interface en anglais, "Go to section based on answer"). Pour "Oui", choisissez "Passer à la section suivante", pour "Non", choisissez "Envoyer le formulaire" [148]. Ainsi, une personne qui refuse ne voit aucune question sur sa santé.

Section 2, Données générales. Le code du participant ("Réponse courte", obligatoire, avec la validation "Expression régulière", "Correspond à", le modèle P[0-9][0-9][0-9], le message d'erreur "Le code a la forme P001") ; l'âge ("Réponse courte", validation "Nombre", "Entre", 18 et 80, message "Saisissez l'âge en années révolues, entre 18 et 80") ; le sexe ("Choix multiples" : F, M) ; le milieu de résidence ("Choix multiples" : urbain, rural) ; le niveau d'études ("Choix multiples", dans l'ordre : collège, lycée ou école professionnelle, universitaire, post-universitaire). Les règles de validation s'activent depuis le menu à trois points de la question, option "Validation de la réponse" ("Response validation") [45].

Section 3, Mesures. La taille en cm (validation "Nombre", "Entre" 120 et 220, message "La taille en centimètres, par exemple 168") ; le poids en kg (entre 30 et 250) ; la PAS (entre 70 et 250 mmHg) et la PAD (entre 40 et 150 mmHg). N'ajoutez pas de question pour l'IMC : il se calcule dans la feuille de calcul à partir du poids et de la taille. Observez comment chaque règle aurait arrêté l'une des erreurs du fichier mg2026_brut.csv (taille 1,68, poids 8,6, PAS 1450).

Section 4, Mode de vie. La question "Fumez-vous actuellement ?" ("Choix multiples" : "Oui", "Non, j'ai arrêté", "Non, je n'ai jamais fumé"), avec embranchement : "Oui" mène à la section 4a, les autres options à la section 5. La section 4a contient une seule question : "Combien de cigarettes fumez-vous en moyenne par jour ?" (nombre entre 1 et 80). Viennent ensuite les minutes d'activité physique par semaine ("Réponse courte", nombre entre 0 et 3000) et l'item de Likert ("Échelle linéaire", de 1 à 5, avec les libellés "pas du tout satisfait" et "très satisfait").

Section 5, Fin. Un texte de remerciement. Dans les paramètres (l'icône en forme de roue dentée), désactivez la collecte des adresses e-mail : pour une étude anonyme, l'adresse e-mail serait une donnée d'identification inutile.

Vérifiez le formulaire avec le bouton "Aperçu" (l'icône en forme d'œil) : essayez de saisir un âge de 15 ans, une taille de 1,75 et la réponse "Je n'accepte pas" au consentement. Le formulaire doit refuser les deux premières saisies et se terminer immédiatement à la troisième. Résultat attendu : un formulaire fonctionnel, avec 5 sections (plus la 4a), au moins 3 règles de validation et 2 embranchements.

L13.3 Simulation du recueil et centralisation des données dans Google Sheets

Dans l'onglet "Réponses" du formulaire, choisissez "Associer à Sheets" ("Link to Sheets") et créez une nouvelle feuille [46]. Chaque membre de l'équipe remplit le formulaire 5 fois, avec des données fictives mais plausibles, comme s'il s'agissait de patients différents ; au moins l'un d'eux refuse de participer, et au moins deux sont fumeurs. L'équipe obtient ainsi 15-20 enregistrements.

Ouvrez la feuille des réponses et effectuez les vérifications suivantes, dans un nouvel onglet nommé "verification".

1. Renommez les colonnes avec des noms courts, sans espaces et sans signes diacritiques (cod, varsta, sex, mediu, educatie, inaltime, greutate, tas, tad, fumat, tigari, activitate, satisfactie). La première colonne, "Horodateur", est conservée pour la traçabilité.
2. Comptez les enregistrements complets : =COUNTA(B2:B21). Comptez les refus : =COUNTIF(B2:B21;"Je n'accepte pas").
3. Vérifiez les doublons de code : =COUNTIF(B$2:B$21;B2) dans une colonne auxiliaire ; toute valeur supérieure à 1 signale un doublon.
4. Calculez l'IMC dans une nouvelle colonne : =ROUND(G2/(F2/100)^2;1), en supposant le poids dans la colonne G et la taille dans la colonne F.
5. Calculez les statistiques descriptives de la PAS : =AVERAGE(H2:H21), =STDEV(H2:H21), =MEDIAN(H2:H21), =MIN(H2:H21), =MAX(H2:H21).
6. Construisez un tableau croisé dynamique ("Insertion" > "Tableau croisé dynamique") avec le sexe en lignes et le tabagisme en colonnes, avec les valeurs comptées [44].
7. Créez un onglet "modifications", avec les colonnes : date, auteur, ligne, variable, ancienne valeur, nouvelle valeur, motif, source. Introduisez volontairement une erreur (par exemple, remplacez manuellement une PAS par 1500), puis corrigez-la et notez la correction dans le journal.

Discutez en équipe : quelles règles de validation ont fonctionné, quelles erreurs sont tout de même passées (par exemple une PAS de 95 avec une PAD de 110, toutes deux dans l'intervalle autorisé, mais cliniquement incompatibles) et quelle vérification logique il faudrait ajouter. Résultat

attendu : la feuille de calcul avec les onglets "reponses", "verification" et "modifications", partagée avec l'enseignant en lecture seule.

L13.4 Nettoyage du fichier mg2026_brut.csv

Importez le fichier mg2026_brut.csv dans une nouvelle feuille Google Sheets ("Fichier" > "Importer" > "Importer"). Sans regarder le chapitre 13, cherchez les erreurs en utilisant uniquement les outils ci-dessous, puis comparez votre liste avec le tableau 13.3.

- Le nombre de lignes de données : vous devez obtenir 161 (une ligne est dupliquée).
- Les doublons : triez selon la colonne id ou utilisez =COUNTIF(A$2:A$162;A2) ; le patient P091 apparaît deux fois.
- Les variables catégorielles : pour sex, mediu et fumat, utilisez "Données" > "Créer un filtre" et regardez la liste des valeurs distinctes. Vous devez trouver : pour sex, "f" et "Feminin" en plus de F et M ; pour mediu, "Urban" ; pour fumat, "Fumator " (avec une espace).
- Les variables numériques : =MIN() et =MAX() sur chaque colonne. Vous devez trouver tas0 = 1540, une taille de 1,72 et un poids de 7,26. Utilisez =COUNT() comparé à =COUNTA() : la différence révèle les valeurs stockées comme texte (la glycémie "87 mg/dl", l'HbA1c "5,2").
- Les valeurs manquantes : =COUNTBLANK() sur chaque colonne ; il manque l'âge pour P042 et la glycémie à 6 mois pour P134.

Notez dans le journal des modifications comment vous traiteriez chaque erreur. Rappelons la règle : les erreurs de forme se corrigent directement, les valeurs impossibles ne se corrigent qu'à partir du document source. Dans cet exercice, le "document source" est le fichier mg2026_curat.csv : vérifiez-y les valeurs des patients P034, P057 et P102 et notez, pour chacun, quelle correction vous auriez devinée et ce que montre la source. Pour la taille et le poids, utilisez aussi l'IMC conservé dans le tableau comme vérification indépendante. Résultat attendu : 12 problèmes identifiés, chacun avec sa méthode de détection et la décision de correction.

L13.5 Analyse complète du jeu de données mg2026

Travaillez sur mg2026_curat.csv. Pour chaque analyse, écrivez dans le rapport les éléments (a)-(g) du format des devoirs de la discipline : l'hypothèse de recherche, les hypothèses statistiques, la méthode et sa justification, le logiciel et la procédure de calcul, le résultat avec un graphique, la conclusion statistique et la conclusion de recherche. Les valeurs de contrôle ci-dessous vous permettent de vérifier vos résultats.

1. La description. Dans JASP : "Descriptives" > "Descriptive Statistics", variables varsta, imc0, tas0, glicemie0, trigliceride ; dans "Statistics", cochez la moyenne, l'écart-type, la médiane, les quartiles, l'asymétrie, l'aplatissement et "Shapiro-Wilk test" ; dans "Plots", "Distribution plots" et "Boxplots". Valeurs de contrôle : âge 56,29 ± 8,77 ans (Shapiro-Wilk p = 0,687) ; IMC médiane 28,25, EIQ 26,2-31,4, asymétrie 1,42, p < 0,001 ; PAS $131{,}13 \pm 13{,}88$ mmHg$(p = 0{,}303)$; triglycérides médiane 151,5 mg/dL, asymétrie 0,95. Pour les variables catégorielles ("Frequencies" > "Frequency tables") : F 86, M 74 ; A 53, B 53, C 54 ; fumeurs 53, anciens fumeurs 33, non-fumeurs 74.
2. Le tableau 1 par bras. Refaites la description avec "Split" selon brat. Comparez les résultats avec le tableau 13.4 et expliquez pourquoi vous n'ajoutez pas de valeurs p.
3. L'intervalle de confiance. Dans JASP, "T-Tests" > "One Sample T-Test" sur tas0, avec "Location parameter" et "Confidence interval" cochés : IC à 95% 128,96-133,30 mmHg. Dans Google Sheets : =AVERAGE(M2:M161) - T.INV.2T(0,05;159)*STDEV(M2:M161)/SQRT(160), en supposant que tas0 se trouve dans la colonne M.
4. Le critère de jugement principal. Créez la variable dglic = glicemie0 - glicemie6 (dans JASP, bouton "+" de l'en-tête du tableau de données, "Computed column"). "ANOVA" > "ANOVA", variable dépendante dglic, facteur brat ; dans "Assumption Checks", cochez "Homogeneity tests" ;

dans "Post Hoc Tests", placez brat et cochez "Tukey" ; dans "Additional Options", "Estimates of effect size" (eta²). Valeurs de contrôle : moyennes 3,0 / 11,1 / 16,3 mg/dL ; Levene p = 0,163 ; F(2 ; 157) = 22,93, p < 0,001 ; $\eta^2 = 0{,}226$; Tukey B-A 8,1 (3,4-12,8), C-A 13,3 (8,6-18,0), C-B 5,2 (0,6-9,9). Dans la section "Nonparametrics" du même menu, Kruskal-Wallis : H = 36,63, p < 0,001.

5. Données appariées et mesures répétées. "T-Tests" > "Paired Samples T-Test", glicemie0 et glicemie6, avec "Normality" coché : différence moyenne de 10,19 mg/dL (IC à 95% 8,38-11,99), t = 11,14, ddl 159, p < 0,001, Shapiro-Wilk sur les différences p = 0,725. "ANOVA" > "Repeated Measures ANOVA" avec glicemie0, glicemie3, glicemie6 comme niveaux du facteur temps : F = 64,97, p < 0,001 ; dans "Nonparametrics", Friedman $\chi^2 = 74{,}67$, p < 0,001.
6. Proportions et risque. "Frequencies" > "Contingency Tables", lignes brat, colonnes scadere5 ; dans "Statistics", cochez χ^2, dans "Cells" les effectifs attendus et les pourcentages en ligne : A 6/53, B 16/53, C 16/54 ; $\chi^2 = 6{,}77$, ddl 2, p = 0,034 ; plus petit effectif attendu 12,6. Pour C contre A, dans MedCalc : "Tests" > "Relative risk", avec 16 événements sur 54 et 6 sur 53 : RR 2,62 (IC à 95% 1,11-6,17), NST 5,5 (soit 6). La même analyse dans Epi Info, "StatCalc" > "2 x 2 tables".
7. Variables ordinales. "ANOVA" > "ANOVA" > "Nonparametrics", satisfactie selon brat : Kruskal-Wallis H = 32,16, p < 0,001. "T-Tests" > "Independent Samples T-Test" avec "Mann-Whitney" coché, uniquement pour les bras A et C (filtrez les données) : U = 2265,5, p < 0,001, corrélation bisériale des rangs 0,58.
8. Corrélation et régression. "Regression" > "Correlation", imc0 et tas0, avec Pearson et Spearman : r = 0,405, rho = 0,372, tous deux p < 0,001 ; nuage de points depuis "Plots". "Regression" > "Linear Regression", variable dépendante tas0, covariable imc0 : PAS = 89,47 + 1,435 x IMC, $R^2 = 0{,}164$. Ajoutez varsta, sex et fumat : $R^2 = 0{,}308$, coefficient de l'IMC 1,389.
9. L'association obésité-diabète. Créez la variable obez (imc0 ≥ 30). "Contingency Tables" obez x diabet_ogtt, avec "Log odds ratio" coché : 18/54 contre 11/106 ; $\chi^2 = 12{,}70$, p < 0,001 ; OR 4,32 (IC à 95% 1,86-10,03) ; rapport de prévalences, calculé dans MedCalc, 3,21 (IC à 95% 1,64-6,31). "Regression" > "Logistic Regression", variable dépendante diabet_ogtt, covariables obez, varsta, sex, avec "Odds ratios" coché : OR de l'obésité 6,01, de l'âge 1,11 par an.
10. Le test diagnostique. Dans MedCalc : "Statistics" > "ROC curves" > "ROC curve analysis", variable findrisc, classification diabet_ogtt ; cochez "Criterion values and coordinates of the ROC curve". Valeurs de contrôle : AUC 0,852 (IC à 95% 0,762-0,943), seuil de Youden > 10, c'est-à-dire ≥ 11 (Se 93,1%, Sp 66,4%). Au seuil ≥ 12, construisez à la main le tableau 2 x 2 (VP 22, FP 33, FN 7, VN 98) et calculez Se 75,9%, Sp 74,8%, VPP 40,0%, VPN 93,3%, RV+ 3,01, RV- 0,32.

Attention à la dernière analyse : MedCalc rapporte le seuil sous la forme "> 10", ce qui, pour un score à valeurs entières, signifie "≥ 11". De nombreux étudiants écrivent un seuil erroné dans leur rapport précisément pour cette raison.

L13.6 Le rapport final au format IMRaD

Le rapport individuel, de 2-3 pages, suit la structure d'un article scientifique : Introduction, Méthodes, Résultats et Discussion (IMRaD). Il s'appuie sur l'analyse de L13.5, et non sur les données fictives de l'équipe.

L'introduction (un paragraphe) présente le problème (le risque cardiométabolique en médecine générale), ce que l'on sait et la question de l'étude, au format PICO. Les méthodes décrivent le schéma d'étude (essai randomisé pragmatique simulé, 3 bras, 12 cabinets), les participants et les critères, les interventions, le critère de jugement principal et les critères secondaires, l'analyse statistique (les tests, le seuil $\alpha = 0{,}05$, les logiciels avec leur version) et précisent que les données sont simulées. Les résultats commencent par le flux des participants (figure 13.3) et par le tableau 1 (un tableau de 8-10 lignes suffit), se poursuivent avec le critère principal (description, test, taille d'effet, IC) et avec deux critères secondaires au choix, puis avec une analyse observationnelle (obésité-diabète ou FINDRISC). Incluez un seul graphique, choisi

pour le critère principal (boîtes à moustaches par bras ou diagramme des différences avec IC). La discussion résume le résultat principal en une phrase, l'interprète cliniquement (par exemple, le NST), discute les limites (absence d'insu, critère de substitution, suivi court, données simulées sans perdus de vue) et formule la conclusion.

Respectez les règles de présentation des résultats de la section 14.4 : valeurs p exactes, taille d'effet avec IC à 95%, unités de mesure, "on ne rejette pas H0" au lieu de "on accepte H0", pas de conclusions causales tirées des analyses observationnelles. Un modèle de phrase pour le résultat principal : "À 6 mois, la baisse moyenne de la glycémie à jeun a été de 3,0 mg/dL dans le bras A, de 11,1 mg/dL dans le bras B et de 16,3 mg/dL dans le bras C (ANOVA, $F(2;\ 157) = 22{,}93$, $p < 0{,}001$; $\eta^2 = 0{,}23$)."

Exercices supplémentaires

1. Sur le fichier BD_Studenti disponible sur Moodle, concevez le formulaire Google Forms qui aurait produit ce fichier : pour chaque variable, choisissez le type de question et la règle de validation. Identifiez les variables qui ont probablement été recueillies par classes et proposez la question qui aurait permis de recueillir la valeur exacte.
2. Sur le fichier BPOC (étude croisée avec des mesures au début de l'étude et après le traitement médicamenteux), écrivez le plan d'analyse pour la question "la fréquence cardiaque se modifie-t-elle après le traitement ?" et pour la question "la capacité vitale après le traitement diffère-t-elle selon le médicament ?". Précisez, pour chacune, pourquoi le test choisi est un test pour données appariées ou pour groupes indépendants. Appliquez les tests et interprétez.
3. Sur le fichier Date_HbA1c, écrivez le tableau 1 pour deux groupes de votre choix, en choisissant correctement entre moyenne ± ÉT et médiane (EIQ) pour chaque variable, selon le test de normalité.
4. Une équipe veut estimer la prévalence du tabagisme parmi les patients d'un cabinet, supposée de 25%, avec une marge de ±5 points de pourcentage et un niveau de confiance de 95%. Calculez n. Solution : $n = \frac{1{,}96^2 \times 0{,}25 \times 0{,}75}{0{,}05^2} = \frac{3{,}8416 \times 0{,}1875}{0{,}0025} = 288{,}1$, soit 289 patients. Avec un taux de réponse estimé à 80%, il faut inviter 289/0,8 = 361,3, c'est-à-dire 362 patients.
5. Dans un formulaire, la question "À quelle fréquence faites-vous du sport ? rarement / parfois / souvent" a produit des réponses difficiles à interpréter. Réécrivez-la de manière à produire une variable numérique et expliquez ce que vous y gagnez pour l'analyse. Solution modèle : "Au cours des 7 derniers jours, combien de minutes au total avez-vous consacrées à une activité physique modérée ou intense ? ___ minutes" ; la variable numérique permet de calculer la moyenne, la médiane, la corrélation avec l'IMC et une dichotomisation ultérieure à un seuil (150 minutes), alors que les options "rarement / parfois / souvent" sont interprétées différemment par chaque répondant.

Ce qu'il faut rendre et barème indicatif

Vous déposez sur Moodle, dans la section de la séance 13 : la fiche de projet (avec le dictionnaire des données), le lien du formulaire (avec droit de lecture), la feuille de calcul avec les onglets "reponses", "verification" et "modifications", la liste des erreurs de mg2026_brut.csv et le rapport IMRaD individuel, au format PDF.

Tableau L13.1. Barème indicatif pour les travaux pratiques 13 (100 points).

Composante	Critères	Points
Fiche de projet	PICO correct ; schéma d'étude nommé, avec ses limites ; hypothèses statistiques en mots et sous forme mathématique ; taille de l'échantillon correctement calculée ; aspects éthiques	15
Dictionnaire des données	toutes les variables, avec type, unité,	5

Composante	Critères	Points
	valeurs autorisées et codes	
Formulaire Google Forms	page de consentement avec embranchement de sortie ; au moins 3 règles de validation avec des messages clairs ; embranchement pour les fumeurs ; types de questions appropriés ; pas de collecte d'adresses e-mail	15
Centralisation dans Sheets	colonnes renommées ; vérifications (doublons, IMC, statistiques descriptives) ; tableau croisé dynamique ; journal des modifications	10
Nettoyage de mg2026_brut	erreurs identifiées (au moins 10 sur 12), avec la méthode de détection et la décision correcte	10
Analyse de mg2026	tests choisis et justifiés correctement ; hypothèses d'application vérifiées ; résultats concordant avec les valeurs de contrôle ; format (a)-(g)	25
Rapport IMRaD	structure ; tableau 1 ; un graphique approprié ; résultats avec taille d'effet et IC ; limites ; conclusion sans exagération causale	20
Total		100

Erreurs qui font perdre des points quelle que soit la composante : la formulation "on accepte H0" ; un test pour groupes indépendants appliqué à des données appariées, ou l'inverse ; moyenne ± ÉT pour des variables visiblement asymétriques sans discussion ; OR présenté comme un risque relatif ; conclusions causales tirées de l'analyse obésité-diabète ; seuil FINDRISC rapporté de façon erronée ("> 10" transcrit en "≥ 10").

Variante hors ligne

Si vous n'avez pas accès à Internet, la partie conception se fait sur papier : dessinez le formulaire par sections, avec des flèches pour les embranchements et la règle de validation écrite à côté de chaque champ, et simulez le "recueil" en faisant remplir à chaque étudiant 5 formulaires imprimés. La centralisation se fait dans un fichier Excel ou LibreOffice Calc, avec les mêmes formules (le séparateur d'arguments reste ";" avec les paramètres régionaux roumains). Le nettoyage du fichier mg2026_brut.csv et l'analyse de mg2026 se font dans JASP et MedCalc, installés localement sur les ordinateurs du laboratoire ; Epi Info peut remplacer MedCalc pour les tableaux 2 x 2. Pour la courbe ROC sans MedCalc, construisez à la main le tableau avec les seuils 8, 10, 12, 14, 15 et 16 du score FINDRISC (le tableau du chapitre 12) et placez les points (1 - Sp, Se) sur du papier millimétré ; l'aire peut être approchée par la règle des trapèzes.

Travaux pratiques 14. L'examen pratique : sujet type, barème et plan de préparation

Objectif

La dernière séance de travaux pratiques est consacrée à l'examen pratique. Vous découvrirez comment il est organisé, vous traiterez entièrement deux sujets types, avec les solutions complètes et les valeurs attendues, vous verrez comment chaque consigne est notée et quelles sont les erreurs qui coûtent le plus de points. La séance se termine par un plan de préparation sur deux semaines et par une liste de vérification des compétences pratiques dans les logiciels. Un étudiant capable de résoudre seul, dans le temps imparti, les deux sujets types est prêt pour l'examen.

À la fin, l'étudiant doit être capable :

- de décrire la structure de l'examen pratique et son poids dans la note finale ;
- de résoudre, dans le temps imparti, un sujet comportant 6-8 consignes qui couvrent la description, l'estimation, les tests, la corrélation et la régression, le risque et le test diagnostique ;
- de rédiger chaque solution au format (a)-(g) de la discipline ;
- de vérifier ses résultats et d'éviter les erreurs fréquentes recensées dans le barème ;
- d'organiser sa préparation à partir de la liste des compétences.

Outils

- JASP 0.98.1 [47], MedCalc 24.0 [50], Epi Info 7.2.7 [51], Google Sheets avec les paramètres régionaux pour la Roumanie (séparateur d'arguments ";") [43].
- Le fichier mg2026_curat.csv et les fichiers classiques de la discipline sur Moodle [52] ; les recueils d'applications pratiques [14,149] et le guide de travail dans JASP [49].
- Les chapitres 2-12 et 14 de cet ouvrage ; le tableau de synthèse des formules (tableau 14.2).
- Durée indicative : 100 minutes, dont 60 pour la résolution d'un sujet type en conditions d'examen et 40 pour sa correction à l'aide du barème.

Déroulement

L14.1 Comment se déroule l'examen pratique

La note finale de biostatistique se compose de trois parties, conformément à la fiche de la discipline : l'examen de type QCM de 50 questions, corrigé selon le système du concours de résidanat (50%), l'activité et les devoirs au cours du semestre (20%) et l'examen pratique en laboratoire (30%). L'examen pratique vérifie exactement ce que vous avez pratiqué en séance : ouvrir un fichier de données, choisir et appliquer la méthode appropriée dans un logiciel, interpréter le résultat et le rédiger correctement.

L'examen se passe individuellement, sur ordinateur, dans la salle de travaux pratiques. Vous recevez un sujet comportant 6-8 consignes, portant sur un jeu de données connu (du type mg2026 ou des fichiers des séances) ou sur un petit jeu de données nouveau, dont les valeurs sont fournies dans l'énoncé. Vous disposez des logiciels utilisés pendant le semestre. Les réponses s'écrivent dans un document (ou sur la copie d'examen), dans le format demandé pour les devoirs : (a) l'hypothèse de recherche ; (b) les hypothèses statistiques en mots et sous forme mathématique ; (c) la méthode choisie et sa justification ; (d) la procédure de calcul et le logiciel ; (e) le résultat, avec un graphique lorsqu'il est demandé ; (f) la conclusion statistique ; (g) la conclusion de recherche en mots. Toutes les consignes n'exigent pas les sept éléments : une consigne de statistique

descriptive ne comporte pas d'hypothèses statistiques, et une consigne d'estimation demande un intervalle de confiance, et non un test.

Ce qui est noté : le choix correct de la méthode et sa justification (généralement la plus grande partie de la note), l'exactitude des chiffres, l'interprétation statistique et l'interprétation médicale. Un chiffre correct obtenu avec une méthode erronée rapporte peu ; une méthode correcte avec une petite erreur de calcul, bien interprétée, rapporte presque tous les points. À l'examen, la perte de points la plus fréquente ne vient pas du calcul, mais de conclusions mal formulées ("on accepte H0", "la corrélation prouve que...").

L14.2 Sujet type 1 (le jeu de données mg2026)

Ouvrez le fichier mg2026_curat.csv. Considérez l'échantillon comme représentatif des adultes à risque cardiométabolique suivis dans les cabinets de médecine générale de l'ouest du pays. Traitez les consignes, avec α = 0,05.

1. Décrivez les variables HDL-cholestérol (hdl) et tour de taille (talie). Choisissez les indicateurs appropriés, justifiez votre choix par la forme de la distribution et réalisez pour chacune un graphique approprié.
2. Estimez la moyenne du HDL dans la population, avec un niveau de confiance de 95%. Dans quel intervalle se trouvent environ 95% des valeurs individuelles du HDL ?
3. Le HDL diffère-t-il entre les femmes et les hommes ?
4. Étudiez la relation entre le tour de taille et la glycémie à jeun à l'inclusion (glicemie0). Réalisez le nuage de points avec la droite de régression et estimez la glycémie d'un patient ayant un tour de taille de 100 cm.
5. La pression artérielle systolique à l'inclusion (tas0) diffère-t-elle entre les fumeurs, les anciens fumeurs et les non-fumeurs ?
6. Le tabagisme actuel est-il associé à l'hypertension artérielle (tas0 ≥ 140 mmHg) à l'inclusion ? Calculez la mesure d'association appropriée.
7. Dans le bras C, la PAS a-t-elle changé entre l'inclusion et 6 mois (tas0 par rapport à tas6) ?
8. Évaluez la glycémie à jeun ≥ 110 mg/dL comme test du diabète confirmé par HGPO (diabet_ogtt) : construisez le tableau 2 x 2 et calculez la Se, la Sp, la VPP, la VPN, l'exactitude, le taux d'erreur et les rapports de vraisemblance.

L14.3 Solution du sujet 1

Consigne 1. HDL : n = 160, moyenne 49,52 mg/dL, ÉT 11,78, erreur standard 0,931, médiane 49, EIQ 41-59, minimum 20, maximum 78, asymétrie 0,04 ; Shapiro-Wilk W = 0,991, p = 0,371. Nous ne rejetons pas la normalité, la moyenne et la médiane sont pratiquement égales ; nous décrivons donc par moyenne ± ÉT : 49,52 ± 11,78 mg/dL. Le graphique approprié est l'histogramme (avec la courbe normale superposée). Tour de taille : moyenne 88,54 cm, ÉT 12,11, médiane 86, EIQ 80,75-95, minimum 65, maximum 140, asymétrie 0,86 ; Shapiro-Wilk W = 0,961, p < 0,001 (p = 0,0002). La distribution est asymétrique à droite, la moyenne est supérieure à la médiane ; nous décrivons donc par la médiane et l'EIQ : 86 (80,75-95) cm. Graphique : boîte à moustaches ou histogramme. Dans JASP : "Descriptives" > "Descriptive Statistics", avec Shapiro-Wilk coché.

Consigne 2. (c) Estimation par intervalle de confiance pour la moyenne, avec la loi t (n = 160, ddl 159, t = 1,975). (d) ES $= \frac{11,78}{\sqrt{160}} = 0,931$; IC 95% $= 49,52 \pm 1,975 \times 0,931 = 49,52 \pm 1,84$. (e) IC à 95% de la moyenne : 47,68-51,36 mg/dL. L'intervalle pour les valeurs individuelles est $m \pm 1,96 \times s = 49,52 \pm 23,09$, soit 26,4-72,6 mg/dL. (g) Avec une confiance de 95%, la moyenne du HDL dans la population se situe entre 47,7 et 51,4 mg/dL ; environ 95% des patients

ont un HDL compris entre 26 et 73 mg/dL. Les deux intervalles répondent à des questions différentes, et les confondre est une erreur sanctionnée par le barème.

Consigne 3. (a) Le HDL diffère entre les sexes. (b) H0 : μF = μM (il n'y a pas de différence entre les moyennes de HDL chez les femmes et chez les hommes) ; H1 : μF ≠ μM. (c) Variable numérique, deux groupes indépendants ; la normalité n'est rejetée dans aucun groupe (Shapiro-Wilk p = 0,413 chez les femmes, p = 0,477 chez les hommes), mais le test de Levene rejette l'égalité des variances (p = 0,034) ; nous utilisons donc le test t de Welch. (d) JASP : "T-Tests" > "Independent Samples T-Test", avec "Welch" et "Equality of variances" cochés. (e) Femmes $53{,}92 \pm 11{,}70$ mg/dL$(n\,86)$, hommes $44{,}41 \pm 9{,}66$ mg/dL$(n\,74)$; t = 5,63, ddl 157,7, p < 0,001 ; différence de 9,51 mg/dL (IC à 95% 6,18-12,85) ; d de Cohen 0,88. (f) Nous rejetons H0. (g) Les femmes ont un HDL plus élevé d'environ 9,5 mg/dL en moyenne que les hommes, une différence importante (d 0,88), concordante avec la physiologie connue. La variante de Student aurait donné t = 5,55, avec la même conclusion ; le choix du test de Welch se justifie par le résultat du test de Levene.

Consigne 4. (a) Un tour de taille plus grand est associé à une glycémie plus élevée. (b) H0 : ρ = 0 (les variables ne sont pas corrélées) ; H1 : ρ ≠ 0. (c) Les deux variables sont numériques, mais aucune ne suit une loi normale (talie p = 0,0002 ; glicemie0 p = 0,037) ; la corrélation principale est donc celle de Spearman ; nous rapportons aussi le coefficient de Pearson et la droite de régression, explicitement demandée. (d) JASP : "Regression" > "Correlation" (Pearson et Spearman), puis "Regression" > "Linear Regression", variable dépendante glicemie0, covariable talie. (e) rho = 0,430, p < 0,001 ; r = 0,488, p < 0,001 ; $R^2 = 0{,}238$. La droite : glicemie0 = 49,17 + 0,636 x talie (IC à 95% de la pente 0,457-0,814) ; erreur standard de l'estimation 13,80 mg/dL. Pour un tour de taille de 100 cm : 49,17 + 63,55 = 112,7 mg/dL. (f) Nous rejetons H0. (g) Il existe une corrélation positive modérée entre le tour de taille et la glycémie : chaque centimètre supplémentaire est associé à une glycémie plus élevée d'environ 0,6 mg/dL ; le tour de taille explique environ 24% de la variabilité de la glycémie. La prédiction individuelle est imprécise (erreur standard de l'estimation proche de 14 mg/dL), et l'association ne prouve pas la causalité.

Consigne 5. (a) La PAS diffère selon le statut tabagique. (b) H0 : μfumeurs = μanciens = μnon-fumeurs ; H1 : au moins une moyenne diffère. (c) Variable numérique, trois groupes indépendants ; Shapiro-Wilk dans chaque groupe : p = 0,37 (anciens fumeurs), p = 0,86 (fumeurs), p = 0,23 (non-fumeurs) ; Levene p = 0,114 : ANOVA à un facteur avec post hoc de Tukey. (d) JASP : "ANOVA" > "ANOVA", variable dépendante tas0, facteur fumat, avec "Homogeneity tests", "Tukey" et eta². (e) Anciens fumeurs $127{,}5 \pm 14{,}4$ mmHg$(n\,33)$, fumeurs $137{,}0 \pm 14{,}6(n\,53)$, non-fumeurs $128{,}5 \pm 11{,}8(n\,74)$; F(2 ; 157) = 7,85, p < 0,001 (p = 0,0006) ; $\eta^2 = 0{,}091$. Tukey : fumeurs par rapport aux anciens fumeurs +9,5 mmHg (IC à 95% 2,5-16,5 ; p = 0,005) ; fumeurs par rapport aux non-fumeurs +8,5 mmHg (IC à 95% 2,9-14,2 ; p = 0,001) ; anciens fumeurs par rapport aux non-fumeurs 1,0 mmHg (IC à 95% -5,6 à 7,6 ; p = 0,94). Kruskal-Wallis, à titre de vérification, H = 12,77, p = 0,002. (f) Nous rejetons H0. (g) Les fumeurs actuels ont une PAS plus élevée de 8-10 mmHg que les anciens fumeurs et les non-fumeurs, qui ne diffèrent pas entre eux. La différence est cliniquement pertinente. S'agissant de données transversales, nous ne pouvons pas affirmer que le tabagisme produit cette différence ; la régression multiple du chapitre 9 montre toutefois qu'elle persiste à IMC, âge et sexe égaux (+6,7 mmHg par rapport aux anciens fumeurs).

Consigne 6. (a) Le tabagisme est associé à l'HTA. (b) H0 : la proportion d'HTA est la même chez les fumeurs et chez les non-fumeurs (ou RP = 1) ; H1 : les proportions diffèrent. (c) Deux variables dichotomiques (fumeur oui/non, HTA oui/non), schéma transversal : khi-deux ; la mesure d'association est le rapport de prévalences (RP), l'OR étant rapporté en complément. (d) Créez les variables fum (fumat = "fumator") et hta (tas0 ≥ 140) dans JASP ("Computed column"),

puis "Frequencies" > "Contingency Tables", avec χ^2, les effectifs attendus et "Log odds ratio" ; le RP dans MedCalc, "Tests" > "Relative risk". (e) HTA : 21/53 (39,6%) chez les fumeurs, 20/107 (18,7%) chez les autres ; plus petit effectif attendu 13,6 ; $\chi^2 = 8{,}15$, ddl 1, p = 0,004 ; RP = 2,12 (IC à 95% 1,26-3,55) ; OR = 2,85 (IC à 95% 1,37-5,95). (f) Nous rejetons H0. (g) La prévalence de l'HTA est environ 2 fois plus élevée chez les fumeurs actuels ; l'OR surestime cette relation, car l'HTA est fréquente (25,6% dans l'échantillon).

Consigne 7. (a) La PAS a changé dans le bras C. (b) H0 : µd = 0 (la moyenne des différences tas0 - tas6 est égale à 0) ; H1 : µd ≠ 0. (c) Données appariées (les mêmes patients, deux temps de mesure), n = 54 ; la normalité des différences n'est pas rejetée (Shapiro-Wilk p = 0,196) : test t apparié. (d) Filtrez le bras C dans JASP (icône en forme d'entonnoir, brat = "C"), puis "T-Tests" > "Paired Samples T-Test", tas0 et tas6, avec "Normality". (*e*) PAS $129{,}1 \pm 14{,}0$ mmHg à l'inclusion et 122,1 ± 16,8 mmHg à 6 mois ; différence moyenne de 6,96 mmHg (ÉT 8,18 ; IC à 95% 4,73-9,20) ; t = 6,26, ddl 53, p < 0,001 ; dz = 0,85 ; Wilcoxon, à titre de vérification, p < 0,001. (f) Nous rejetons H0. (g) Dans le bras C, la PAS a baissé en moyenne d'environ 7 mmHg en 6 mois. Attention à l'interprétation : une comparaison avant-après dans un seul bras ne montre pas l'effet de l'intervention par rapport au contrôle ; pour cela, on compare les baisses entre les bras (dans mg2026, l'ANOVA sur la baisse de la PAS donne p = 0,056).

Consigne 8. (c) Test diagnostique par rapport au standard de référence HGPO, seuil fixé. (d) Créez la variable test = glicemie0 ≥ 110 et construisez le tableau test x diabet_ogtt. (e) Le tableau :

Tableau L14.1. La glycémie à jeun ≥ 110 mg/dL par rapport au diabète confirmé par HGPO (mg2026, n = 160).

	Diabète HGPO oui	**Diabète HGPO non**	**Total**
Glycémie ≥ 110	24	39	63
Glycémie < 110	5	92	97
Total	29	131	160

$\text{Se} = \frac{24}{29} = 82{,}8\%$; $\text{Sp} = \frac{92}{131} = 70{,}2\%$; $\text{VPP} = \frac{24}{63} = 38{,}1\%$; $\text{VPN} = \frac{92}{97} = 94{,}8\%$; exactitude = (24 + 92)/160 = 72,5% ; taux d'erreur = (39 + 5)/160 = 27,5% ; $\text{RV}^+ = \frac{0{,}828}{0{,}298} = 2{,}78$; $\text{RV}^- = \frac{0{,}172}{0{,}702} = 0{,}25$. (g) À ce seuil, la glycémie à jeun manque environ un patient diabétique sur six et donne un résultat faussement positif chez près de 30% des non-diabétiques. Un résultat négatif rend le diabète peu probable (VPN 94,8%), un résultat positif n'élève la probabilité qu'à 38%. La performance globale de la glycémie (AUC 0,844) est proche de celle du score FINDRISC (AUC 0,852).

L14.4 Sujet type 2 (un petit jeu de données nouveau)

Un médecin généraliste teste un programme de 8 semaines de réduction du sodium alimentaire chez 12 patients hypertendus. Il mesure la PAS avant et après le programme et note l'IMC au début du programme.

Tableau L14.2. Les données du sujet 2 (PAS en mmHg, IMC en kg/m²).

Patient	**PAS avant**	**PAS après**	**IMC**
1	152	140	29,4
2	148	141	30,8
3	160	150	33,5
4	145	140	26,1
5	158	145	28,9

Patient	PAS avant	PAS après	IMC
6	150	142	31,6
7	165	150	32,2
8	142	138	27,5
9	155	146	34,0
10	149	143	26,8
11	162	151	30,1
12	147	145	29,3

Consignes (α = 0,05) :

1. Décrivez la PAS avant le programme (moyenne, ÉT, médiane, minimum, maximum), avec le calcul à la main de la moyenne et de l'écart-type.
2. Estimez la moyenne de la PAS avant le programme dans la population des patients hypertendus similaires, avec un niveau de confiance de 95% (t(11 ; 0,975) = 2,201).
3. La PAS a-t-elle baissé après le programme ? Appliquez le test approprié, avec le calcul à la main de la statistique du test.
4. L'IMC est-il corrélé à la PAS avant le programme ? Calculez r à la main, testez-le (t(10 ; 0,975) = 2,228) et écrivez la droite de régression.
5. Dans une cohorte de 400 patients hypertendus suivis pendant 5 ans, 150 sont fumeurs. Un AVC est survenu chez 18 fumeurs et 10 non-fumeurs. Calculez le RR avec son IC à 95%, l'OR, la différence de risques, le NNH et la fraction attribuable chez les exposés ; testez l'association.
6. Un test rapide de l'hypertrophie ventriculaire gauche a une Se de 90% et une Sp de 80%. Il est appliqué à 1000 patients hypertendus, avec une prévalence de l'hypertrophie de 15%. Reconstituez le tableau 2 x 2, calculez la VPP, la VPN, l'exactitude, le taux d'erreur, le RV+ et le RV-. Quelle serait la VPP pour une prévalence de 2% ?
7. Quel test choisissez-vous pour : a) comparer la PAS après le programme entre les hommes et les femmes (6 et 6 patients, distributions asymétriques) ; b) comparer la proportion de patients ayant une PAS < 140 mmHg avant et après le programme chez les mêmes patients ; c) comparer la satisfaction (1-5) entre trois programmes différents ; d) suivre l'évolution de la PAS à 0, 4 et 8 semaines chez les mêmes patients, avec des distributions normales ?

L14.5 Solution du sujet 2

Consigne 1. La somme des valeurs de PAS avant est 1833, donc $m = \frac{1833}{12} = 152{,}75$ mmHg. La somme des carrés des écarts à la moyenne est 574,25, donc $s^2 = \frac{574{,}25}{11} = 52{,}20$ et s = 7,23 mmHg. Les valeurs ordonnées : 142, 145, 147, 148, 149, 150, 152, 155, 158, 160, 162, 165 ; la médiane est la moyenne des valeurs en 6e et 7e positions : (150 + 152)/2 = 151 mmHg ; minimum 142, maximum 165. Shapiro-Wilk p = 0,805 : nous ne rejetons pas la normalité ; la moyenne ± ÉT est donc une description appropriée : 152,75 ± 7,23 mmHg.

Consigne 2. $\text{ES} = \frac{7{,}23}{\sqrt{12}} = 2{,}086$; $\text{IC}\,95\% = 152{,}75 \pm 2{,}201 \times 2{,}086 = 152{,}75 \pm 4{,}59$, soit 148,16-157,34 mmHg. Avec un échantillon de 12 patients, nous utilisons la loi t à 11 degrés de liberté, et non la valeur 1,96.

Consigne 3. (a) Le programme fait baisser la PAS. (b) H0 : μd = 0 ; H1 : μd ≠ 0 (ou, en unilatéral, μd > 0, si cela a été décidé à l'avance). (c) Données appariées, n = 12 ; les différences (avant - après) sont 12, 7, 10, 5, 13, 8, 15, 4, 9, 6, 11, 2 ; Shapiro-Wilk sur les différences p > 0,99 : test t apparié. (d) Moyenne des différences $\text{md} = \frac{102}{12} = 8{,}5$ mmHg ; la somme des carrés des écarts des différences est 167, donc $\text{sd} = \sqrt{\frac{167}{11}} = \sqrt{15{,}18} = 3{,}90$; $\text{ES} = \frac{3{,}90}{\sqrt{12}} = 1{,}125$; $t =$

$\frac{8,5}{1,125} = 7,56$, ddl 11. La valeur critique bilatérale est 2,201 ; 7,56 est bien supérieur, donc p < 0,001 (le logiciel donne p = 0,00001). IC à 95% pour md : $8,5 \pm 2,201 \times 1,125 = 8,5 \pm 2,48$, soit 6,0-11,0 mmHg ; dz $= \frac{8,5}{3,90} = 2,18$. Le test de Wilcoxon donne, à titre de vérification, p < 0,001 (les 12 différences sont toutes positives). (f) Nous rejetons H0. (g) La PAS a baissé en moyenne de 8,5 mmHg (IC à 95% 6,0-11,0). En l'absence de groupe contrôle, la baisse ne peut pas être attribuée avec certitude au programme : la régression vers la moyenne et l'effet d'être suivi sont des explications alternatives.

Consigne 4. La moyenne de l'IMC est de 30,02 kg/m². Les sommes de travail : $S_{xy} = \sum(x - m_x)\left(y - m_y\right) = 121,85$; $S_{xx} = \sum(x - m_x)^2 = 70,06$; Syy = 574,25.

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \frac{121,85}{\sqrt{70,06 \times 574,25}} = \frac{121,85}{200,57} = 0,608 t = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,608 \times 3,162}{\sqrt{0,630}} = \frac{1,922}{0,794} = 2,42$$

Comme 2,42 > 2,228, nous rejetons H0 : ρ = 0 (p = 0,036). $R^2 = 0,369$. Pente $b = \frac{S_{xy}}{S_{xx}} = \frac{121,85}{70,06} = 1,739$; ordonnée à l'origine a = 152,75 - 1,739 x 30,02 = 100,54. La droite : PAS = 100,54 + 1,739 x IMC ; pour un IMC de 30, PAS prédite de 152,7 mmHg. La corrélation de Spearman (rho = 0,629, p = 0,028) mène à la même conclusion. Interprétation : corrélation positive modérée à bonne ; l'IMC explique environ 37% de la variabilité de la PAS dans ce petit groupe. Avec 12 patients, l'intervalle de confiance de la pente est très large (0,14-3,34 mmHg par kg/m²) ; l'estimation est donc imprécise.

Consigne 5. Le tableau : fumeurs 18 avec AVC, 132 sans ; non-fumeurs 10 avec AVC, 240 sans. Risques : 18/150 = 0,12 et 10/250 = 0,04.

$$\text{RR} = \frac{0,12}{0,04} = 3,0;\ \text{ES}(\ln \text{RR}) = \sqrt{\frac{1}{18} - \frac{1}{150} + \frac{1}{10} - \frac{1}{250}} = \sqrt{0,0556 - 0,0067 + 0,1000 - 0,0040} = \sqrt{0,1449} = 0,381;\ \text{IC}\,95\%\,\text{RR} = \exp(1,099 \pm 1,96 \times 0,381) = \exp(0,352) \text{ à } \exp(1,845) = 1,42 \text{ à } 6,33;\ \text{OR} = \frac{18 \times 240}{132 \times 10} = \frac{4320}{1320} = 3,27(;\ \text{IC}\,95\%\,1,47 - 7,30);\ \text{DR} = 0,12 - 0,04 = 0,08(;\ \text{IC}\,95\%\,0,023 - 0,137)$$; $\text{NNH} = \frac{1}{0,08} = 12,5$, soit 13 FA chez les exposés = (3 - 1)/3 = 0,667

Les effectifs attendus (le plus petit étant de 10,5) permettent le khi-deux : $\chi^2 = 9,22$, ddl 1, p = 0,002 (avec la correction de Yates 8,03, p = 0,005). Conclusion : dans cette cohorte, les fumeurs hypertendus ont un risque d'AVC 3 fois plus élevé ; pour 13 fumeurs, un AVC supplémentaire survient en 5 ans par rapport aux non-fumeurs ; si la relation est causale, deux tiers des AVC des fumeurs seraient attribuables au tabagisme. S'agissant d'un schéma de cohorte, le RR est la mesure appropriée ; l'OR (3,27) en est proche, car l'AVC est relativement rare.

Consigne 6. Malades : 150 ; sujets sains : 850. VP = 0,90 x 150 = 135 ; FN = 15 ; VN = 0,80 x 850 = 680 ; FP = 170.

Tableau L14.3. Le tableau reconstitué pour le test rapide (n = 1000, prévalence 15%).

	Hypertrophie oui	**Hypertrophie non**	**Total**
Test positif	135	170	305
Test négatif	15	680	695
Total	150	850	1000

$\text{VPP} = \frac{135}{305} = 44{,}3\%$; $\text{VPN} = \frac{680}{695} = 97{,}8\%$; exactitude = 815/1000 = 81,5% ; taux d'erreur = (170 + 15)/1000 = 18,5% ; $\text{RV}^{+} = \frac{0{,}90}{0{,}20} = 4{,}5$; $\text{RV}^{-} = \frac{0{,}10}{0{,}80} = 0{,}125$. Pour une prévalence de 2% (20 malades sur 1000) : VP 18, FP 196, donc $\text{VPP} = \frac{18}{214} = 8{,}4\%$, et la VPN monte à 99,7%. Interprétation : un résultat négatif exclut pratiquement l'hypertrophie ; un résultat positif doit être confirmé par échocardiographie, surtout lorsque la prévalence est faible.

Consigne 7. a) Le test de Mann-Whitney (deux petits groupes indépendants, distributions asymétriques). b) Le test de McNemar (proportions sur données appariées). c) Le test de Kruskal-Wallis (variable ordinale, trois groupes indépendants). d) L'ANOVA à mesures répétées (trois temps de mesure, les mêmes patients, distributions normales) ; l'alternative non paramétrique serait le test de Friedman.

L14.6 Le barème et les erreurs fréquentes

Chaque sujet est noté sur 100 points, répartis entre les consignes proportionnellement à leur difficulté. Au sein d'une consigne de test, la répartition indicative des points est la suivante.

Tableau L14.4. Répartition indicative des points pour une consigne de test statistique (10 points).

Élément	Ce qui est demandé	Points
(a)-(b) Les hypothèses	hypothèse de recherche ; H0 et H1 en mots et sous forme mathématique	2
(c) La méthode	le bon test, justifié par le type de la variable, le nombre de groupes, l'indépendance et la vérification des hypothèses d'application	3
(d)-(e) Le calcul et le résultat	le logiciel et le menu ; la statistique du test, les degrés de liberté, p, la taille d'effet ou l'IC ; le graphique demandé	3
(f) La conclusion statistique	"on rejette H0" ou "on ne rejette pas H0", avec le seuil	1
(g) La conclusion de recherche	l'interprétation médicale, avec le sens et l'ampleur de l'effet, sans exagération	1

Pour les consignes de description, on note le choix des indicateurs selon la forme de la distribution, les chiffres et le graphique. Pour les consignes d'estimation, la formule ou le menu, l'intervalle correct et son interprétation. Pour les consignes de risque et de test diagnostique, le tableau 2 x 2 correctement orienté, la mesure adaptée au schéma d'étude et l'interprétation.

Les erreurs qui coûtent le plus de points, d'après l'expérience des sessions précédentes :

- Un test pour groupes indépendants appliqué à des données appariées (ou l'inverse). C'est l'erreur la plus coûteuse, car elle invalide la consigne entière.
- L'absence de vérification de la normalité avant un test paramétrique, ou sa vérification sur la mauvaise variable (pour le test t apparié, la normalité se vérifie sur les différences, et non à chaque temps de mesure).
- "On accepte H0" au lieu de "on ne rejette pas H0" ; "le résultat prouve qu'il n'y a pas de différence".
- La confusion entre l'intervalle de confiance de la moyenne et l'intervalle des valeurs individuelles.
- Un OR rapporté dans une étude de cohorte alors que l'on demande un RR, ou un RR calculé dans une étude cas-témoins.

- Le tableau 2 x 2 mal orienté (les malades en lignes et le test en colonnes, puis les formules appliquées comme si c'était l'inverse), ce qui intervertit la Se et la VPP.
- Le taux d'erreur calculé comme (a + d)/N ; la formule correcte est (b + c)/N.
- Des conclusions causales tirées d'une corrélation ou d'une étude transversale.
- Des valeurs p écrites "p = 0,000" (on écrit p < 0,001) et des décimales en excès.
- Le seuil du score de MedCalc mal transcrit ("> 10" signifie "≥ 11" pour des valeurs entières).
- Un graphique inapproprié : diagramme circulaire pour une variable numérique, histogramme pour une variable nominale.

L14.7 Le plan de préparation sur deux semaines

Le plan suppose environ une heure et demie par jour, cinq jours par semaine, plus une simulation d'examen à la fin de chaque semaine.

Semaine 1, consolidation.

- Jour 1 : les types de variables, la statistique descriptive et les graphiques (chapitre 2, séances 1-2). Sur mg2026, décrivez toutes les variables et choisissez les indicateurs ; refaites la consigne 1 du sujet 1.
- Jour 2 : la loi normale, les tests de normalité, le score z, la transformation logarithmique (chapitre 3). Sur mg2026 : la PAS, l'IMC, les triglycérides avant et après transformation logarithmique.
- Jour 3 : les intervalles de confiance et la taille de l'échantillon (chapitre 4). Calculez à la main et avec le logiciel les IC pour la PAS, le HDL, la proportion de diabète (Wald et Wilson).
- Jour 4 : les tests t, l'ANOVA et les équivalents non paramétriques (chapitres 6-8). Refaites les consignes 3, 5 et 7 du sujet 1 et la consigne 3 du sujet 2, sans regarder la solution.
- Jour 5 : le khi-deux, le risque relatif, le rapport de cotes, le NST (chapitres 7 et 10). Les problèmes tabac-BPCO, écran-acuité visuelle, contraceptifs oraux-polyarthrite rhumatoïde des séances ; la consigne 6 du sujet 1 et la consigne 5 du sujet 2.
- Fin de semaine : le sujet type 1 en 60 minutes, chronométré, puis correction à l'aide du barème.

Semaine 2, intégration.

- Jour 6 : la corrélation et la régression (chapitre 9). Calculez r à la main pour le sujet 2 ; sur mg2026, IMC-PAS, âge-PAS, tour de taille-glycémie ; la régression multiple.
- Jour 7 : le raisonnement bayésien et les tests diagnostiques (chapitres 11-12). Reconstituez des tableaux 2 x 2 à partir de la Se, de la Sp et de la prévalence, ou à partir de la Se et de la VPP ; la courbe ROC du score FINDRISC dans MedCalc.
- Jour 8 : le choix du test (chapitre 8 et tableau 14.1). Prenez 20 questions parmi les exercices des chapitres et, pour chacune, écrivez seulement le test et sa justification, en 1 minute.
- Jour 9 : la rédaction (section 14.4). Réécrivez les conclusions de toutes les consignes résolues jusqu'ici au format (f)-(g), en vérifiant les formulations interdites.
- Jour 10 : révision sur les fichiers des séances (BD_Studenti, BPOC, Date_HbA1c), une question de chaque type.
- Fin de semaine : le sujet type 2 en 60 minutes, puis une variante personnelle : changez les variables du sujet 1 (par exemple, le cholestérol au lieu du HDL, le milieu de résidence au lieu du sexe) et résolvez-la.

L14.8 La liste de vérification des compétences pratiques dans les logiciels

Ne cochez chaque compétence qu'après l'avoir mise en œuvre seuls, sans la fiche de travail.

Google Sheets :

☐ importation d'un CSV, paramètres régionaux pour la Roumanie, séparateur ";" ;

- ☐ =AVERAGE, =MEDIAN, =STDEV, =MIN, =MAX, =QUARTILE, =COUNTIF, =COUNTBLANK ;
- ☐ IC de la moyenne avec =T.INV.2T et =SQRT ;
- ☐ tableau croisé dynamique pour les fréquences et les tableaux de contingence ;
- ☐ histogramme et boîte à moustaches depuis le menu "Insertion" > "Graphique".

JASP :

- ☐ ouverture d'un CSV et vérification du type de chaque variable (nominal, ordinal, scale) ;
- ☐ colonne calculée (différences, IMC, dichotomisation à un seuil) et filtrage sur un sous-groupe ;
- ☐ Descriptive Statistics avec Shapiro-Wilk, histogrammes, boîtes à moustaches, "Split" par groupe ;
- ☐ One Sample, Independent Samples (avec Welch et Mann-Whitney) et Paired Samples T-Test (avec Wilcoxon) ;
- ☐ ANOVA avec Levene, Tukey et eta² ; Kruskal-Wallis ; Repeated Measures ANOVA et Friedman ;
- ☐ Contingency Tables avec χ^2, effectifs attendus, pourcentages en ligne et log odds ratio ;
- ☐ Correlation (Pearson, Spearman) avec le nuage de points ; Linear Regression simple et multiple ; Logistic Regression avec OR.

MedCalc :

- ☐ Tests > Relative risk et Odds ratio à partir d'effectifs saisis directement ;
- ☐ Statistics > ROC curves > ROC curve analysis, avec le tableau des seuils et l'indice de Youden ;
- ☐ le calculateur de taille d'échantillon pour une moyenne et pour une proportion.

Epi Info :

- ☐ StatCalc > 2 x 2 tables (RR, OR, khi-deux, Fisher) ;
- ☐ StatCalc > Sample size pour une étude transversale.

Calcul à la main (avec une calculatrice de poche) :

- ☐ la moyenne, l'ÉT, l'ES, l'IC à 95% d'une moyenne et d'une proportion ;
- ☐ le t apparié à partir des différences ; r à partir des sommes de produits ;
- ☐ RR, OR, DR, NST avec IC ; Se, Sp, VPP, VPN, RV, taux d'erreur ; probabilité post-test par les cotes.

Exercices supplémentaires

1. Sur le fichier BD_Studenti, rédigez un sujet d'examen de 6 consignes sur le modèle du sujet 1 (une description, un IC, un test pour deux groupes indépendants, un test pour données appariées, une association entre variables catégorielles, une corrélation) et échangez-le avec un collègue ; chacun résout le sujet de l'autre et le corrige selon le barème de L14.6.
2. Sur le fichier BPOC, résolvez au format (a)-(g) : "La fréquence cardiaque se modifie-t-elle significativement après le traitement ?" et "Un âge supérieur à 50 ans est-il un facteur de risque de PAS initiale supérieure à 140 mmHg ?". Précisez le schéma d'étude de chaque question et pourquoi la mesure appropriée pour la seconde est le rapport de prévalences ou l'OR, et non le RR.
3. Sur le fichier Date_HbA1c, choisissez une variable asymétrique et une variable symétrique et rédigez leur description complète, avec un graphique.
4. Problème avec données dans l'énoncé. Un test a une Se de 80% et une VPP de 50% dans un échantillon de 400 personnes, avec une prévalence de 10%. Reconstituez le tableau. Solution : malades 40, VP 32, FN 8 ; d'après la VPP de 50%, VP + FP = 64, donc FP 32 ; VN = 360 - 32 = 328. Sp $= \frac{328}{360} = 91{,}1\%$; VPN $= \frac{328}{336} = 97{,}6\%$; exactitude (32 + 328)/400 = 90,0% ; taux d'erreur (32 + 8)/400 = 10,0% ; $RV^+ = \frac{0{,}80}{0{,}089} = 9{,}0$.

5. Problème avec données dans l'énoncé. Chez 10 patients, les différences de glycémie avant et après un traitement (mg/dL) sont : 15, 8, 22, -3, 10, 12, 5, 18, 9, 14. Calculez la moyenne des différences, l'ÉT, t et la conclusion (t(9 ; 0,975) = 2,262). Solution : md $= \frac{110}{10} = 11{,}0$; la somme des carrés des écarts à la moyenne est 442, sd $= \sqrt{\frac{442}{9}} = \sqrt{49{,}11} = 7{,}01$; ES $= \frac{7{,}01}{\sqrt{10}} = 2{,}22$; $t = \frac{11{,}0}{2{,}22} = 4{,}96 > 2{,}262$, donc nous rejetons H0 ($p < 0{,}001$) ; IC 95% $11{,}0 \pm 2{,}262 \times 2{,}22 = 6{,}0 - 16{,}0$ mg/dL. La normalité des différences n'est pas rejetée (Shapiro-Wilk p = 0,97).

Ce qu'il faut rendre et barème indicatif

Lors de la simulation en laboratoire, vous déposez sur Moodle le document contenant la solution du sujet type 1, au format (a)-(g), avec les graphiques insérés. L'enseignant le corrige selon le barème de L14.6 et vous indique, pour chaque consigne, quel élément manquait. À l'examen proprement dit, la note obtenue est convertie en note de l'examen pratique, qui représente 30% de la note finale.

Tableau L14.5. Répartition indicative des 100 points entre les consignes du sujet 1.

Consigne	Contenu	Points
1	description, choix des indicateurs, graphiques	12
2	IC de la moyenne et intervalle des valeurs individuelles	10
3	test pour deux groupes indépendants	14
4	corrélation et régression, prédiction	14
5	ANOVA avec post hoc	14
6	khi-deux et mesure d'association	12
7	test pour données appariées	12
8	test diagnostique, tableau 2 x 2	12
Total		100

Variante hors ligne

Si les logiciels ne peuvent pas être utilisés, l'examen peut se dérouler dans la variante "sur papier", avec un sujet de type 2 : le petit jeu de données est fourni dans l'énoncé, et les calculs se font avec une calculatrice de poche, à l'aide du tableau de synthèse des formules (tableau 14.2) et des tables de valeurs critiques t, χ^2 et z remises avec le sujet. Les graphiques sont esquissés à la main (histogramme, boîte à moustaches, nuage de points avec la droite de régression). Pour vous préparer, résolvez entièrement le sujet 2 à la main, puis vérifiez les chiffres dans JASP ou MedCalc, lorsque vous y avez accès.

Annexe A. Réponses aux questions d'autoévaluation et aux exercices proposés

Les réponses sont données par chapitre. Pour les questions à réponses multiples, toutes les options correctes sont indiquées ; pour les exercices de calcul, le résultat et la méthode sont indiqués, et la solution complète se refait sur le modèle des problèmes résolus du chapitre.

Chapitre 1

Questions : 1 b ; 2 c ; 3 c ; 4 b ; 5 c ; 6 a, b, d ; 7 b, c, d ; 8 a, b, d ; 9 a, b, d ; 10 a, c, d

Explications : Pour la question 3, le piège est l'option d (0,621), qui correspond à $P(\text{obèse} \mid \text{diabète}) = \frac{18}{29}$, et non à $P(\text{diabète} \mid \text{obèse}) = \frac{18}{54} = 0{,}333$; la condition indique le total auquel la cellule est rapportée. Pour la question 8, deux événements mutuellement exclusifs de probabilités non nulles ne peuvent pas être indépendants : si B s'est produit, la probabilité de A devient nulle, donc P(A|B) = 0 ≠ P(A) ; et P(A) + P(B) = P(A ou B) ne peut pas dépasser 1. Pour la question 7, un intervalle de confiance étroit indique une erreur aléatoire faible, mais ne dit rien de l'erreur systématique ; une très grande étude peut être très précise et néanmoins biaisée.

Exercices proposés : 1. $P(D \mid \text{urbain}) = \frac{19}{95} = 0{,}200$; $P(D \mid \text{rural}) = \frac{10}{65} = 0{,}154$; $P(\text{urbain et } D) = \frac{19}{160} = 0{,}119$; $P(\text{urbain}) \times P(D) = 0{,}594 \times 0{,}181 = 0{,}108$. Les valeurs sont proches, mais pas identiques ; la différence est faible et peut être due au hasard, de sorte que les événements sont approximativement indépendants dans l'échantillon (la décision formelle exige un test du khi-deux). 2. $P(\text{fumeur} \mid \text{femme}) = \frac{30}{86} = 0{,}349$; $P(\text{femme} \mid \text{fumeur}) = \frac{30}{53} = 0{,}566$; $P(\text{femme ou fumeur}) = \frac{86+53-30}{160} = \frac{109}{160} = 0{,}681$. Les deux premières diffèrent parce qu'elles sont rapportées à des totaux différents (respectivement 86 femmes et 53 fumeurs). 3. Loi binomiale n = 8, p = 0,25 : $(a)\ 0{,}75^8 = 0{,}100$; $(b)\ 8 \times 0{,}25 \times 0{,}75^7 = 0{,}267$; (c) 1 - 0,100 - 0,267 = 0,633. 4. $k = \frac{2\,400}{120} = 20$; les dossiers 9, 29, 49, 69, 89,..., le dernier étant 9 + 20 x 119 = 2 389. 5. Fraction de sondage 90/1 500 = 6% ; 600 x 0,06 = 36 (18-39 ans), 36 (40-64 ans), 300 x 0,06 = 18 (≥ 65 ans). 6. (a) biais d'information (de déclaration, de désirabilité sociale : la consommation est sous-déclarée) ; (b) confusion (l'activité physique est liée à la fois aux compléments alimentaires et au risque de fracture) ; (c) biais de sélection par perdus de vue (les sujets restants ne représentent plus le groupe randomisé) ; (d) biais de sélection de type spectre (patients à un stade avancé comparés à des volontaires sains ; la performance du test sera surestimée). 7. Unité statistique : l'ordonnance ; n = 1 498 ; population cible : les prescriptions (et, indirectement, les personnes âgées traitées) du milieu urbain de l'ouest de la Roumanie ; étude transversale ; variables numériques : l'âge, le nombre de médicaments par ordonnance ; variables qualitatives : le sexe, la présence d'un médicament potentiellement inapproprié (oui/non) ; si un patient a plusieurs ordonnances, les observations ne sont plus indépendantes, et les patients ayant de nombreuses ordonnances pèsent davantage dans les résultats. 8. Loi de Poisson λ = 1,5 : $P(X = 0) = e^{-1{,}5} = 0{,}223$; $P(X \geq 4) = 1 - P(X \leq 3) = 1 - 0{,}934 = 0{,}066$. 9. Probabilité totale : $P(\text{BPCO}) = 0{,}30 \times 0{,}15 + 0{,}70 \times 0{,}03 = 0{,}045 + 0{,}021 = 0{,}066 (6{,}6\%)$; Bayes : $P(\text{fumeur} \mid \text{BPCO}) = \frac{0{,}045}{0{,}066} = 0{,}682 (68{,}2\%)$. 10. (a) P : adultes à risque cardiométabolique suivis dans les cabinets de médecine générale ; I : conseil + application + téléconsultation mensuelle (bras C) ; C : conseil standard (bras A) ; O : variation de la pression systolique entre l'inclusion et 6 mois (tas0 - tas6) ; T : 6 mois. (b) Les données sont pseudonymisées (il existe une

table de correspondance qui permet la réidentification) ; elles restent des données à caractère personnel, relevant de la catégorie particulière des données concernant la santé, soumises au RGPD (GDPR) et à la loi roumaine 190/2018.

Chapitre 2

Questions : 1 b ; 2 c ; 3 b ; 4 b ; 5 b ; 6 a, b, c ; 7 a, b, d, e ; 8 a, b, c, e ; 9 a, b, c, e ; 10 a, b, d, e

Explications : Pour la question 2, l'écart-type (7,0) est supérieur à la moitié de la moyenne (11,6) pour une variable qui ne peut pas être négative ; la distribution est donc asymétrique à droite et la médiane décrit mieux la valeur typique. Pour la question 5, n = 8 est pair, et la médiane est la moyenne des valeurs en 4e et 5e positions, (4 + 5)/2 = 4,5. Pour la question 7, l'option c est fausse : l'écart-type décrit la variabilité des patients et ne diminue pas systématiquement lorsque n augmente ; seule l'erreur standard, $\frac{s}{\sqrt{n}}$, diminue. Pour la question 10, l'option c est fausse : une valeur extrême biologiquement plausible et vérifiée n'est pas éliminée ; la règle 1,5 x EIQ signale des valeurs à vérifier.

Exercices proposés : 1. Groupe sanguin : nominale ; score d'Apgar : ordinale (score discret 0-10) ; nombre de grossesses : numérique discrète, échelle de rapport ; température corporelle : numérique continue, échelle d'intervalle ; classe NYHA : ordinale ; HbA1c : numérique continue, échelle de rapport ; degré d'accord sur une échelle de Likert : ordinale ; code du patient : identifiant, non analysé ; durée d'hospitalisation : numérique (discrète lorsqu'on compte des jours entiers), échelle de rapport ; tabagisme oui/non : nominale dichotomique. 2. A = 0,35 x 40 = 14 ; B = 15,0 + 35,0 = 50,0% ; $C = \frac{10}{40} = 25{,}0\%$ (vérification 50,0 + 25,0 = 75,0) ; D = 0,125 x 40 = 5 ; E = 40 - 6 - 14 - 10 - 5 = 5 (12,5%) ; glycémie ≥ 126 mg/dL : 12,5 + 12,5 = 25,0% (10 patients). 3. $m = \frac{800}{6} = 133{,}33$ mmHg ; médiane (130 + 136)/2 = 133 mmHg ; EMA $= \frac{56}{6} = 9{,}33$ mmHg ; somme des carrés des écarts 693,33 ; $s^2 = \frac{693{,}33}{5} = 138{,}67$ mmHg^2 ; s = 11,78 mmHg ; CV = 8,8% ; ES $= \frac{11{,}78}{\sqrt{6}} = 4{,}81$ mmHg. 4. Moyenne 85/10 = 8,5 jours ; médiane (5 + 6)/2 = 5,5 jours ; deux modes, 4 et 5 (distribution bimodale dans l'échantillon) ; Q1 = 4, Q3 = 8 (médianes des deux moitiés ; les logiciels donnent 4,25 et 7,75), EIQ = 4 ; limites -2 et 14, donc 31 jours est une valeur extrême (12 ne l'est pas) ; sans elle, la moyenne devient 54/9 = 6,0 jours ; on rapporte la médiane 5,5 (Q1-Q3 4-8). 5. k = 1 + 3,322 x log10(250) = 1 + 3,322 x 2,398 = 8,97, soit 9 classes ; $h = \frac{412-68}{9} = 38{,}2$ mg/dL, arrondi à 40 mg/dL (par exemple, classes 60-99, 100-139,..., 400-439). 6. Si 3 est l'ES, ÉT = 3 x √64 = 24 mg/dL (CV 21%, plausible) ; si 3 était l'ÉT, le CV serait de 2,7%, une homogénéité biologiquement invraisemblable ; c'est donc l'interprétation comme ES qui est plausible ; ES de la PAS chez les femmes = 14,47/√86 = 1,56 mmHg. 7. Moyenne pondérée = $\frac{120\times45{,}2+60\times61{,}8+20\times70{,}4}{200} = \frac{10\,540}{200} = 52{,}7$ ans ; la moyenne simple des moyennes, 59,13 ans, surestime l'âge parce qu'elle donne au petit service (20 patients) le même poids qu'au grand service. 8. $\frac{(a)\,26}{53} = 49{,}1\%$; (b) l'IMC, les triglycérides et le score FINDRISC sont rapportés en médiane (Q1-Q3) parce qu'ils sont asymétriques ou qu'il s'agit de scores ; (c) après randomisation, les différences initiales entre les bras sont par définition dues au hasard ; un test vérifierait donc une hypothèse nulle que nous savons vraie ; on juge l'ampleur clinique des déséquilibres, et non la valeur p. 9. (a) diagramme en barres (un diagramme circulaire à 4 catégories est acceptable) ; (b) boîtes à moustaches juxtaposées par bras ; (c) nuage de points ; (d) graphique linéaire ; (e) histogramme ; (f) barres empilées à 100%, une pour chaque sexe, avec les catégories dans l'ordre de l'échelle. 10. Moyenne arithmétique = 1 111 000/4 = 277 750 copies/mL ; les logarithmes décimaux sont 3, 4, 5, 6, de moyenne 4,5 ; la moyenne géométrique est donc 10 puissance 4,5 = 31 623 copies/mL ; la charge virale varie de façon multiplicative (distribution approximativement

log-normale), et sur une échelle logarithmique une baisse de "1 log" signifie une réduction d'un facteur 10, facile à comparer d'un patient à l'autre.

Chapitre 3

Questions : 1 c ; 2 b ; 3 b ; 4 b ; 5 c ; 6 a, b, c, e ; 7 b, c, e ; 8 a, b, c, d ; 9 b, c, d, e ; 10 a, b, d, e

Explications : Pour la question 4, m - 2s = 11,6 - 14,0 = -2,4 jours ; une variable qui ne peut pas être négative et dont m - 2s est inférieur à zéro ne peut pas être normale, et les durées d'hospitalisation ont presque toujours une longue queue à droite. Pour la question 7, un p de 0,30 signifie seulement qu'on ne rejette pas l'hypothèse de normalité ; avec 12 observations, le test a une faible puissance, il ne "prouve" donc rien et n'oblige pas à utiliser un test paramétrique. Pour la question 10, l'option c est fausse : par construction, 5% des personnes en bonne santé ont des valeurs en dehors de l'intervalle, et 1 - 0,95^20 = 0,64 explique l'option e.

Exercices proposés :

1. z(125) = (125 - 131,13) / 13,88 = -0,44, Φ = 0,3300, 33e percentile ; z(150) = 1,36, Φ = 0,9131, 91e percentile. Dans Sheets : =NORM.S.DIST(-0,44; TRUE) et =NORM.S.DIST(1,36; TRUE).
2. z(240) = 0,83, proportion théorique 1 - 0,7967 = 20,3% (observée 23,8%) ; $z(190) = -0{,}53$, proportion théorique 29,8% (observée 30,0%). Bonne adéquation ; la queue droite est un peu plus épaisse que ne le prédit le modèle, une différence de l'ordre des fluctuations d'échantillonnage.
3. $49{,}52 \pm 1{,}96 \times 11{,}78 = 49{,}52 \pm 23{,}09$, soit 26,4-72,6 mg/dL ; théoriquement 95% (observé 94,4%). HDL < 40 : z = -0,81, Φ = 0,2090, soit 20,9% (observé 19,4%).
4. Loi binomiale n = 6, p = 0,33 : moyenne 1,98 ; variance 6 x 0,33 x 0,67 = 1,33 ; $P(0) = 0{,}67^6 = 0{,}0905$; $P(1) = 6 \times 0{,}33 \times 0{,}67^5 = 0{,}2673$; $P(X \geq 2) = 1 - 0{,}0905 - 0{,}2673 = 0{,}6422$.
5. Loi de Poisson λ = 0,5 : $P(X \geq 1) = 1 - e^{-0{,}5} = 1 - 0{,}6065 = 0{,}3935$; $P(X \geq 3) = 1 - (0{,}6065 + 0{,}3033 + 0{,}0758) = 0{,}0144$.
6. $4{,}3 \pm 1{,}96 \times 0{,}35 = 4{,}3 \pm 0{,}69$, soit 3,61-4,99 mmol/L ; $z = \frac{5{,}1-4{,}3}{0{,}35} = 2{,}29$, Φ = 0,9890, 98,9e percentile ; la valeur est en dehors de l'intervalle, et le dosage est répété (l'hémolyse de l'échantillon peut augmenter faussement la kaliémie) avant toute décision.
7. m - 2s = 9,1 - 11,6 = -2,5 jours : la distribution est asymétrique à droite ; il aurait été plus approprié d'utiliser la médiane et l'EIQ (éventuellement le minimum et le maximum), et de faire la comparaison avec un test non paramétrique ou sur une échelle logarithmique.
8. 0,39 / 0,192 = 2,03 et 0,74 / 0,381 = 1,94 : écart limite, confirmé par Shapiro-Wilk p = 0,034 ; le score est discret, avec de nombreuses valeurs égales. On rapporte la médiane (EIQ), éventuellement aussi la moyenne ± ÉT ; pour des comparaisons avec n = 160, le test t est robuste, mais un test non paramétrique (Mann-Whitney, Kruskal-Wallis) est un choix sûr ; idéalement, on indique que les deux mènent à la même conclusion.
9. Moyenne géométrique $e^{4{,}97} = 144{,}0$; facteur $e^{0{,}49} = 1{,}63$; intervalle à 95% : de $e^{4{,}97-0{,}96} = e^{4{,}01} = 55{,}1$ à $e^{4{,}97+0{,}96} = e^{5{,}93} = 376{,}3$ (environ 55-376, asymétrique par rapport à 144).
10. EIQ = 23 ; les barrières 93 - 34,5 = 58,5 et 116 + 34,5 = 150,5, donc 154 est une valeur extrême selon la règle de Tukey ; $z = \frac{154-105{,}44}{15{,}76} = 3{,}08$, donc aussi selon le critère z. La valeur est plausible dans une population comportant des diabètes non diagnostiqués : on la vérifie sur le compte rendu de laboratoire, on la conserve et on effectue une analyse de sensibilité.
11.

    (a) points sur la droite ; (b) courbe en forme de "U" : points au-dessus de la droite aux deux extrémités, avec une montée abrupte à droite (asymétrie positive) ; (c) forme de "S" couché : points sous la droite à gauche et au-dessus de la droite à droite ; (d) points sur la droite, avec un seul point nettement au-dessus de la droite dans le coin supérieur droit.

Chapitre 4

Questions : 1 b ; 2 c ; 3 b ; 4 d ; 5 c ; 6 a, c, d, e ; 7 a, b, c, e ; 8 a, b, c, d ; 9 a, b, c, e ; 10 a, b, c, d

Explications : Pour la question 3, l'intervalle 0,90-2,18 inclut la valeur nulle 1, donc p ≥ 0,05 ; cela ne prouve pas l'absence d'association (option c), mais montre que les données sont compatibles aussi bien avec une diminution qu'avec un doublement de la cote. Pour la question 4, on applique la règle de trois : 3 / 60 = 0,05, soit 5%. Pour la question 9, l'option d est fausse : un seuil α plus permissif (0,10) diminue la valeur critique et donc l'échantillon nécessaire, au prix d'un plus grand nombre de résultats faussement positifs. Pour la question 10, l'option e confond la précision avec la taille de l'effet : un intervalle large indique une estimation imprécise.

Exercices proposés :

1. ES = $\frac{36,91}{\sqrt{160}} = 2,918$ mg/dL ; marge 1,975 x 2,918 = 5,76 ; IC à 95% 203,63-215,15 mg/dL.
2. ES = $\frac{11,78}{\sqrt{160}} = 0,931$ mg/dL. Valeurs individuelles : 68,3% 37,74-61,30 ; 95,4% 25,96-73,08 ; 99,7% 14,18-84,86 mg/dL. Moyenne : 68,3% 48,59-50,45 ; 95,4% 47,66-51,38 ; 99,7% 46,73-52,31 mg/dL.
3. ES = $\frac{1,5}{\sqrt{25}} = 0,30\%$; t(24 ; 0,975) = 2,064 (n petit, σ estimé à partir de l'échantillon) ; marge 0,62 ; IC à 95% 6,58-7,82%.
4. $p = \frac{41}{160} = 25,6\%$; ES = 0,0345 ; Wald 18,9-32,4% ; Wilson 19,5-32,9%.
5. Règle de trois : 3 / 120 = 2,5% (la méthode exacte donne 3,0%).
6. Différence de 8 mmHg ; ES = $\sqrt{\frac{15^2}{40} + \frac{14^2}{45}} = \sqrt{5,625 + 4,356} = 3,16$ mmHg ; IC à 95% d'environ 8 ± 6,19, soit 1,8-14,2 mmHg (avec le t de Welch, environ 80 ddl, 1,7-14,3). L'intervalle exclut 0 ; la différence est donc significative au seuil de 0,05, mais l'estimation est imprécise.
7. OR = $\frac{20\times40}{30\times10} = 2,67$; ln OR = 0,981 ; ES = $\sqrt{\frac{1}{20} + \frac{1}{30} + \frac{1}{10} + \frac{1}{40}} = 0,456$; 0,981 ± 0,895 donne 0,086-1,876 ; IC à 95% 1,09-6,52. Il exclut 1 (association significative), mais la borne inférieure est proche de 1.
8. (a) $n = \left(\frac{1,96\times10}{2}\right)^2 = 96,04$, donc 97 ; $(b)\, n = \frac{1,96^2\times0,25}{0,04^2} = 600,25$, donc 601.
9. $n = \frac{2\times(1,96+0,84)^2}{0,4^2} = \frac{15,68}{0,16} = 98,0$, donc 99 par groupe (G*Power : 100) ; avec 20% de perdus de vue, 99 / 0,8 = 123,75, donc 124 par groupe (125 en partant de 100).
10. La valeur 1,1 est presque certainement l'erreur standard ; ÉT = $1,1 \times \sqrt{160} = 13,9$ mmHg. OR 1,75 (IC à 95% 0,95-3,10) : l'intervalle inclut 1 ; l'association n'est donc pas significative au seuil de 0,05.
11. La largeur diminue avec $\sqrt{n}$: pour la réduire de moitié, il faut multiplier n par 4, soit environ 160 patients.

Chapitre 5

Questions : 1 c ; 2 b ; 3 b ; 4 d ; 5 d ; 6 a, b, c, e ; 7 a, b, d ; 8 b, c, e ; 9 a, b, d ; 10 a, b, d, e

Explications : Pour la question 3, la valeur 132 mmHg se trouve dans l'IC à 95% (128,96-133,30) ; le test bilatéral à α = 0,05 ne rejette donc pas H0 ; l'option c est fausse parce que "accepter H0" n'est pas une conclusion permise, et la moyenne de la population n'est pas connue avec certitude. Pour la question 7, passer d'un test unilatéral justifié à un test bilatéral répartit α sur deux queues et diminue la puissance ; l'augmentation de l'écart-type accroît le bruit et diminue la

puissance. Pour la question 8, l'option d généralise abusivement : un résultat non significatif ne démontre pas l'absence d'effet, et l'intervalle de confiance montre seulement quelles différences sont exclues dans la population étudiée.

Exercices proposés : 1. Test t pour une moyenne : ES $= \frac{0,45}{\sqrt{160}} = 0,0356$; $t = \frac{5,58-5,7}{0,0356} = -3,37$ (avec la moyenne exacte 5,5775, t = -3,48), ddl = 159, p < 0,001 ; on rejette H0. IC à 95% de la moyenne 5,51-5,65%. La moyenne de l'HbA1c dans la population étudiée est significativement inférieure à 5,7%. 2. ES $= \frac{4,27}{12,649} = 0,338$; $t = \frac{10,37-12}{0,338} = -4,83$, p < 0,001 ; IC à 95% 9,70-11,04 ; $d = -\frac{1,63}{4,27} = -0,38$ (effet faible à moyen). On rejette H0 ; la moyenne du FINDRISC est inférieure à 12. 3. ES $= \frac{9,30}{12,649} = 0,735$; $t = \frac{2,71}{0,735} = 3,68$, p < 0,001 (environ 0,0003) ; IC à 95% 73,26-76,16 battements/min ; d = 0,29. Résultat extrêmement significatif sur le plan statistique, mais une différence d'environ 3 battements/min par rapport à 72 n'a pas de pertinence clinique. 4. Test z : ES $= \frac{1,2}{\sqrt{36}} = 0,2$ g/dL ; $z = \frac{11,2-11,8}{0,2} = -3,0$; |z| > 1,96 (et > 2,576) ; p = 0,0027. On rejette H0, différence très significative : l'hémoglobine moyenne des enfants de la communauté est inférieure à la moyenne de la population générale. 5. (a) erreur de type I (faux positif) ; (b) erreur de type II (faux négatif), favorisée par le petit échantillon ; (c) décision correcte (H0 vraie, non rejetée). 6. $(a)\ 1 - 0,95^8 = 0,337$; $\frac{(b)\ 0,05}{8} = 0,00625$; $(c)\ 1 - 0,95^4 = 0,185$, seuil de Bonferroni 0,05 / 4 = 0,0125. 7. Sans correction : 4 (0,002 ; 0,009 ; 0,018 ; 0,035). Bonferroni (seuil 0,01) : 2. Holm : 0,002 < 0,01 ; 0,009 < 0,0125 ; 0,018 > 0,0167, arrêt : 2. Benjamini-Hochberg (seuils 0,01 ; 0,02 ; 0,03 ; 0,04 ; 0,05) : 0,035 ≤ 0,04, donc 4. 8. (a) T.DIST.2T(2,5;12) = 0,028 (entre les valeurs critiques 2,179 et 3,055) ; (b) bilatéral 0,097, unilatéral 0,048 ; (c) T.DIST.2T(2,9;159) = 0,004 (la fonction reçoit la valeur absolue de t). 9. (a) L'IC à 95% contient 0, donc p > 0,05. (b) Effet indéterminé : l'intervalle englobe à la fois 0 et des différences supérieures au seuil clinique de 5 mmHg (jusqu'à 9 mmHg). (c) Une étude plus grande, dimensionnée pour une différence de 5 mmHg (environ 123-125 patients par groupe pour σ = 14, puissance de 80%). 10. (a) ES $= \frac{3,91}{12,649} = 0,309$; $t = \frac{29,04-30}{0,309} = -3,11$, p = 0,002. (b) La distribution de l'IMC est asymétrique (1,42) et non normale ; on préfère le test des rangs signés de Wilcoxon pour la médiane par rapport à 30 (p < 0,001) ou l'analyse sur données log-transformées. (c) Médiane 28,25 kg/m² (EIQ 26,2-31,4), test de Wilcoxon, p < 0,001 : l'IMC typique des patients est inférieur au seuil de 30 kg/m². 11. $n = \frac{2\times(1,96+0,84)^2\times14^2}{10^2} = \frac{2\times7,84\times196}{100} = 30,7$, soit 31 patients par groupe (le calcul exact avec la loi t donne 32) ; $d = \frac{10}{14} = 0,71$.

Chapitre 6

Questions : 1 b ; 2 c ; 3 b ; 4 c ; 5 b ; 6 a, b, c, e ; 7 a, b, c ; 8 a, b, d ; 9 a, b, c ; 10 a, b, c, e

Explications : Pour la question 2, les degrés de liberté du test apparié se calculent à partir du nombre de paires (25 - 1 = 24), et non du nombre total de mesures. Pour la question 8, l'option c est fausse : l'ANOVA montre seulement qu'au moins une moyenne diffère ; la paire C - B est identifiée par le test post hoc. L'option e est fausse parce que $\eta^2 = 0,226$ indique un effet important, qui ne dépend pas de n. Pour la question 10, l'alternative non paramétrique pour les mesures répétées est le test de Friedman ; le test de Kruskal-Wallis s'applique aux groupes indépendants.

Exercices proposés : 1. Les différences : 0,6 ; 0,3 ; 0,1 ; 0,9 ; 0,6 ; 0,3 ; 0,8 ; 0,4 ; mD = 0,50% ; somme des carrés des écarts 0,52 ; $s_D = \sqrt{\frac{0,52}{7}} = 0,273$; ESD = 0,096 ; t = 5,19, ddl = 7,

p = 0,0013 (t critique 2,365) ; IC à 95% 0,27-0,73% ; dz = 1,83. On rejette H0 : l'HbA1c a baissé de façon extrêmement significative, en moyenne de 0,5 point de pourcentage (étude sans groupe contrôle). 2. $s^2_{\text{commune}} = \frac{24\times144+29\times196}{53} = 172{,}45$; scommun = 13,13 ; ES = $13{,}13 \times \sqrt{\frac{1}{25}+\frac{1}{30}} = 3{,}56$; $t = \frac{7}{3{,}56} = 1{,}97$, ddl = 53, p = 0,054 (t critique 2,006) ; IC à 95% -0,13 à 14,13 mmHg ; d = 0,53. On ne rejette pas H0 (à la limite) ; l'intervalle englobe à la fois 0 et des différences importantes ; le résultat est donc indéterminé, et l'étude manque de puissance pour une différence de 7 mmHg. 3. Student : $s^2_{\text{commune}} = \frac{11\times1{,}21+39\times10{,}24}{50} = 8{,}25$; ES = 0,946 ; t = -1,37, ddl = 50, p = 0,175. Welch : ES = $\sqrt{\frac{1{,}21}{12}+\frac{10{,}24}{40}} = 0{,}597$; t = -2,18, ddl = 48,9, p = 0,034. La variance commune est dominée par le grand groupe, dont la variance est beaucoup plus grande (rapport des ÉT proche de 3) ; pour le petit groupe, de faible variance, l'erreur standard est surestimée, et le test de Student devient trop conservateur. Le test de Welch est le bon ; conclusion : différence significative. 4. M = 126,33 ; SS inter = $10 \times [(120-126{,}33)^2 + (128-126{,}33)^2 + (131-126{,}33)^2] = 646{,}7 (\text{ddl } 2,\ \text{MS } 323{,}3)$; SS intra = 9 × (100 + 144 + 121) = 3285 (ddl 27, MS 121,7) ; F = 2,66 < 3,35 ; p = 0,088 ; $\eta^2 = 0{,}16$. On ne rejette pas H0. 5. ddl inter = 3, ddl intra = 36, ddl total = 39 ; MS inter = 120, MS intra = 40 ; F = 3,0 > 2,87 ; p = 0,043 ; $\eta^2 = \frac{360}{1800} = 0{,}20$. On rejette H0 (différence significative) ; les tests post hoc suivent. 6. Welch : ES = $\sqrt{\frac{6{,}49^2}{52}+\frac{4{,}60^2}{84}} = \sqrt{0{,}810+0{,}252} = 1{,}030$; $t = \frac{3{,}68}{1{,}030} = 3{,}57$, ddl ≈ 82,7, p ≈ 0,0006. scommun = 5,40, d ≈ 0,68 (effet moyen à important). La variante de Student donnerait t ≈ 3,86. Les deux valeurs sont compatibles avec le p = 0,001 rapporté ; les différences peuvent provenir du test utilisé (éventuellement non paramétrique) ou des arrondis. 7. Welch : ES = $\sqrt{\frac{7{,}0^2}{224}+\frac{5{,}8^2}{62}} = \sqrt{0{,}219+0{,}543} = 0{,}873$; $t = \frac{2{,}5}{0{,}873} = 2{,}87$, ddl ≈ 115, p ≈ 0,005 (calcul didactique, et non résultat rapporté par les auteurs). L'ÉT représente environ 60% de la moyenne, et la moyenne moins 2 ÉT serait négative : la distribution de la durée d'hospitalisation est asymétrique à droite. Sur les données individuelles, on préfère la médiane avec l'EIQ et le test de Mann-Whitney, ou l'analyse sur données log-transformées. 8. ESD = $\frac{9{,}05}{12{,}649} = 0{,}716$; $t = \frac{4{,}78}{0{,}716} = 6{,}68$, ddl = 159, p < 0,001 ; IC à 95% 3,36-6,19 mmHg ; dz = 0,53. La baisse moyenne d'environ 5 mmHg est pertinente au niveau de la population, mais l'analyse sur l'ensemble de l'échantillon n'a pas de groupe contrôle ; l'effet des interventions s'évalue en comparant les bras. 9. (a) t apparié / test des rangs signés de Wilcoxon ; (b) t pour échantillons indépendants (Student ou Welch) / Mann-Whitney ; (c) ANOVA à mesures répétées / Friedman ; (d) t pour échantillons indépendants / Mann-Whitney ; (e) ANOVA à un facteur / Kruskal-Wallis ; (f) t pour une seule moyenne / Wilcoxon pour une médiane. 10. HSD = $3{,}51 \times \sqrt{\frac{121{,}7}{10}} = 3{,}51 \times 3{,}49 = 12{,}2$ mmHg. Les différences : 8, 11 et 3 mmHg, toutes inférieures à 12,2 ; aucune paire ne diffère, ce qui est cohérent avec un F non significatif. 11. Le test apparié utilise l'écart-type des différences individuelles, qui dépend de la corrélation entre les deux mesures et ne peut pas être déduit des ÉT à chaque temps de mesure. Il faudrait : le nombre de patients (non indiqué dans le résumé) et la différence moyenne avec l'ÉT des différences ou avec l'IC à 95% (ou la médiane des différences avec l'EIQ, si le test de Wilcoxon a été utilisé).

Chapitre 7

Questions : 1 c ; 2 b ; 3 d ; 4 a ; 5 c ; 6 a, c, e ; 7 a, c, d ; 8 a, b, c, e ; 9 a, c, d, e ; 10 a, b, d, e Explications : pour la question 2, les trois valeurs 7 occupent les positions 3, 4 et 5 et reçoivent chacune le rang moyen 4, et la somme des rangs (21) confirme le calcul. Pour la question 7, un effectif observé de 0 n'est pas un problème si les effectifs attendus sont grands (b), et des groupes inégaux n'invalident pas le test (e) ; ce qui compte, ce sont les effectifs attendus, l'indépendance des observations et l'utilisation des nombres absolus. Pour la question 8, la satisfaction n'est pas une mesure du contrôle glycémique (d). Pour la question 10, une petite valeur p ne mesure pas la force de l'association (c) ; la force se lit dans φ = 0,28 ou dans le rapport de prévalences. Exercices proposés : 1. Les rangs : 4 ; 6 ; 6 ; 1,5 ; 8 ; 6 ; 3 ; 1,5 ; somme $36 = \frac{8\times9}{2}$. 2. R1 = 11, R2 = 34 ; U1 = 11 - 10 = 1, U2 = 34 - 15 = 19 ; U = 1 ≤ 1, donc on rejette H0 (p exact 0,032) ; l'anxiété est plus faible dans le groupe ayant reçu l'information vidéo (médiane 4 contre 8). 3. Les différences : 8, 2, -2, 12, 6, 0, 7, 9 ; le patient 6 a une différence de 0 et est éliminé, n = 7 ; rangs des valeurs absolues : 2 et 2 → 1,5 ; 6 → 3 ; 7 → 4 ; 8 → 5 ; 9 → 6 ; 12 → 7 ; W+ = 26,5, W- = 1,5 ; T = 1,5 ≤ 2, on rejette H0 (p exact 0,047), la pression a baissé (médiane des différences 7 mmHg). 4. Somme $40 + 60 + 71 = 171 = \frac{18\times19}{2}$; $H = \frac{\left[\frac{12}{18\times19}\right]\times\left(40^2+60^2+71^2\right)}{6} - 57 = 2{,}89 <$ 5,99, on ne rejette pas H0 (p = 0,24). 5. $\chi^2 F = \left[\frac{12}{10\times3\times4}\right] \times (14^2 + 20^2 + 26^2) - 120 = 7{,}2 >$ 5,99, on rejette H0 (p = 0,027) ; $W = \frac{7{,}2}{10\times2} = 0{,}36$. 6. Vaccinés : 10 avec grippe / 190 sans ; non vaccinés : 24 / 126 ; attendus 19,43 / 180,57 / 14,57 / 135,43 ; χ^2 = 11,83, p < 0,001 (Yates 10,60, p = 0,001) ; RR $= \frac{5{,}0\%}{16{,}0\%} = 0{,}31$ (la vaccination réduit le risque d'environ 69% dans cet échantillon) ; φ = 0,18. 7. N = 70 ; totaux des colonnes 20, 36, 14 ; attendus sur la ligne 3 : 2,57 ; 4,63 ; 1,80 (la cellule ligne 1, colonne 3 vaut exactement 5,0) ; 3 cellules sur 9 (33%) ont E < 5, au-delà de la limite de 20% ; le khi-deux n'est donc pas applicable tel quel ; solutions : regroupement de la ligne 3 avec une catégorie voisine si cela a un sens clinique, test exact (Fisher-Freeman-Halton) ou davantage de données. 8. McNemar (données appariées) : $\chi^2 = \frac{(12-4)^2}{16} =$ 4,00, p = 0,046 ; avec correction (8 - 1)²/16 = 3,06, p = 0,080 ; variante exacte p = 0,077 ; avec seulement 16 paires discordantes, on préfère la variante corrigée ou exacte ; le résultat est donc à la limite, non significatif. 9. Attendus 70, 80, 34, 16 ; χ^2 = 0 + 0,31 + 0,47 + 0,06 = 0,85, ddl 3, p = 0,84 ; la distribution observée est compatible avec les proportions hypothétiques. 10. Attendus 1,5 et 10,5 sur chaque ligne, donc test exact de Fisher ; p = 0,217, on ne rejette pas H0 ; avec 3 événements sur 24 patients, la puissance est très faible, et le résultat ne prouve pas l'innocuité du médicament. 11. Médianes 4 (EIQ 3-4) dans B et 3 (EIQ 2-3) dans A ; U = 2000 (ou 809, si le logiciel rapporte le U du groupe A) ; p < 0,001 ; $r_{\text{rb}} = 1 - \frac{2\times809}{2809} = 0{,}42$ (effet moyen à important).

Chapitre 8

Questions : 1 b ; 2 c ; 3 c ; 4 c ; 5 b ; 6 a, b, d ; 7 a, b, d ; 8 a, b, c, e ; 9 a, b, d ; 10 a, b, c, e Explications : pour la question 4, le risque global est de 1 - 0,95³ = 0,143, soit environ 14%, et non 15% (3 x 5% n'est qu'une approximation). Pour la question 7, un Shapiro-Wilk significatif pour n = 300 n'oblige pas à utiliser un test non paramétrique (c), et choisir le test d'après la valeur p est une erreur de procédure (e). Pour la question 8, les changements au sein des bras ne prouvent pas l'efficacité (d) ; seule la comparaison entre les bras montre la supériorité d'une intervention. Exercices proposés : 1. La variable : nombre de consultations (numérique discrète, fortement asymétrique) ; deux groupes indépendants ; test U de Mann-Whitney (alternative : le test t,

uniquement si les groupes sont grands et l'asymétrie modérée ; à un niveau avancé, régression de Poisson ou binomiale négative) ; H0 : les distributions du nombre de consultations sont identiques, H1 : elles diffèrent. 2. SpO2 (numérique), les mêmes 25 patients à deux temps de mesure, données appariées ; test t apparié si les différences sont normales, sinon Wilcoxon (la SpO2 est bornée supérieurement et souvent asymétrique) ; H0 : μd = 0. 3. La variable vacciné oui/non (nominale), trois groupes indépendants ; khi-deux sur un tableau 3 x 2, ddl = 2 (les effectifs attendus sont grands) ; H0 : p1 = p2 = p3. 4. Le score 0-21 (numérique ou ordinal), les mêmes 30 internes dans trois conditions, mesures répétées ; ANOVA à mesures répétées si la distribution le permet, sinon Friedman, puis post hoc avec correction. 5. L'accord de deux évaluateurs sur une variable binaire : coefficient kappa de Cohen (le test de McNemar montre seulement si un radiologue dit "pneumonie" systématiquement plus souvent que l'autre, et le khi-deux montre l'association, non l'accord). 6. Le but est la prédiction d'un critère binaire au cours du temps : si tous les patients ont été suivis complètement pendant 10 ans, régression logistique (OR) ; si la durée de suivi varie ou s'il y a des perdus de vue, régression de Cox (HR) sur le délai jusqu'au diabète. 7. $1 - 0{,}95^3 = 0{,}143$ (14,3%) ; seuil de Bonferroni 0,05/3 = 0,0167 ; aucune comparaison ne reste significative (0,03 et 0,04 > 0,0167) ; la bonne démarche était une ANOVA suivie du test de Tukey. 8. Le plus petit effectif attendu $= \frac{12 \times 10}{60} = 2{,}0 < 5$, donc test exact de Fisher. 9. Femmes $208{,}08 \pm 35{,}37$ mg/dL$(n\,86)$, hommes $210{,}91 \pm 38{,}82 (n\,74)$; SW p = 0,61 et 0,93 ; Levene p = 0,65 ; t de Student = -0,48, ddl 158, p = 0,63 (Welch p = 0,63) ; différence F - M -2,82 mg/dL (IC à 95% -14,50 à 8,85) ; d = -0,08 ; Mann-Whitney p = 0,76 ; nous ne rejetons pas H0, le cholestérol ne diffère pas entre les sexes. 10. Rural 10/65 (15,4%), urbain 19/95 (20,0%) ; plus petit effectif attendu 11,8 ; $\chi^2 = 0{,}55$, ddl 1, p = 0,46 (Yates p = 0,59, Fisher p = 0,53) ; nous ne rejetons pas H0, le diabète n'est pas associé au milieu de résidence dans cet échantillon. 11. Réponse ouverte ; on vérifie la cohérence entre le schéma d'étude, le type de variable et le test.

Travaux pratiques 8 - réponses attendues (pour l'enseignant)

L1. Le tour de taille selon le sexe. Femmes $86{,}34 \pm 11{,}67$ cm (médiane 84, EIQ 78-93), hommes $91{,}09 \pm 12{,}19$ cm (médiane 90, EIQ 83-97). SW : femmes p = 0,008, hommes p = 0,005 (légère asymétrie), mais les groupes sont grands (86 et 74) ; les deux tests sont donc acceptables si le choix est justifié. Levene p = 0,80. Test t de Student t = -2,52, ddl 158, p = 0,013 (Welch t = -2,51, p = 0,013) ; différence F - M -4,76 cm (IC à 95% -8,50 à -1,01) ; d = -0,40. Mann-Whitney p = 0,0096, r_rb = 0,24. Conclusion : nous rejetons H0 ; les hommes ont un tour de taille plus grand d'environ 5 cm en moyenne (effet faible à moyen) ; les seuils cliniques du tour de taille sont différents selon le sexe.

L2. Les triglycérides selon le diabète. Sans diabète : médiane 154 mg/dL (EIQ 104,5-209,5), SW p < 0,001 ; avec diabète : médiane 143 (EIQ 112-179), SW p = 0,11. Test approprié : Mann-Whitney (distribution asymétrique, groupes inégaux de 131 et 29) ; alternative : le test t sur les triglycérides log-transformés. U = 2033,5 (rapporté pour le groupe sans diabète), p = 0,55 ; r_rb = 0,07 ; Welch sur les valeurs brutes p = 0,30. Nous ne rejetons pas H0. En (g), on accepte le commentaire selon lequel, dans les populations réelles, les triglycérides sont généralement plus élevés chez les diabétiques ; sur 29 cas et avec des données simulées, la différence n'apparaît pas.

L3. L'HbA1c à l'inclusion selon le bras. A 5,61 ± 0,45 ; B 5,57 ± 0,43 ; C 5,55 ± 0,46% ; Levene p = 0,58 ; ANOVA F(2 ; 157) = 0,27, p = 0,76, $\eta^2 = 0{,}003$; Kruskal-Wallis p = 0,55. Nous ne rejetons pas H0. Le point clé du problème est la discussion : dans un essai randomisé, les différences à l'inclusion sont par définition dues au hasard ; le tableau des caractéristiques à l'inclusion se décrit, il ne se teste pas (CONSORT). Ce qui compte, c'est l'ampleur clinique d'une éventuelle différence (ici négligeable : 0,06 point de pourcentage).

L4. Le cholestérol à l'inclusion selon le bras. A 209,32 ± 39,01 ; B 202,85 ± 37,55 ; C 215,87 ± 33,56 mg/dL ; Levene p = 0,61 ; ANOVA F(2 ; 157) = 1,68, p = 0,19, η^2 = 0,021 ; Kruskal-Wallis p = 0,19. Nous ne rejetons pas H0. Même discussion que pour L3 ; la différence maximale entre les bras (13 mg/dL) n'est pas cliniquement importante pour les résultats glycémiques.

L5. La perte de poids selon le bras. A 0,71 ± 2,48 kg (médiane 1,1) ; B 2,84 ± 2,36 (médiane 2,3) ; C 3,37 ± 2,39 (médiane 3,55) ; SW par bras p = 0,70, 0,26, 0,17 ; Levene p = 0,79. ANOVA F(2 ; 157) = 18,25, p < 0,001, η^2 = 0,189. Tukey : B - A 2,13 kg (IC à 95% 1,02-3,24 ; p < 0,001) ; C - A 2,66 kg (IC à 95% 1,56-3,77 ; p < 0,001) ; C - B 0,53 kg (IC à 95% -0,57 à 1,63 ; p = 0,49). Kruskal-Wallis H = 27,50, p < 0,001. Conclusion : les deux bras avec application ont entraîné une perte de poids plus importante que le simple conseil (de 2,1-2,7 kg), et la téléconsultation n'a pas apporté de bénéfice supplémentaire significatif par rapport à l'application seule.

L6. L'IMC selon le statut tabagique. Anciens fumeurs médiane 28,0 (EIQ 26,5-30,9) ; fumeurs 28,3 (26,3-31,4) ; non-fumeurs 28,35 (26,1-31,08). SW p < 0,05 dans tous les groupes (0,029, < 0,001, 0,003) ; Levene p = 0,048. Test approprié : Kruskal-Wallis (asymétrie dans tous les groupes, le groupe des anciens fumeurs ne compte que 33 patients) ; H = 0,46, ddl 2, p = 0,80. L'ANOVA (p = 0,35) et le test de Welch (p = 0,38) mènent à la même conclusion. Nous ne rejetons pas H0.

L7. La satisfaction selon le sexe. Médiane 4 (EIQ 3-4) chez les deux sexes ; Mann-Whitney U = 3221, p = 0,89 ; r_rb = 0,01. Nous ne rejetons pas H0. On retire des points pour la présentation des moyennes (3,56 et 3,49) comme description principale et pour l'utilisation du test t.

L8. L'âge selon le milieu de résidence. Rural 57,38 ± 9,08 ans (n 65), urbain $55{,}55 \pm 8{,}52 (n\,95)$; SW p = 0,87 et 0,69 ; Levene p = 0,64 ; t de Student = 1,30, ddl 158, p = 0,19 (Welch t = 1,29, p = 0,20) ; différence rural - urbain 1,84 an (IC à 95% -0,98 à 4,66) ; d = 0,21 ; Mann-Whitney p = 0,17. Nous ne rejetons pas H0.

L9. La perte de poids ≥ 5% selon le sexe. Femmes 23/86 (26,7%), hommes 15/74 (20,3%). Effectifs attendus : femmes 20,4 avec perte de poids et 65,6 sans ; hommes 17,6 et 56,4. χ^2 = 0,92, ddl 1, p = 0,34 (Yates p = 0,44, Fisher p = 0,36) ; φ = 0,08. Nous ne rejetons pas H0.

L10. Le tabagisme selon le milieu de résidence. Rural : anciens fumeurs 14 (21,5%), fumeurs 16 (24,6%), non-fumeurs 35 (53,8%) ; urbain : 19 (20,0%), 37 (38,9%), 39 (41,1%). Plus petit effectif attendu 13,4. χ^2 = 3,80, ddl 2, p = 0,15 ; V de Cramér = 0,15. Nous ne rejetons pas H0 ; la tendance (davantage de fumeurs en milieu urbain) n'atteint pas le seuil de signification.

L11. Le tabagisme selon le bras. A : anciens fumeurs 5 (9,4%), fumeurs 18 (34,0%), non-fumeurs 30 (56,6%) ; B : 11 (20,8%), 18 (34,0%), 24 (45,3%) ; C : 17 (31,5%), 17 (31,5%), 20 (37,0%). Plus petit effectif attendu 10,9. χ^2 = 8,60, ddl 4, p = 0,072 ; V de Cramér = 0,16. Nous ne rejetons pas H0. Discussion attendue : le déséquilibre (par exemple 9,4% d'anciens fumeurs dans A contre 31,5% dans C) est dû au hasard, il résulte de la randomisation ; il n'"invalide" pas l'étude ; si le tabagisme influence le critère de jugement principal, il peut être inclus comme variable d'ajustement dans l'analyse, décision qui devrait être prise à l'avance.

L12. La glycémie ≥ 100 à 0 et 6 mois, McNemar dans chaque bras. Bras A : ≥ 100 aux deux temps 32, uniquement à l'inclusion 3, uniquement à 6 mois 1, à aucun des deux 17 ; proportion de 35/53 (66,0%) à l'inclusion et de 33/53 (62,3%) à 6 mois ; χ^2 de McNemar = 1,00, p = 0,32 (avec correction 0,25, p = 0,62 ; exact p = 0,63) ; nous ne rejetons pas H0. Bras B : aux deux temps 19, uniquement à l'inclusion 15, uniquement à 6 mois 0, à aucun des deux 19 ; 34/53 (64,2%) contre 19/53 (35,8%) ; χ^2 = 15,00, p < 0,001 (avec correction 13,07, p < 0,001 ; exact p < 0,001) ; nous

rejetons H0. Discussion : les résultats au sein des bras décrivent le changement dans chaque bras ; la comparaison de l'efficacité se fait entre les bras (par exemple un khi-deux sur la proportion de patients passés sous 100, ou l'analyse de la baisse de la glycémie selon le bras). On retire des points pour le khi-deux habituel sur le tableau "temps x glycémie ≥ 100".

L13. La glycémie et l'HbA1c à l'inclusion. Le nuage de points montre une relation linéaire, sans valeurs extrêmes importantes ; l'HbA1c suit une loi normale (SW p = 0,30), la glycémie présente un léger écart (SW p = 0,037 pour n = 160, asymétrie 0,37). Corrélation de Pearson r = 0,738, p < 0,001 (Spearman rho = 0,727, p < 0,001) ; $R^2 = 0{,}54$. Nous rejetons H0 : ρ = 0. Association forte, positive, attendue sur le plan physiologique (les deux mesurent la glycémie, à court terme et à long terme) ; il ne s'agit pas d'une mesure de l'accord entre les deux, puisque ce sont des grandeurs différentes.

L14. La glycémie à 0, 3 et 6 mois dans le bras A. Moyennes 107,8 / 105,7 / 104,8 mg/dL ; médianes 111 / 104 / 105. SW à chaque temps p = 0,20, 0,62, 0,71 (normales) ; test de Mauchly p = 0,020 (sphéricité violée, ε de Greenhouse-Geisser 0,88). Analyse principale justifiée : ANOVA à mesures répétées avec correction de Greenhouse-Geisser : F(2 ; 104) = 2,73, p = 0,078 (sans correction p = 0,070) ; nous ne rejetons pas H0. L'alternative, Friedman : χ^2 = 6,04, ddl 2, p = 0,049 ; les comparaisons de Wilcoxon avec Bonferroni ne restent pas significatives (plus petit p ajusté 0,075). La réponse complète reconnaît la discordance, rapporte le test décidé à l'avance et conclut que, dans le bras A, il n'existe pas de preuve convaincante d'une baisse de la glycémie au cours du temps. On retire des points pour le choix du test de Friedman justifié uniquement par le fait qu'il donne p < 0,05.

L15. L'âge et le FINDRISC. Le score FINDRISC est un score à valeurs entières, légèrement asymétrique (SW p = 0,034) ; on accepte Spearman (rho = 0,456, p < 0,001) ou Pearson (r = 0,442, p < 0,001), avec justification. Nous rejetons H0. Discussion attendue : l'âge est l'une des composantes du score FINDRISC (les patients reçoivent des points selon leur classe d'âge) ; une partie de la corrélation est donc "construite" par la définition du score ; la corrélation ne montre pas que l'âge apporte une information indépendante des autres composantes du score.

Chapitre 9

Questions : 1 b ; 2 c ; 3 d ; 4 b ; 5 c ; 6 a, b, e ; 7 a, b, d ; 8 a, b, d ; 9 a, b, c, e ; 10 a, b, c, e

Explications : Pour la question 3, $R^2 = 0{,}60^2 = 0{,}36$; la proportion due à d'autres facteurs est donc 1 - 0,36 = 0,64 (64%) ; l'option "40%" apparaît lorsqu'on soustrait de 1 r, et non r². Pour la question 6, l'option c est fausse parce que r ne mesure que la relation linéaire (une relation en forme de U peut avoir un r proche de zéro), et l'option d est fausse parce que r ne dépend pas des unités de mesure. Pour la question 8, $R^2 = 0{,}088^2 = 0{,}0077$, soit moins de 1%, et non 8,8% ; la corrélation n'est significative qu'en raison de la grande taille de l'échantillon et ne démontre pas une relation causale. Pour la question 9, la normalité est exigée pour les résidus, et non pour la variable X.

Exercices proposés : 1. mx = 53, my = 130,33 ; Sxy = 660, Sxx = 868, Syy = 553,33 ; $r = \frac{660}{\sqrt{868 \times 553{,}33}} = 0{,}952$ (corrélation positive forte) ; t = 6,24, ddl 4, t critique 2,776, p = 0,003, on rejette H0 ; $b = \frac{0{,}760 \text{ mmHg}}{\text{an}}$, a = 90,03 ; PAS = 90,03 + 0,760 x âge ; à 60 ans, 135,7 mmHg ; $R^2 = 0{,}907$. 2. n = 40 : $t = \frac{0{,}30 \times \sqrt{38}}{\sqrt{0{,}91}} = 1{,}94$, ddl 38, t critique 2,024, p = 0,060, on ne rejette pas H0. n = 200 : t = 4,43, p < 0,001, on rejette H0. Le même r devient significatif avec un échantillon plus grand ; selon Colton, la corrélation est faible dans les deux cas. 3. $z = 0{,}5 \times \ln\left(\frac{1{,}62}{0{,}38}\right) = 0{,}725$; $\text{ES} = \frac{1}{\sqrt{47}} = 0{,}146$; IC sur l'échelle z 0,439-1,011 ; IC à 95% de ρ : 0,41-0,77. La borne

supérieure dépasse 0,75 ; la corrélation dans la population pourrait donc être forte, mais la borne inférieure (0,41) est aussi compatible avec une corrélation faible. 4. Rangs EVA : 1, 4, 2, 6, 5, 7, 3 ; rangs d'anxiété : 2, 4, 1, 7, 6, 5, 3 ; di : -1, 0, 1, -1, -1, 2, 0 ; $\sum {d_i}^2 = 8$; $r_s = 1 - \frac{6\times8}{7\times48} = 0{,}857$, p = 0,014 ; corrélation positive forte. 5. $R^2 = 0{,}738^2 = 0{,}545$: environ 54,5% de la variation de l'HbA1c est associée linéairement à la variation de la glycémie à jeun ; $1 - R^2 = 0{,}455$: 45,5% de la variation de l'HbA1c relève d'autres facteurs. Selon Colton, corrélation modérée, très proche du seuil de 0,75. 6. 0,81 positive forte ; -0,573 négative modérée ; 0,256 positive faible ; -0,088 absence de corrélation ; 0,729 positive modérée ; -0,832 négative forte ; 0,107 absence de corrélation ; 0,50 positive modérée (à la limite entre faible et modérée). 7. Moyenne des différences (biais) = 3,0 mmHg ; ÉT = 3,46 mmHg ; limites d'agrément de $3{,}0 \pm 1{,}96 \times 3{,}46 = -3{,}8$ à 9,8 mmHg. Le coefficient de corrélation montrerait seulement si les patients sont ordonnés de façon semblable ; une différence systématique de quelques mmHg ne ferait pas diminuer r. 8. 5 ans : 1,11^5 = 1,69 ; 10 ans : 1,11^10 = 2,84 (avec le coefficient non arrondi, environ 2,9). Ce n'est pas correct : l'OR porte sur les cotes (p / (1 - p)), et non sur les probabilités ; les deux ne sont proches que pour des événements rares. 9. Femme non fumeuse, 45 ans, IMC 27 : 64,51 + 1,389 x 27 + 0,449 x 45 - 1,119 = 121,1 mmHg. Homme fumeur, 65 ans, IMC 35 : 64,51 + 1,389 x 35 + 0,449 x 65 - 1,481 + 6,738 = 147,6 mmHg. Différence de 26,5 mmHg : l'IMC y contribue pour 8 x 1,389 = 11,1 mmHg (la plus grande part), l'âge pour 20 x 0,449 = 9,0 mmHg, le tabagisme pour 6,738 + 1,119 = 7,9 mmHg, le sexe pour -1,5 mmHg. 10. (a) Chaque centimètre supplémentaire est associé à un poids plus élevé de 0,95 kg en moyenne. (b) 1 - 0,38 = 0,62, soit 62%. (c) Pour 170 cm : -95 + 0,95 x 170 = 66,5 kg. Pour 120 cm, la formule donne 19 kg, une extrapolation absurde en dehors du domaine 155-195 cm ; le modèle ne s'applique pas.

Chapitre 10

Questions : 1 c ; 2 c ; 3 b ; 4 b ; 5 c ; 6 a, b, d, e ; 7 a, b, c, e ; 8 a, b, d, e ; 9 a, b, c, e ; 10 a, b, d, e

Explications : Pour la question 1, l'étude part de la maladie (cas et témoins) ; il s'agit donc d'une étude cas-témoins et on ne calcule que l'OR. Pour la question 2, la différence de risques est 0,20 - 0,10 = 0,10, et $\text{NNH} = \frac{1}{0{,}10} = 10$; l'option "5" apparaît lorsqu'on divise par le risque chez les exposés. Pour la question 3, RR = 0,70 signifie une réduction relative du risque de 30% (1 - 0,70), et non de 70%. Pour la question 4, la maladie est fréquente (prévalences de 45% et de 18,75%) ; l'OR (3,55) est donc plus grand que le rapport de prévalences (2,40). Pour la question 8, le facteur de confusion ne se situe pas sur la voie causale ; une variable intermédiaire (médiateur) n'est pas ajustée comme un facteur de confusion. Pour la question 10, la valeur p ne fait pas partie des critères de Bradford Hill.

Exercices proposés : 1. Étude cas-témoins (sélection selon l'infertilité). a = 89 (infertiles avec DIU), b = 640 (témoins avec DIU), c = 194 (infertiles sans DIU), d = 3193 (témoins sans DIU). $\text{OR} = \frac{89\times3193}{640\times194} = 2{,}29 (;\ \text{IC}\,95\%\ 1{,}76 - 2{,}98)$; χ^2 = 39,35, p < 0,001. La cote d'avoir utilisé un DIU est 2,3 fois plus élevée chez les femmes infertiles ; association significative. 2. a = 23, b = 84, c = 117, d = 800 ; $\text{OR} = \frac{23\times800}{84\times117} = 1{,}87 (;\ \text{IC}\,95\%\ 1{,}14 - 3{,}09)$; χ^2 = 6,20, p = 0,013. Le syndrome de Down est rare dans la population ; si l'échantillon a été constitué comme une étude cas-témoins, l'OR est donc une approximation du RR ; le RR ne peut pas être calculé directement dans ce schéma d'étude. 3. Bras B : 16 / 53 = 30,2% ; bras A : 6 / 53 = 11,3% ; RR = 2,67 (IC à 95% 1,13-6,29) ; différence absolue 18,9 points de pourcentage (IC à 95% 3,9-33,9) ; NST = 5,3, arrondi à 6. Résultat très proche de celui du bras C par rapport au bras A (RR 2,62 ; NST 6). 4. a = 10, b = 43, c = 19, d = 88. $\text{RP} = \frac{18{,}9\%}{17{,}8\%} = 1{,}06 (;\ \text{IC}\,95\%\ 0{,}53 - 2{,}12)$; OR = 1,08 (IC à 95% 0,46-2,52) ; χ^2 = 0,03, p = 0,86. On ne rejette pas H0 ; dans mg2026, le tabagisme n'est pas

associé au diabète. Conclusion clinique : les données ne montrent pas d'association, mais l'intervalle large n'exclut pas non plus un effet modéré. 5. a = 21, b = 33, c = 20, d = 86. RP = $\frac{38,9\%}{18,9\%} = 2,06(;\ \text{IC}\,95\%\,1,23 - 3,46)$; OR = 2,74 (IC à 95% 1,32-5,69) ; χ^2 = 7,52, p = 0,006. L'OR dépasse le RP parce que l'événement est fréquent (25,6% ont une PAS ≥ 140 mmHg). 6. Exposés : 10 / 400 = 2,5 cas pour 100 personnes-années ; non-exposés : 8 / 800 = 1,0 pour 100 personnes-années ; IRR = 2,5. 7. P ≈ I x D = 0,005 x 4 = 0,02, soit 2%. Si la durée double pour atteindre 8 ans, à incidence égale la prévalence augmente à environ 4%, bien que le traitement constitue un progrès. 8. Différence de risques = 6% - 2% = 4 points de pourcentage ; RR = 3 ; $\text{NNH} = \frac{1}{0,04} = 25$: pour 25 patients traités, un effet indésirable supplémentaire apparaît par rapport au placebo. 9. FAe = (3 - 1) / 3 = 0,667 (66,7%) ; $\text{FAP} = \frac{0,40\times2}{1+0,40\times2} = \frac{0,8}{1,8} = 0,444(44,4\%)$. 10. OR de la strate $1 = \frac{20\times60}{80\times10} = 1,5$; OR de la strate $2 = \frac{30\times60}{20\times60} = 1,5$. Le tableau combiné : a = 50, b = 100, c = 70, d = 120, OR brut = 0,86. $\text{OR(MH)} = \frac{\frac{20\times60}{170}+\frac{30\times60}{170}}{\frac{80\times10}{170}+\frac{20\times60}{170}} = \frac{17,65}{11,76} = 1,5$. Les mesures dans les strates sont égales entre elles et diffèrent beaucoup de la mesure brute : le facteur de stratification est un facteur de confusion, qui, dans l'analyse brute, inverse même le sens de l'association. 11. a = 10, b = 55, c = 19, d = 76 ; OR = 0,73 (IC à 95% 0,31-1,69) ; χ^2 = 0,55, p = 0,46. L'intervalle contient 1 : on ne peut pas affirmer que le milieu rural est protecteur ; le résultat est non significatif.

Chapitre 11

Questions : 1 c ; 2 b ; 3 b ; 4 d ; 5 c ; 6 a, c, d, e ; 7 a, b, d ; 8 a, b, e ; 9 a, b, c ; 10 a, b, c, d

Explications : Pour la question 2, sur 1000 patients, 100 sont malades (80 tests positifs) et 900 sont sains (90 faux positifs), donc $\text{VPP} = \frac{80}{170} = 47\%$; l'option a) confond la VPP avec la sensibilité. Pour la question 5, BF10 = 0,05 signifie BF01 = 20, c'est-à-dire des preuves fortes en faveur de H0 ; le facteur de Bayes n'est pas une valeur p. Pour la question 7, l'option c) est fausse : les rapports de vraisemblance proviennent de la sensibilité et de la spécificité, qui ne dépendent pas en principe de la prévalence (mais peuvent changer avec le spectre des patients). Pour la question 9, le biais de vérification est un biais de conception des études diagnostiques (tous les patients ne reçoivent pas le standard de référence), et non une erreur cognitive au lit du malade.

Exercices proposés :

1. a = 45, b = 15, c = 88, d = 352, N = 500. $p(B^+) = \frac{60}{500} = 0,12$; $p(S^+/B^+) = \frac{45}{60} = 0,75$; $p(S^+/B^-) = \frac{88}{440} = 0,20$; $p(B^+/S^+) = \frac{45}{133} = 0,338$. Bayes : $\frac{0,75\times0,12}{0,75\times0,12+0,20\times0,88} = \frac{0,090}{0,266} = 0,338$.
2. Sur 10 000 femmes : 80 avec un cancer (72 positives, 8 faux négatifs), 9920 sans cancer (environ 694 faux positifs, 9226 négatives). $\text{VPP} = \frac{72}{766} = 9,4\%$; $\text{VPN} = \frac{9226}{9234} = 99,9\%$. Pour chaque cancer détecté, on explore de façon complémentaire environ 766/72 = 10,6, soit à peu près 11 femmes ayant une mammographie positive.
3. 0,40 -> 0,40/0,60 = 0,667 ; 0,05 -> 0,0526 ; cote 3 -> 3/4 = 0,75 ; cote 0,25 -> 0,25/1,25 = 0,20.
4. (a) $\text{RV}^+ = \frac{0,703}{0,340} = 2,07$; $\text{RV}^- = \frac{0,297}{0,660} = 0,45$. (b) Cote pré-test 0,429 ; avec un dNLR au-dessus du seuil, 0,429 x 2,07 = 0,886, probabilité 47,0% ; en dessous du seuil, 0,429 x 0,45 = 0,193, probabilité 16,2%. Le test déplace peu la probabilité, ce qui concorde avec une AUC de 0,624 (satisfaisante sur l'échelle du cours). (c) Trois décimales (70,312%) suggèrent une précision qu'un échantillon de 336 patients n'offre pas ; une décimale (70,3%) et un intervalle de confiance auraient été appropriés.

5. $RV^- = \frac{0,07}{0,86} = 0,081$. À 33,8% : 0,512 x 0,081 = 0,042, probabilité post-test 4,0%. À 5% : 0,0526 x 0,081 = 0,0043, probabilité post-test 0,43%.
6. En série : Se = 0,90 x 0,85 = 0,765 ; Sp = 1 - 0,20 x 0,05 = 0,99. En parallèle : Se = 1 - 0,10 x 0,15 = 0,985 ; Sp = 0,80 x 0,95 = 0,76. Triage aux urgences : en parallèle (on ne manque pas de malades). Confirmation avant un traitement toxique : en série (peu de faux positifs).
7. Produits : migraine 0,70 x 0,05 x 0,05 x 0,70 = 0,001225 ; céphalée de tension 0,25 x 0,02 x 0,10 x 0,10 = 0,00005 ; HSA 0,05 x 0,80 x 0,60 x 0,70 = 0,0168. Somme 0,018075. Probabilités a posteriori : migraine 6,8% ; céphalée de tension 0,3% ; HSA 92,9%. Décision : imagerie en urgence (scanner cérébral), quelles que soient les réserves concernant l'indépendance des signes.
8. Cote au seuil 0,10/0,90 = 0,111. Seuil de test : 0,111/4 = 0,0278, soit 2,7%. Seuil de test-traitement : 0,111/0,25 = 0,444, soit 30,8%.
9.
    (a) SOFA : 0,771 bon, 0,873 très bon ; NEWS2 : 0,892 très bon, 0,729 bon ; APACHE II : 0,698 satisfaisant, 0,806 à la limite entre bon et très bon (0,81 après arrondi). (b) NEWS2 à l'admission, SOFA au cinquième jour ; la performance d'un score dépend du moment de son application. (c) La régression logistique produit des rapports de cotes (OR) ; les HR proviennent d'un modèle de Cox, qui exige des données de délai jusqu'à l'événement. Le fait de rapporter des HR issus d'une régression logistique est un point à discuter lors de la lecture critique.
10.
    (a) $BF_{01} = \frac{1}{0,16} = 6,25$: preuves modérées en faveur de l'absence de corrélation. (b) Cote a posteriori 1 x 0,16 = 0,16 ; probabilité 0,16/1,16 = 13,8%.
11. P : adultes non explorés à risque cardiométabolique suivis en médecine générale ; I : HbA1c (par exemple seuil ≥ 6,5%) ; C : HGPO comme standard de référence ; O : sensibilité, spécificité, rapports de vraisemblance pour le diabète de type 2. Schéma d'étude approprié : étude transversale d'exactitude diagnostique sur des patients consécutifs, qui reçoivent tous les deux tests, rapportée selon STARD 2015.

Chapitre 12

Questions : 1 c ; 2 b ; 3 d ; 4 c ; 5 c ; 6 a, b, d ; 7 a, b, c, e ; 8 a, b, d ; 9 a, b, d, e ; 10 a, c, e

Explications : Pour la question 1, l'option b) est la formule erronée qui apparaît parfois dans les notes de cours ; d contient les vrais négatifs, c'est-à-dire des personnes correctement classées. Pour la question 4, l'AUC n'est ni l'exactitude ni la sensibilité ; c'est la probabilité que, dans une paire malade-sain choisie au hasard, le malade ait la valeur la plus élevée. Pour la question 6, l'option c) décrit le biais de vérification, et l'option e) est fausse parce qu'un échantillon plus grand avec le même spectre inapproprié donne seulement une estimation plus précise d'une valeur biaisée. Pour la question 8, le taux d'erreur est (7 + 33)/160 = 0,25, et la spécificité est 98/131 (98/105 est la VPN).

Exercices proposés :

1. a = 45, b = 5, c = 30, d = 120. Prévalence 50/200 = 0,25 ; Se 0,90 ; $\frac{\text{Sp}\,120}{150} = 0,80$; $\frac{\text{VPP}\,45}{75} = 0,60$; $\frac{\text{VPN}\,120}{125} = 0,96$; $\frac{\text{Ex}\,165}{200} = 0,825$; $\frac{\text{TE}\,35}{200} = 0,175$; J 0,70 ; RV+ 4,5 ; RV- 0,125 ; $\frac{\text{DOR}(45\times120)}{5\times30} = 36$.
2. a = 80, b = 20 ; à partir de $\frac{80}{80+c} = \frac{4}{5}$, c = 20 ; à partir de $\frac{d}{20+d} = \frac{9}{10}$, d = 180. Se 0,80 ; $\frac{\text{Sp}\,180}{200} = 0,90$; N = 300.
3. d = 150 et Sp = $\frac{3}{4}$, donc c + d = 200 et c = 50 ; à partir de $\frac{150}{b+150} = \frac{15}{16}$, b = 10 ; a = 40. $\frac{\text{Se}\,40}{50} = 0,80$; $\frac{\text{VPP}\,40}{90} = 0,444$; N = 250.

4. 100 malades : 75 VP, 25 FN ; 900 non-malades : 855 VN, 45 FP. $\frac{\text{VPP } 75}{120} = 62{,}5\%$; $\frac{\text{VPN } 855}{880} = 97{,}2\%$.
5. HPV : a = 17, b = 1, c = 19, d = 27 ; $\frac{\text{Ex } 44}{64} = 68{,}8\%$; J 0,531. p16/Ki-67 : a = 12, b = 6, c = 7, d = 39 ; $\frac{\text{Ex } 51}{64} = 79{,}7\%$; J 0,514. Le test HPV convient mieux pour exclure (Se 94,4%, RV- 0,095), p16/Ki-67 pour confirmer (Sp 84,8%, RV+ 4,38). Des indices de Youden presque égaux masquent ces rôles différents.
6. Trapèzes : 0,010 + 0,025 + 0,070 + 0,170 + 0,282 + 0,297 = 0,854 ; très bon (0,81-0,90).
7. Po $= \frac{175}{200} = 0{,}875$; A "pneumonie" 45/200 = 0,225, $\frac{B\,40}{200} = 0{,}20$; Pe = 0,225 x 0,20 + 0,775 x 0,80 = 0,665 ; κ = 0,210/0,335 = 0,63, accord substantiel.
8.
   (a) Décès 159 x 0,2515 = 40, survivants 119 ; VP 0,878 x 40 = 35, FN 5 ; VN 0,7372 x 119 = 88, FP 31. (b) VPP $= \frac{35}{66} = 53{,}0\%$, proche des 53,73% rapportés. (c) Wilson pour 35/40 : environ 73,9-94,5%, près de 80 fois plus large que l'intervalle rapporté ; l'intervalle du résumé ne peut pas provenir d'une méthode habituelle pour une proportion estimée sur 40 décès.
9. $\frac{\text{Se } 40}{50} = 80\%$; $\frac{\text{Sp } 360}{450} = 80\%$; précision 40/130 = 30,8% ; $F_1 = \frac{2\times0{,}308\times0{,}80}{0{,}308+0{,}80} = 0{,}44$; $\frac{\text{Ex } 400}{500} = 80\%$. Un classificateur qui déclarerait tout bénin aurait une exactitude de 450/500 = 90%, supérieure à celle de l'algorithme, ce qui montre pourquoi l'exactitude ne s'utilise pas seule lorsque les classes sont déséquilibrées ; la faible précision reflète la prévalence de 10%.
10. 0,624 satisfaisant ; 0,752 bon. Le test H0 : AUC = 0,5 montre seulement que le marqueur discrimine mieux que le hasard ; sur 336 patients, même une discrimination modeste devient significative. L'utilité clinique se juge d'après l'ampleur de l'AUC, les rapports de vraisemblance au seuil et les conséquences des décisions.
11. Cote pré-test 0,111 ; $\text{RV}^+ = \frac{0{,}801}{0{,}143} = 5{,}60$, donc VPP $= \frac{0{,}622}{1{,}622} = 38{,}4\%$; $\text{RV}^- = \frac{0{,}199}{0{,}857} = 0{,}232$, probabilité post-test 2,5%, donc VPN 97,5%.
12. $\text{RV}^+ = \frac{0{,}379}{0{,}053} = 7{,}1$; $\text{RV}^- = \frac{0{,}621}{0{,}947} = 0{,}66$. Un résultat ≥ 126 mg/dL multiplie la cote par environ 7 (utile pour confirmer), mais un résultat inférieur au seuil ne change presque pas la probabilité (RV- 0,66) ; ce seuil manque donc la majorité des diabétiques identifiés par l'HGPO et n'est pas adapté au dépistage.

Chapitre 13

Questions : 1 c ; 2 b ; 3 c ; 4 c ; 5 b ; 6 a, b, d ; 7 a, c, d ; 8 a, b, d, e ; 9 a, b, c ; 10 a, c, d

Explications : Pour la question 4, NNT $= \frac{1}{0{,}183} = 5{,}5$ et l'on arrondit toujours vers le haut, à 6 ; les valeurs 3 et 29 sont les bornes de l'intervalle de confiance, et non l'estimation ponctuelle. Pour la question 6, l'option c est fausse : l'ITT dilue l'effet avec les patients non observants ; elle tend donc à sous-estimer l'effet biologique, et non à le surestimer ; l'option e décrit l'analyse per protocole. Pour la question 9, une PAS de 1540 et une taille de 1,72 sont des erreurs dans des variables numériques et se détectent par le minimum, le maximum et la boîte à moustaches, et non par un tableau de fréquences des variables catégorielles. Pour la question 10, le test de Bartlett doit être significatif (items corrélés), et l'alpha mesure la cohérence interne, et non la validité.

Exercices proposés : 1. $n = \frac{2\times(1{,}96+1{,}28)^2\times12^2}{8^2} = \frac{2\times10{,}50\times144}{64} = 47{,}2$, donc 48 patients par groupe ; avec 10% de perdus de vue, 48/0,9 = 53,3, donc 54 par groupe. 2. p1 = 0,35, p2 = 0,20, $\bar{p}$ = 0,275 ; $n = \frac{\left[1{,}96\times\sqrt{2\times0{,}275\times0{,}725}+0{,}84\times\sqrt{0{,}35\times0{,}65+0{,}20\times0{,}80}\right]^2}{0{,}15^2} = 137{,}9$, donc 138 patients par

groupe. 3. Effet de plan = 1 + (20 - 1) x 0,02 = 1,38 ; 64 x 1,38 = 88,3, donc 89 patients par bras (environ 5 cabinets de 20 patients par bras). 4. $\alpha = \frac{5}{4} \times \left(1 - \frac{5,5}{18,0}\right) = 1,25 \times 0,694 = 0,868$; bonne cohérence interne. 5. Risque de 30/150 = 20% avec le programme, de 45/150 = 30% dans le groupe contrôle ; RR = 0,67 (réduction relative du risque de 33%) ; réduction absolue du risque de 10 points de pourcentage ; NNT $= \frac{1}{0,10} = 10$. 6. 1 - 0,95^8 = 0,337 (environ 34%) ; le seuil de Bonferroni est de 0,05/8 = 0,00625 pour chaque test. 7. a) la borne inférieure -2,4 est au-dessus de -3 : non-infériorité démontrée ; b) la borne inférieure -4,1 descend sous -3 : résultat non concluant, la non-infériorité n'est pas démontrée (mais l'infériorité non plus) ; c) l'intervalle entier est au-dessus de 0 : non-infériorité et même supériorité, si le test de supériorité était prévu dans le protocole. 8. a) essai croisé (cross-over), analyse pour données appariées ; b) essai randomisé en grappes (écoles) ; c) étude pilote à un seul bras, mesures répétées (ANOVA à mesures répétées ou Friedman) ; d) plan factoriel 2 x 2 ; e) cohorte observationnelle (l'exposition n'a pas été attribuée par l'investigateur). 9. 1,75 : unité erronée (mètres) ; 1680 : chiffre en trop ; "1 m 70" : texte dans une colonne numérique. Toutes se vérifient dans le document source et ne se corrigent qu'à partir de celui-ci ; si la source n'est pas disponible, la valeur est déclarée manquante. Prévention : validation dans le formulaire "nombre entre 120 et 220", avec le message "en centimètres". Les valeurs 168, 172, 159, 165, 181 sont plausibles. 10. 3/300 = 0,01, c'est-à-dire que la borne supérieure de l'IC à 95% est d'environ 1% ; l'effet indésirable pourrait survenir jusqu'à environ 1 cas pour 100 patients sans avoir été observé. Les effets plus rares ne peuvent être détectés qu'en phase IV, sur des milliers ou des dizaines de milliers de patients. 11. $p = \frac{16}{53} = 30,2\%$; ES $= \sqrt{\frac{0,302 \times 0,698}{53}} = 0,063$; IC à 95% de Wald 17,8-42,5% ; IC à 95% de Wilson 19,5-43,5% (asymétrique, recommandé pour les petits échantillons ou les proportions éloignées de 50%). 12. RR $= \frac{\frac{16}{53}}{\frac{6}{53}} = 2,67 (;\ \text{IC}\,95\%\ 1,13 - 6,29)$; différence de risques 18,9 points de pourcentage (IC à 95% 3,9-33,9) ; NST = 5,3, soit 6. Résultats pratiquement identiques à ceux du bras C par rapport au bras A (RR 2,62, NST 6) : sur ce critère, la téléconsultation n'apporte pas de bénéfice au-delà de l'application (RR C par rapport à B 0,98, IC à 95% 0,55-1,75).

Chapitre 14

Questions : 1 b ; 2 c ; 3 b ; 4 b ; 5 c ; 6 b ; 7 b ; 8 c ; 9 b ; 10 b ; 11 b ; 12 b ; 13 b ; 14 b ; 15 b ; 16 a, b, c ; 17 a, c, e ; 18 a, b, c, e ; 19 a, b, c ; 20 a, b, c, e ; 21 a, b, d ; 22 a, b, c, d ; 23 a, b, c, e ; 24 a, b, c, e ; 25 a, b, c, e ; 26 a, b, d, e ; 27 a, b, c, e ; 28 a, b, c, d ; 29 a, b, d, e ; 30 a, b, c, e

Explications : Pour la question 2, les options a et b sont les pièges classiques : l'intervalle de confiance porte sur la moyenne de la population, et non sur les valeurs individuelles, et la moyenne de l'échantillon se trouve toujours dans l'intervalle (elle en est le centre). Pour la question 9, sur 150 malades, 135 sont positifs, et sur 850 sujets sains, 170 sont faussement positifs ; VPP $= \frac{135}{305} = 44,3\%$. Pour la question 13, changer de test d'après le résultat (option d) relève de la pêche à la significativité, et l'option a utilise la formulation erronée "on accepte H0". Pour la question 14, RR $= \frac{\frac{18}{150}}{\frac{10}{250}} = \frac{0,12}{0,04} = 3,0$; 3,27 est l'OR. Pour la question 21, l'ordonnée à l'origine correspond à un IMC de 0, valeur impossible et située loin du domaine observé (23,5-48,3) ; elle n'a donc pas d'interprétation clinique. Pour la question 24, le seuil de Youden maximise Se + Sp, et non la VPP. Pour la question 29, r = 0 n'exclut qu'une relation linéaire ; une relation curvilinéaire (par exemple en forme de U) peut donner un r proche de 0.

Exercices proposés : 1. Total 250 : non-fumeurs 120 (48,0% du total ; 52,2% des valeurs valides), fumeurs 70 (28,0% ; 30,4%), anciens fumeurs 40 (16,0% ; 17,4%), données manquantes 20 (8,0%) ; les pourcentages valides se calculent sur 230. 2. OR $= \frac{56\times96}{64\times24} = 3{,}5$; IC 95% 1,97-6,21 ; χ^2 = 19,20, p < 0,001. La cote du tabagisme est 3,5 fois plus élevée chez les patients atteints de BPCO. 3. Risque 30/200 = 15% chez les exposés, 15/300 = 5% chez les non-exposés ; RR = 3,0 (IC 95% 1,66-5,43) ; DR = 10 points de pourcentage ; NNH = 10 ; χ^2 = 14,65, p < 0,001. 4. Malades 200 : VP 180, FN 20 ; sujets sains 9 800 : VN 9 310, FP 490. VPP $= \frac{180}{670} = 26{,}9\%$; VPN $= \frac{9310}{9330} = 99{,}8\%$. Avec une prévalence faible, près des trois quarts des résultats positifs sont des faux positifs, bien que le test ait une Sp de 95% ; un résultat positif demande une confirmation, un résultat négatif exclut la maladie presque à coup sûr. 5. ES $= \frac{18}{\sqrt{36}} = 3$ mg/dL ; IC 95% $= 220 \pm 2{,}03 \times 3 = 213{,}9 - 226{,}1$ mg/dL. 6. $R^2 = 0{,}36$; $t = \frac{0{,}6\times\sqrt{25}}{\sqrt{1-0{,}36}} = \frac{3}{0{,}8} = 3{,}75 > 2{,}06$, donc r est significatif (p ≈ 0,001). 7. PAS = 89,47 + 1,435 x 27 = 128,2 mmHg. Un IMC de 15 se situe en dehors du domaine observé (23,5-48,3 kg/m²), et l'extrapolation de la droite en dehors des données n'est pas justifiée. 8. Cote pré-test = 0,30/0,70 = 0,429 ; cote post-test = 0,429 x 6 = 2,57 ; probabilité post-test = 2,57/3,57 = 72%. 9. a) ANOVA à un facteur (avec Tukey) ; b) test de Wilcoxon pour données appariées (le score de douleur est ordinal) ; c) khi-deux (Fisher si les effectifs attendus sont faibles) ; d) corrélation de Pearson (ou de Spearman) et régression linéaire ; e) test de Mann-Whitney ; f) test de Friedman. 10. Modèle : "Le HDL-cholestérol moyen était de 53,9 mg/dL (ÉT 11,7) chez les femmes et de 44,4 mg/dL (ÉT 9,7) chez les hommes. La différence de 9,5 mg/dL (IC 95% 6,2-12,9) était statistiquement significative (test t de Welch, choisi en raison de variances inégales, Levene p = 0,034 ; t = 5,63, ddl 157,7, p < 0,001)." 11. Version correcte : "La prévalence du diabète était de 33,3% chez les patients obèses et de 10,4% chez les non-obèses (χ^2 = 12,70, p < 0,001 ; rapport de prévalences 3,21, IC 95% 1,64-6,31 ; OR 4,32, IC 95% 1,86-10,03). S'agissant d'une étude transversale, le résultat montre une association, et non une relation causale." Erreurs initiales : causalité, confusion de l'OR avec le risque, absence d'IC. 12. Les écarts-types sont presque égaux aux moyennes, pour une variable qui ne peut pas être négative : la distribution est fortement asymétrique. Recommandations : description par la médiane et l'EIQ, comparaison par le test de Mann-Whitney (ou test t sur données log-transformées), présentation de la différence des médianes.

Annexe B. Glossaire

Le glossaire réunit les termes employés dans le cours. Les entrées sont classées dans l'ordre alphabétique des termes français ; chaque terme est suivi, entre parenthèses, de l'abréviation utilisée dans l'ouvrage, lorsqu'elle existe, du terme roumain et du terme anglais (utile pour la lecture des articles et pour le travail avec les logiciels statistiques, dont l'interface est en anglais), puis d'une définition courte et du chapitre où la notion est expliquée en détail. Les définitions sont formulées pour la révision ; elles ne remplacent pas les explications, les exemples et les pièges discutés dans les chapitres.

- Absence de preuve (roum. "absența dovezii" ; angl. absence of evidence) : un résultat non significatif ne prouve pas que l'effet est absent ; il peut simplement refléter une étude trop petite. L'intervalle de confiance montre quels effets restent compatibles avec les données. (Chap. 5)
- Aire sous la courbe ROC (AUC ; roum. "aria de sub curba ROC" ; angl. area under the curve) : indicateur global de la capacité de discrimination du test, entre 0,5 (test inutile) et 1 (test parfait). C'est la probabilité qu'un malade choisi au hasard ait un résultat plus "anormal" qu'un non-malade choisi au hasard. (Chap. 12)
- Algorithme de choix du test (roum. "algoritmul de alegere a testului" ; angl. test selection algorithm) : suite de questions qui conduit au test approprié : de quel type est la variable de résultat, combien de groupes ou de mesures sont comparés, sont-ils indépendants ou appariés, les conditions du test paramétrique sont-elles remplies. (Chap. 8)
- Analyse de la variance (ANOVA ; roum. "analiza varianței" ; angl. analysis of variance) : test qui compare les moyennes de trois groupes indépendants ou plus, en décomposant la variabilité totale en variabilité intergroupes et variabilité intragroupe. Un résultat significatif montre seulement qu'au moins une moyenne diffère. (Chap. 6)
- Analyse en intention de traiter (ITT ; roum. "analiza după intenția de tratament" ; angl. intention-to-treat analysis) : analyse dans laquelle chaque participant est inclus dans le bras auquel il a été attribué par randomisation, qu'il ait reçu ou non l'intervention. Elle préserve les avantages de la randomisation et constitue l'analyse principale des études de supériorité. (Chap. 13)
- Analyse intégrée des données (roum. "analiza integrată a datelor" ; angl. integrated data analysis) : traitement complet d'un jeu de données, du nettoyage et de la statistique descriptive jusqu'au choix des tests, à l'interprétation et à la présentation des résultats, dans un rapport cohérent qui répond à la question de recherche. (Chap. 13)
- Analyse intermédiaire (roum. "analiză interimară" ; angl. interim analysis) : analyse effectuée avant la fin de l'étude, selon des règles fixées dans le protocole, qui peut conduire à un arrêt précoce pour bénéfice, absence d'effet ou risque. (Chap. 13)
- Analyse per protocole (PP ; roum. "analiza per protocol" ; angl. per-protocol analysis) : analyse limitée aux participants qui ont respecté le protocole. Elle peut surestimer l'effet, car ceux qui abandonnent diffèrent souvent de ceux qui restent. (Chap. 13)
- Anonymisation (roum. "anonimizare" ; angl. anonymisation) : traitement des données de sorte que la personne ne puisse plus être identifiée par aucun moyen raisonnable. Les données véritablement anonymes ne sont plus des données à caractère personnel. (Chap. 1)
- ANOVA à mesures répétées (roum. "ANOVA cu măsurători repetate" ; angl. repeated measures ANOVA) : extension du test t apparié à trois mesures ou plus chez les mêmes individus (par exemple la glycémie à 0, 3 et 6 mois). Elle suppose, entre autres, la sphéricité. (Chap. 6)
- Aplatissement (kurtosis ; roum. "exces" ; angl. kurtosis) : mesure du "poids" des queues d'une distribution par rapport à la loi normale. Un excès d'aplatissement positif indique davantage de valeurs extrêmes qu'on ne l'attendrait pour une distribution normale. (Chap. 3)

- Asymétrie (roum. "asimetrie" ; angl. skewness) : mesure de l'absence de symétrie d'une distribution. L'asymétrie positive (queue vers la droite) est fréquente pour les variables biologiques qui ne peuvent pas descendre sous zéro, comme les triglycérides ; la moyenne dépasse alors la médiane. (Chap. 3)
- Biais d'information (roum. "bias de informație" ; angl. information bias) : erreur systématique dans la mesure de l'exposition ou du résultat, par exemple du fait d'instruments non étalonnés ou de questions formulées différemment pour les cas et pour les témoins. (Chap. 10)
- Biais de mémorisation (roum. "bias de memorie" ; angl. recall bias) : forme de biais d'information propre aux études cas-témoins, dans laquelle les malades se souviennent de leurs expositions passées mieux (ou autrement) que les témoins sains. (Chap. 10)
- Biais de sélection (roum. "bias de selecție" ; angl. selection bias) : erreur systématique qui apparaît lorsque les individus inclus diffèrent systématiquement de la population cible ou lorsque les groupes comparés sont constitués différemment (par exemple, les volontaires sont en meilleure santé que la moyenne). (Chap. 10)
- Biais de spectre (roum. "bias de spectru" ; angl. spectrum bias) : estimation erronée de la sensibilité et de la spécificité lorsque les patients étudiés ne reflètent pas l'éventail de ceux rencontrés en pratique (par exemple uniquement des cas sévères et des volontaires en parfaite santé). (Chap. 12)
- Biais de vérification (roum. "bias de verificare" ; angl. verification bias) : erreur systématique des études d'exactitude diagnostique qui apparaît lorsque le standard de référence n'est appliqué qu'à une partie des patients, habituellement à ceux dont le test est positif. (Chap. 12)
- Biostatistique (roum. "biostatistică" ; angl. biostatistics) : branche de la statistique qui s'occupe de la collecte, de la description, de l'analyse et de l'interprétation des données de la biologie et de la médecine. Elle part de la variabilité biologique et offre des méthodes grâce auxquelles l'incertitude est mesurée au lieu d'être ignorée. (Chap. 1)
- Bonnes pratiques cliniques (BPC ; roum. "buna practică în studiul clinic" ; angl. Good Clinical Practice, GCP) : norme internationale d'éthique et de qualité scientifique pour la conception, la conduite, l'enregistrement et la publication des résultats des essais cliniques, décrite dans la ligne directrice ICH E6. (Chap. 13)
- Box-plot (boîte à moustaches ; roum. "box-plot" ; angl. box-and-whisker plot) : graphique qui montre la médiane, les quartiles, les moustaches jusqu'aux valeurs non aberrantes les plus éloignées et, séparément, les valeurs aberrantes. Il est idéal pour comparer les distributions de plusieurs groupes. (Chap. 2)
- Bras de l'étude (roum. "braț (al studiului)" ; angl. study arm) : chacun des groupes d'intervention d'un essai clinique (par exemple les bras A, B et C du jeu de données mg2026). (Chap. 13)
- Cahier d'observation (CRF ; roum. "formular de raportare a cazului" ; angl. case report form, CRF) : formulaire, papier ou électronique, dans lequel on enregistre les données de chaque participant exigées par le protocole. Un formulaire bien conçu prévient une grande partie des erreurs de saisie. (Chap. 13)
- Causalité (roum. "cauzalitate" ; angl. causality) : relation dans laquelle la modification de l'exposition entraîne la modification du résultat. Elle se démontre de la façon la plus convaincante par des essais randomisés ; dans les études observationnelles, on l'argumente à l'aide de critères tels que ceux de Bradford Hill. (Chap. 10)
- Classificateur (roum. "clasificator" ; angl. classifier) : règle ou modèle qui attribue chaque individu à l'une des classes (malade ou sain) à partir de certaines caractéristiques ; sa performance se décrit par la sensibilité, la spécificité et la courbe ROC. (Chap. 12)
- Codage (roum. "codificare" ; angl. coding) : transformation des réponses en valeurs standardisées (par exemple 0/1 pour absent/présent, 1-4 pour le niveau d'études), décrite dans le dictionnaire des données. (Chap. 2)

- Coefficient de corrélation de Pearson (r ; roum. "coeficientul de corelație Pearson" ; angl. Pearson correlation coefficient) : mesure de l'intensité et du sens de la relation linéaire entre deux variables numériques, avec des valeurs comprises entre -1 et 1. Il suppose la linéarité, des distributions approximativement normales et l'absence de valeurs extrêmes influentes. (Chap. 9)
- Coefficient de corrélation de Spearman (rho ; roum. "coeficientul de corelație Spearman" ; angl. Spearman rank correlation) : coefficient de Pearson calculé sur les rangs ; il mesure une relation monotone (pas nécessairement linéaire). On l'utilise pour les variables ordinales ou pour les variables numériques non normales. (Chap. 9)
- Coefficient de détermination (R^2 ; roum. "coeficient de determinare" ; angl. coefficient of determination) : proportion de la variabilité de la variable dépendante expliquée par le modèle ; en régression simple, il est égal à r^2. Il ne dit rien du sens de la relation. (Chap. 9)
- Coefficient de détermination ajusté (R^2 ajusté ; roum. "coeficient de determinare ajustat" ; angl. adjusted R-squared) : variante de R^2 qui pénalise l'ajout de prédicteurs inutiles. On l'utilise pour comparer des modèles comportant un nombre différent de variables. (Chap. 9)
- Coefficient de variation (CV ; roum. "coeficient de variație" ; angl. coefficient of variation) : écart-type exprimé en pourcentage de la moyenne (100 x s/m). Il permet de comparer la variabilité de mesures d'unités ou d'ordres de grandeur différents. (Chap. 2)
- Coefficient kappa de Cohen (κ ; roum. "coeficientul kappa al lui Cohen" ; angl. Cohen's kappa) : mesure de l'accord entre deux évaluateurs pour une variable catégorielle, corrigée de l'accord dû au hasard. Les valeurs supérieures à 0,60 indiquent un accord important, celles supérieures à 0,80 un accord presque parfait. (Chap. 12)
- Coefficient tau de Kendall (roum. "coeficientul tau Kendall" ; angl. Kendall's tau) : coefficient de corrélation non paramétrique fondé sur le nombre de paires concordantes et discordantes. Certains auteurs le préfèrent pour les petits échantillons et en présence de nombreuses valeurs égales. (Chap. 9)
- Comité d'éthique de la recherche (roum. "comisie de etică a cercetării" ; angl. research ethics committee) : organisme indépendant qui évalue et approuve le protocole d'une étude chez l'être humain avant son démarrage, en vérifiant le respect des droits, de la sécurité et du bien-être des participants. (Chap. 13)
- Comité indépendant de surveillance des données (roum. "comitet independent de monitorizare a datelor" ; angl. data monitoring committee) : groupe d'experts indépendants du promoteur et des investigateurs qui suit les données de sécurité et d'efficacité au cours de l'étude et recommande sa poursuite, sa modification ou son arrêt. (Chap. 13)
- Comparaisons multiples (roum. "comparații multiple" ; angl. multiple comparisons) : réalisation de plusieurs tests sur les mêmes données, ce qui augmente la probabilité d'obtenir au moins un résultat faussement positif. On les contrôle par des corrections (Bonferroni) ou par des tests post-hoc. (Chap. 5)
- Concordance entre méthodes (roum. "acord între metode" ; angl. method agreement) : mesure dans laquelle deux méthodes de mesure donnent les mêmes valeurs chez les mêmes patients. Pour les variables numériques, elle s'évalue par la méthode de Bland-Altman, et non par le coefficient de corrélation. (Chap. 12)
- Consentement éclairé (roum. "consimțământ informat" ; angl. informed consent) : accord libre, donné par écrit, d'une personne pour participer à une étude, après avoir reçu des informations claires sur le but, les procédures, les risques, les bénéfices et le droit de se retirer à tout moment. (Chap. 1)
- Correction de Bonferroni (roum. "corecția Bonferroni" ; angl. Bonferroni correction) : division du seuil α par le nombre de comparaisons effectuées (α/k). Elle est simple et conservatrice. (Chap. 5)
- Correction de Yates (correction de continuité ; roum. "corecția Yates" ; angl. Yates continuity correction) : ajustement du test du khi-deux pour les tableaux 2 x 2, qui diminue légèrement la

valeur de χ^2. Elle est conservatrice, et l'on préfère souvent aujourd'hui le test exact de Fisher lorsque les effectifs sont faibles. (Chap. 7)

- Corrélation (roum. "corelație" ; angl. correlation) : association entre deux variables, caractérisée par son sens (positif ou négatif) et son intensité. Une corrélation ne prouve pas une relation de cause à effet. (Chap. 9)
- Corrélation bisériale de rangs (roum. "corelația biserială a rangurilor" ; angl. rank-biserial correlation) : taille d'effet du test de Mann-Whitney, comprise entre -1 et 1 ; elle indique avec quelle fréquence les valeurs d'un groupe dépassent celles de l'autre. (Chap. 7)
- Corrélation fallacieuse (corrélation factice ; roum. "corelație falsă" ; angl. spurious correlation) : association statistique entre deux variables sans lien direct, produite par une troisième variable commune ou par le hasard. (Chap. 9)
- Cote (odds ; roum. "cotă" ; angl. odds) : rapport entre la probabilité qu'un événement se produise et la probabilité qu'il ne se produise pas, p/(1 - p). Une probabilité de 0,20 correspond à une cote de 0,25 (1 contre 4). (Chap. 10)
- Cote pré-test et cote post-test (roum. "cota pre-test și cota post-test" ; angl. pre-test odds, post-test odds) : probabilités pré-test et post-test exprimées sous forme de cotes ; cote post-test = cote pré-test x rapport de vraisemblance. (Chap. 11)
- Courbe ROC (roum. "curba ROC" ; angl. receiver operating characteristic curve) : graphique de la sensibilité en fonction de 1 - spécificité pour tous les seuils possibles d'un test numérique. Plus la courbe se rapproche du coin supérieur gauche, meilleur est le test. (Chap. 12)
- Critère de jugement principal (roum. "criteriul principal de evaluare" ; angl. primary outcome, primary endpoint) : résultat principal à partir duquel on juge l'efficacité de l'intervention et on calcule la taille de l'échantillon. Il est fixé dans le protocole, avant la collecte des données. (Chap. 13)
- Critère de jugement secondaire (roum. "criteriu secundar de evaluare" ; angl. secondary outcome) : résultat supplémentaire, analysé pour compléter le tableau des effets de l'intervention ; les conclusions fondées uniquement sur des critères secondaires ont un caractère exploratoire. (Chap. 13)
- Critère de substitution (roum. "criteriu surogat" ; angl. surrogate endpoint) : résultat intermédiaire (pression artérielle, HbA1c, cholestérol) utilisé à la place d'un résultat clinique important (infarctus, décès), en supposant que sa modification prédit l'effet clinique. (Chap. 13)
- Critères d'inclusion et d'exclusion (roum. "criterii de includere și de excludere" ; angl. eligibility criteria) : conditions qu'une personne doit remplir (ou ne pas remplir) pour pouvoir participer à l'étude. Ils définissent la population à laquelle les résultats peuvent s'appliquer. (Chap. 13)
- Critères de Bradford Hill (roum. "criteriile Bradford Hill" ; angl. Bradford Hill criteria) : repères pour juger du caractère causal d'une association : force de l'association, constance, spécificité, séquence temporelle, gradient dose-réponse, plausibilité biologique, cohérence, preuves expérimentales et analogie. Parmi eux, seule la séquence temporelle est obligatoire. (Chap. 10)
- d de Cohen (roum. "Cohen d" ; angl. Cohen's d) : différence entre deux moyennes exprimée en écarts-types. Valeurs indicatives : 0,2 effet faible, 0,5 moyen, 0,8 fort. (Chap. 6)
- Déclaration d'Helsinki (roum. "Declarația de la Helsinki" ; angl. Declaration of Helsinki) : ensemble de principes éthiques pour la recherche médicale impliquant des êtres humains, adopté par l'Association médicale mondiale et révisé périodiquement (dernière révision en 2024). (Chap. 13)
- Degrés de liberté (ddl ; roum. "grade de libertate" ; angl. degrees of freedom, df) : nombre de valeurs qui peuvent varier librement une fois certaines contraintes fixées (par exemple la moyenne). Ils déterminent la forme des lois t, du khi-deux et F ; pour le test t sur un échantillon, ddl = n - 1. (Chap. 4)

- Densité d'incidence (taux d'incidence ; roum. "densitate de incidență" ; angl. incidence rate) : nombre de nouveaux cas rapporté à la somme des temps de suivi des personnes à risque (personnes-années). Elle tient compte du fait que les participants sont suivis pendant des durées différentes. (Chap. 10)
- Diagramme circulaire (roum. "diagramă circulară" ; angl. pie chart) : graphique dans lequel un cercle est divisé en secteurs proportionnels aux fréquences des catégories. Il n'est lisible que pour un petit nombre de catégories et ne s'utilise pas pour les variables numériques. (Chap. 2)
- Diagramme de dispersion (nuage de points ; roum. "diagramă de dispersie" ; angl. scatter plot) : graphique dans lequel chaque individu est un point dont les coordonnées sont données par deux variables numériques. Il est obligatoire avant le calcul d'un coefficient de corrélation. (Chap. 9)
- Diagramme de flux (roum. "diagramă de flux" ; angl. flow diagram) : schéma qui indique le nombre de participants à chaque étape de l'étude ainsi que les motifs des exclusions et des pertes. (Chap. 13)
- Diagramme en barres (roum. "diagramă cu bare" ; angl. bar chart) : graphique pour variables catégorielles, dans lequel la longueur de chaque barre est proportionnelle à la fréquence de la catégorie. Les barres sont séparées, contrairement à l'histogramme. (Chap. 2)
- Dichotomisation (roum. "dihotomizare" ; angl. dichotomisation) : transformation d'une variable numérique en une variable à deux catégories, au moyen d'un seuil (par exemple obèse si IMC $\geq$ 30). Elle simplifie l'interprétation, mais entraîne une perte d'information et diminue la puissance statistique. (Chap. 2)
- Dictionnaire des données (roum. "dicționar de date" ; angl. data dictionary) : document qui décrit chaque variable d'un fichier : nom, signification, type, échelle, unité de mesure ou codage, domaine valide et mode de notation des valeurs manquantes. (Chap. 13)
- Différence de risques (DR ; roum. "diferența de risc" ; angl. risk difference) : différence absolue entre le risque chez les exposés (ou les traités) et le risque chez les non-exposés (ou les témoins). Elle exprime l'effet en points de pourcentage et sert de base au NST. (Chap. 10)
- Différence minimale cliniquement importante (roum. "diferența minimă importantă clinic" ; angl. minimal clinically important difference, MCID) : plus petite différence d'un résultat que les patients ou les médecins jugent pertinente. Elle est fixée lors de la planification de l'étude et sert au calcul de la taille de l'échantillon. (Chap. 5)
- Dissimulation de l'allocation (roum. "ascunderea alocării" ; angl. allocation concealment) : mesures qui empêchent la personne qui inclut le patient de connaître à l'avance le bras auquel il sera attribué (enveloppes scellées opaques, randomisation centralisée). Elle est distincte de l'insu. (Chap. 13)
- Distribution bimodale (roum. "distribuție bimodală" ; angl. bimodal distribution) : distribution à deux pics, qui suggère souvent le mélange de deux sous-populations (par exemple la taille des hommes et des femmes considérés ensemble). (Chap. 2)
- Distribution d'échantillonnage (roum. "distribuție de eșantionare" ; angl. sampling distribution) : distribution qu'aurait une statistique (par exemple la moyenne) si l'on tirait un très grand nombre d'échantillons de même taille dans la même population. Son écart-type est l'erreur standard. (Chap. 4)
- Données à caractère personnel (roum. "date cu caracter personal" ; angl. personal data) : toute information se rapportant à une personne identifiée ou identifiable. Les données de santé font partie des catégories particulières, bénéficiant d'une protection renforcée en vertu du Règlement général sur la protection des données. (Chap. 1)
- Données appariées (données liées ; roum. "date pereche" ; angl. paired data) : observations qui forment des paires naturelles : le même individu mesuré deux fois, les deux yeux d'un même patient, des jumeaux ou des patients appariés par paires. On les analyse avec des tests pour données dépendantes. (Chap. 6)

- Écart absolu moyen (roum. "abatere medie absolută" ; angl. mean absolute deviation) : moyenne des valeurs absolues des écarts à la moyenne. Il est intuitif, mais on l'utilise moins souvent que l'écart-type. (Chap. 2)
- Écart interquartile (EIQ ; roum. "interval intercuartilic" ; angl. interquartile range, IQR) : différence Q3 - Q1, qui comprend les 50% centraux des valeurs. Il accompagne la médiane dans la description des distributions asymétriques et sert de base à la règle de détection des valeurs aberrantes sur le box-plot. (Chap. 2)
- Écart-type (ÉT, DS ; roum. "abatere standard" ; angl. standard deviation, SD) : racine carrée de la variance ; il indique à quel point les valeurs individuelles sont dispersées autour de la moyenne, dans l'unité de mesure de la variable. Il ne diminue pas lorsque la taille de l'échantillon augmente. (Chap. 2)
- Échantillon (roum. "eșantion" ; angl. sample) : sous-ensemble d'individus de la population sur lequel on effectue réellement les mesures. La qualité des conclusions dépend davantage de la façon dont l'échantillon a été choisi que de sa taille. (Chap. 1)
- Échantillon représentatif (roum. "eșantion reprezentativ" ; angl. representative sample) : échantillon qui reproduit la structure de la population (selon l'âge, le sexe, le milieu de résidence, etc.) dans des proportions proches des proportions réelles. On l'obtient par une sélection aléatoire, dans laquelle chaque individu a la même chance d'être choisi et les choix sont indépendants. (Chap. 1)
- Échantillonnage aléatoire simple (roum. "eșantionare aleatoare simplă" ; angl. simple random sampling) : méthode de sélection dans laquelle chaque individu de la population a la même probabilité d'entrer dans l'échantillon, par exemple par tirage au sort avec une table de nombres aléatoires ou avec un générateur de nombres aléatoires. (Chap. 1)
- Échantillonnage de convenance (roum. "eșantionare de conveniență" ; angl. convenience sampling) : inclusion des patients facilement disponibles (ceux qui viennent au cabinet pendant une certaine période, les répondants d'un questionnaire en ligne). Il est fréquent, mais ne garantit pas la représentativité. (Chap. 1)
- Échantillonnage en grappes (roum. "eșantionare pe grupuri" ; angl. cluster sampling) : on tire au hasard des groupes entiers (cabinets, écoles, villages), puis on étudie à l'intérieur de ceux-ci tous les individus ou un sous-échantillon. Il est pratique sur le terrain, mais les individus d'un même groupe se ressemblent, ce qui diminue la précision. (Chap. 1)
- Échantillonnage stratifié (roum. "eșantionare stratificată" ; angl. stratified sampling) : la population est divisée en strates homogènes (par exemple urbaine et rurale), et l'on tire au hasard un échantillon dans chaque strate, généralement proportionnel à la taille de la strate. Il augmente la précision et garantit la présence de toutes les strates. (Chap. 1)
- Échantillonnage systématique (roum. "eșantionare sistematică" ; angl. systematic sampling) : on choisit au hasard un point de départ, puis chaque k-ième individu d'une liste ordonnée (par exemple chaque dixième dossier). Il est simple, mais peut introduire des erreurs si la liste présente une périodicité cachée. (Chap. 1)
- Échantillons indépendants (roum. "eșantioane independente" ; angl. independent samples) : groupes formés d'individus différents, dans lesquels les valeurs d'un groupe ne sont pas liées aux valeurs de l'autre. (Chap. 6)
- Échelle d'intervalle (roum. "scală de interval" ; angl. interval scale) : échelle numérique à distances égales entre les valeurs, mais sans zéro absolu (la température en degrés Celsius). Les différences ont un sens, les rapports non. (Chap. 2)
- Échelle de Likert (roum. "scală Likert" ; angl. Likert scale) : échelle de réponse ordonnée, habituellement à 5 ou 7 niveaux (de "pas du tout d'accord" à "tout à fait d'accord"), utilisée dans les questionnaires. Les données obtenues sont ordinales et s'analysent de préférence par des tests non paramétriques. (Chap. 7)

- Échelle de rapport (roum. "scală de raport" ; angl. ratio scale) : échelle numérique à distances égales et à zéro absolu (le poids, la concentration d'une substance), pour laquelle les rapports ont aussi un sens : 100 kg représentent deux fois plus que 50 kg. (Chap. 2)
- Epi Info (roum. "Epi Info" ; angl. Epi Info) : logiciel gratuit développé par les Centers for Disease Control and Prevention des États-Unis, utilisé en épidémiologie pour les formulaires, les tableaux 2 x 2, le RR, l'OR et le test du khi-deux. (Chap. 10)
- Épidémiologie (roum. "epidemiologie" ; angl. epidemiology) : science qui étudie la fréquence, la répartition et les déterminants des maladies dans les populations et qui applique ces connaissances à la prévention et à la maîtrise des maladies. (Chap. 10)
- Erreur aléatoire (roum. "eroare aleatoare" ; angl. random error) : variation due au hasard des résultats d'un échantillon à l'autre, sans direction privilégiée. Elle diminue lorsque la taille de l'échantillon augmente et se quantifie par l'erreur standard et par l'intervalle de confiance. (Chap. 4)
- Erreur de type I (roum. "eroare de tip I" ; angl. type I error) : rejet de l'hypothèse nulle alors qu'elle est vraie (résultat faussement positif). Sa probabilité est α. (Chap. 5)
- Erreur de type II (roum. "eroare de tip II" ; angl. type II error) : non-rejet de l'hypothèse nulle alors qu'elle est fausse (résultat faussement négatif, effet réel non détecté). Sa probabilité est β. (Chap. 5)
- Erreur standard d'une proportion (roum. "eroarea standard a proporției" ; angl. standard error of a proportion) : $\sqrt{\frac{p_i(1-p_i)}{n}}$; elle mesure la précision avec laquelle une proportion observée estime la proportion dans la population. (Chap. 4)
- Erreur standard de l'estimation (SEE ; roum. "eroarea standard a estimării" ; angl. standard error of the estimate) : écart-type des résidus, exprimé dans l'unité de la variable dépendante ; elle indique à quelle distance de la droite se trouvent, en moyenne, les points. (Chap. 9)
- Erreur standard de la moyenne (ES ; roum. "eroarea standard a mediei" ; angl. standard error of the mean, SEM) : écart-type de la distribution des moyennes des échantillons, estimé par $\frac{s}{\sqrt{n}}$. Elle mesure la précision de la moyenne et diminue lorsque la taille de l'échantillon augmente. (Chap. 4)
- Erreur systématique (biais ; roum. "eroare sistematică" ; angl. bias) : écart constant, dans le même sens, des résultats par rapport à la valeur réelle, produit par des défauts de sélection, de mesure ou d'analyse. Contrairement à l'erreur aléatoire, elle ne diminue pas lorsque la taille de l'échantillon augmente. (Chap. 1)
- Essai clinique (roum. "studiu clinic" ; angl. clinical trial) : étude expérimentale chez l'être humain qui évalue les effets d'une intervention (médicament, dispositif, procédure, programme éducatif) sur la santé. Il se déroule selon un protocole approuvé par le comité d'éthique. (Chap. 13)
- Essai contrôlé randomisé (ECR ; roum. "studiu clinic randomizat" ; angl. randomised controlled trial, RCT) : étude dans laquelle les participants sont répartis au hasard dans différents groupes d'intervention, ce qui rend les groupes comparables aussi bien pour les facteurs connus que pour les facteurs inconnus. (Chap. 13)
- Essai croisé (roum. "studiu încrucișat" ; angl. crossover trial) : étude dans laquelle chaque participant reçoit successivement les deux interventions, dans un ordre randomisé, séparées par une période d'élimination (wash-out). Chaque patient est son propre témoin, ce qui exige des tests pour données appariées. (Chap. 13)
- Essai d'équivalence (roum. "studiu de echivalență" ; angl. equivalence trial) : étude qui vise à montrer que deux interventions ne diffèrent, dans un sens comme dans l'autre, pas plus que d'une marge préétablie. (Chap. 13)
- Essai de non-infériorité (roum. "studiu de non-inferioritate" ; angl. non-inferiority trial) : étude qui vise à montrer qu'une nouvelle intervention n'est pas moins bonne que le traitement standard au-delà d'une marge préétablie, habituellement parce qu'elle présente d'autres avantages (coût, sécurité, mode d'administration). (Chap. 13)

- Essai de supériorité (roum. "studiu de superioritate" ; angl. superiority trial) : étude qui vise à montrer qu'une intervention est meilleure que le comparateur. (Chap. 13)
- Essai explicatif (roum. "studiu explicativ" ; angl. explanatory trial) : essai clinique mené dans des conditions idéales, strictement contrôlées, qui répond à la question "l'intervention peut-elle fonctionner ?" (efficacité théorique, et non efficacité en pratique). (Chap. 13)
- Essai pragmatique (roum. "studiu pragmatic" ; angl. pragmatic trial) : essai clinique mené dans des conditions proches de la pratique courante, avec des critères d'inclusion larges, qui répond à la question "l'intervention fonctionne-t-elle dans la vie réelle ?". (Chap. 13)
- Estimateur sans biais (roum. "estimator nedeplasat" ; angl. unbiased estimator) : estimateur dont la valeur moyenne, sur l'ensemble des échantillons possibles, est égale au paramètre estimé. La variance calculée avec le dénominateur n - 1 est un estimateur sans biais de la variance de la population. (Chap. 4)
- Estimation par intervalle (roum. "estimare prin interval" ; angl. interval estimation) : indication d'un domaine de valeurs plausibles pour le paramètre, accompagné d'un niveau de confiance. La forme habituelle est l'intervalle de confiance. (Chap. 4)
- Estimation ponctuelle (roum. "estimare punctuală" ; angl. point estimate) : valeur unique calculée à partir de l'échantillon, utilisée comme meilleure approximation du paramètre (par exemple la PAS moyenne de 131,13 mmHg). (Chap. 4)
- Êta carré (η^2 ; roum. "eta pătrat" ; angl. eta squared) : proportion de la variabilité totale expliquée par le facteur de regroupement en ANOVA (SSB/SST). Valeurs indicatives : 0,01 faible, 0,06 moyen, 0,14 fort. (Chap. 6)
- Étapes de l'étude statistique (roum. "etapele studiului statistic" ; angl. stages of a statistical study) : succession formulation de la question, planification (schéma d'étude, échantillon, variables), collecte des données, vérification et nettoyage, analyse descriptive, analyse inférentielle, interprétation et présentation des résultats. (Chap. 1)
- Étendue (roum. "amplitudine" ; angl. range) : différence entre la valeur maximale et la valeur minimale. Elle est facile à calculer, mais ne dépend que de deux valeurs, elle est donc très sensible aux valeurs extrêmes. (Chap. 2)
- Étude cas-témoins (roum. "studiu caz-control" ; angl. case-control study) : étude qui part de malades (les cas) et de personnes sans la maladie (les témoins) et compare la fréquence de l'exposition antérieure. Elle est efficace pour les maladies rares, mais ne permet que le calcul de l'OR et est exposée au biais de mémorisation. (Chap. 10)
- Étude de cohorte (roum. "studiu de cohortă" ; angl. cohort study) : étude dans laquelle un groupe de personnes sans la maladie, classées selon l'exposition, est suivi dans le temps pour la survenue de la maladie. Elle permet de calculer l'incidence et le RR ; elle peut être prospective ou rétrospective (historique). (Chap. 10)
- Étude écologique (roum. "studiu ecologic" ; angl. ecological study) : étude dans laquelle l'unité d'analyse est un groupe (pays, département), et non l'individu. Les associations observées au niveau du groupe ne peuvent pas être transposées automatiquement aux individus (l'erreur écologique). (Chap. 10)
- Étude expérimentale (roum. "studiu experimental" ; angl. experimental study) : étude dans laquelle le chercheur attribue l'intervention aux participants, idéalement par randomisation. L'essai contrôlé randomisé en est la forme de référence. (Chap. 13)
- Étude observationnelle (roum. "studiu observațional" ; angl. observational study) : étude dans laquelle le chercheur se contente d'observer l'exposition et le résultat, sans les influencer (transversale, de cohorte, cas-témoins). Elle est exposée aux facteurs de confusion. (Chap. 10)
- Étude pilote (roum. "studiu pilot" ; angl. pilot study) : étude de petite taille qui teste la faisabilité des procédures (recrutement, formulaires, observance) avant l'étude principale ; elle n'est pas conçue pour tester l'efficacité. (Chap. 13)

- Étude transversale (roum. "studiu transversal" ; angl. cross-sectional study) : étude dans laquelle l'exposition et la maladie sont mesurées au même moment, comme une photographie de la population. Elle estime la prévalence, mais ne peut pas établir l'ordre chronologique de l'exposition et de la maladie. (Chap. 10)
- Événement (roum. "eveniment" ; angl. event) : résultat possible d'une observation ou d'une expérience (le patient est diabétique, le test est positif). Les événements se combinent par "et" (intersection) et "ou" (réunion). (Chap. 1)
- Événement contraire (roum. "eveniment contrar" ; angl. complementary event) : l'événement "non-A", qui se produit exactement lorsque A ne se produit pas ; $P(\text{non} - A) = 1 - P(A)$. (Chap. 1)
- Événement indésirable (roum. "eveniment advers" ; angl. adverse event) : tout événement médical défavorable survenu chez un participant au cours de l'étude, qu'il soit lié ou non à l'intervention. Les événements graves sont déclarés immédiatement. (Chap. 13)
- Événements incompatibles (roum. "evenimente incompatibile" ; angl. mutually exclusive events) : événements qui ne peuvent pas se produire simultanément (un patient ne peut pas être à la fois fumeur et non-fumeur). Pour eux, $P(A \text{ ou } B) = P(A) + P(B)$. (Chap. 1)
- Événements indépendants (roum. "evenimente independente" ; angl. independent events) : deux événements pour lesquels la survenue de l'un ne modifie pas la probabilité de l'autre ; alors $P(A \text{ et } B) = P(A) \times P(B)$. (Chap. 1)
- Exactitude (roum. "acuratețe" ; angl. accuracy) : proportion des personnes correctement classées par le test, (a + d)/N. Elle dépend de la prévalence et peut être trompeuse lorsque la maladie est rare. (Chap. 12)
- Exposition (roum. "expunere" ; angl. exposure) : facteur dont on étudie l'effet sur la maladie (un comportement, un traitement, un environnement, une caractéristique génétique). Elle peut être ponctuelle ou chronique, auquel cas la dose et la durée comptent. (Chap. 10)
- Extrapolation (roum. "extrapolare" ; angl. extrapolation) : utilisation d'un modèle pour prédire des valeurs en dehors du domaine des valeurs du prédicteur observées dans les données. Elle est risquée, car la relation peut ne plus y être linéaire. (Chap. 9)
- Facteur de confusion (roum. "factor de confuzie" ; angl. confounder) : variable associée à la fois à l'exposition et au résultat, sans être un maillon intermédiaire entre les deux, qui peut créer ou masquer une association. On le contrôle par la randomisation, la restriction, l'appariement, la stratification ou l'ajustement dans des modèles multivariés. (Chap. 10)
- Facteur de risque (roum. "factor de risc" ; angl. risk factor) : caractéristique, comportement ou exposition qui augmente la probabilité de survenue d'une maladie (le tabagisme pour la BPCO, l'obésité pour le diabète). Une association ne signifie pas automatiquement une causalité. (Chap. 10)
- Facteur protecteur (roum. "factor protector" ; angl. protective factor) : exposition associée à une probabilité plus faible de maladie, avec un RR ou un OR inférieur à 1 (l'activité physique pour le diabète). (Chap. 10)
- Faux négatif (FN ; roum. "fals negativ" ; angl. false negative) : malade dont le test est négatif (case b). La conséquence est le retard du diagnostic et du traitement. (Chap. 12)
- Faux positif (FP ; roum. "fals pozitiv" ; angl. false positive) : personne non malade dont le test est positif (case c). Les conséquences sont l'anxiété et des investigations inutiles. (Chap. 12)
- Feuille de calcul (roum. "foaie de calcul" ; angl. spreadsheet) : logiciel de type tableur (Google Sheets, Excel) dans lequel les données sont organisées avec une ligne pour chaque unité statistique et une colonne pour chaque variable. (Chap. 2)
- Forest plot (graphique en forêt ; roum. "grafic forest" ; angl. forest plot) : représentation graphique des estimations (RR, OR, différences de moyennes) et de leurs intervalles de confiance pour plusieurs études ou sous-groupes, avec une ligne verticale à la valeur "absence d'effet". (Chap. 10)

- Fraction attribuable chez les exposés (roum. "fracție atribuibilă la expuși" ; angl. attributable fraction among the exposed) : proportion des cas survenus chez les exposés qui seraient dus à l'exposition, (RR - 1)/RR. (Chap. 10)
- Fraction attribuable dans la population (FAP ; roum. "fracție atribuibilă în populație" ; angl. population attributable fraction) : proportion de l'ensemble des cas de la population qui disparaîtraient si l'exposition était supprimée, en supposant une relation causale. Elle dépend à la fois du RR et de la fréquence de l'exposition. (Chap. 10)
- Fréquence absolue (roum. "frecvență absolută" ; angl. absolute frequency) : nombre d'individus qui présentent une certaine valeur ou appartiennent à une certaine catégorie. Le test du khi-deux se calcule toujours sur les fréquences absolues, jamais sur les pourcentages. (Chap. 2)
- Fréquence attendue (roum. "frecvență așteptată" ; angl. expected frequency) : nombre d'observations qui se trouveraient dans une case si les variables étaient indépendantes, calculé comme total de la ligne x total de la colonne / N. La condition habituelle du test du khi-deux est que les fréquences attendues soient d'au moins 5. (Chap. 7)
- Fréquence cumulée (roum. "frecvență cumulată" ; angl. cumulative frequency) : somme des fréquences de toutes les catégories jusqu'à la catégorie courante incluse. Elle n'a de sens que pour des variables ordonnées (ordinales ou numériques groupées en classes). (Chap. 2)
- Fréquence relative (roum. "frecvență relativă" ; angl. relative frequency) : rapport entre le nombre d'apparitions d'une valeur et le nombre total d'observations, exprimé sous forme de fraction ou de pourcentage. La somme des fréquences relatives de toutes les catégories est égale à 1 (100%). (Chap. 2)
- Fréquence valide (roum. "frecvență validă" ; angl. valid percent) : pourcentage calculé sur le nombre de réponses valides, en excluant les valeurs manquantes. Lorsque des données manquent, le pourcentage valide diffère du pourcentage calculé sur le total, et le rapport doit préciser lequel a été utilisé. (Chap. 2)
- Graphique Q-Q (roum. "grafic Q-Q" ; angl. quantile-quantile plot) : graphique qui met en regard les quantiles observés et les quantiles théoriques de la loi normale. Si les données sont normales, les points se disposent approximativement sur une droite. (Chap. 3)
- Groupe témoin (roum. "grup de control" ; angl. control group) : groupe qui reçoit l'intervention de comparaison (placebo, traitement standard ou aucune intervention), par rapport auquel on mesure l'effet de l'intervention étudiée. (Chap. 13)
- Hétérogénéité (roum. "heterogenitate" ; angl. heterogeneity) : variation des résultats entre les études incluses dans une méta-analyse, supérieure à ce que le hasard expliquerait ; on la quantifie habituellement par l'indice I^2. (Chap. 11)
- Hiérarchie des preuves (roum. "ierarhia dovezilor" ; angl. hierarchy of evidence) : classement des types d'études selon leur capacité à contrôler les erreurs systématiques, depuis l'opinion d'expert et les séries de cas, en passant par les études observationnelles, jusqu'aux essais randomisés et aux revues systématiques avec méta-analyse. (Chap. 11)
- Histogramme (roum. "histogramă" ; angl. histogram) : graphique formé de rectangles accolés qui montre la distribution d'une variable numérique groupée en classes. Sa forme suggère la symétrie, l'asymétrie ou l'existence de plusieurs pics. (Chap. 2)
- Homogénéité des variances (homoscédasticité ; roum. "omogenitatea varianțelor" ; angl. homogeneity of variance) : condition selon laquelle la variabilité est approximativement la même dans les groupes comparés (ou, en régression, le long de la droite). C'est une condition du test t classique, de l'ANOVA et de la régression linéaire. (Chap. 6)
- Hypothèse alternative (H1, Ha ; roum. "ipoteza alternativă" ; angl. alternative hypothesis) : hypothèse qui contredit H0, en affirmant l'existence d'une différence ou d'une association. Elle peut être bilatérale ($\mu 1 \neq \mu 2$) ou unilatérale ($\mu 1 > \mu 2$). (Chap. 5)

- Hypothèse de recherche (roum. "ipoteză de cercetare" ; angl. research hypothesis) : affirmation que le chercheur veut vérifier, formulée en termes cliniques (par exemple, "la téléconsultation diminue la glycémie davantage que le conseil standard"). Les hypothèses statistiques en découlent. (Chap. 5)
- Hypothèse nulle (H0 ; roum. "ipoteza nulă" ; angl. null hypothesis) : hypothèse statistique d'absence de différence ou d'association ($\mu 1 = \mu 2$, RR = 1, $\rho = 0$). Le test calcule dans quelle mesure les données sont compatibles avec elle ; on rejette ou on ne rejette pas H0, mais on ne l'"accepte" pas. (Chap. 5)
- Incidence cumulée (risque ; roum. "incidență cumulativă" ; angl. cumulative incidence) : proportion des personnes initialement saines qui développent la maladie au cours d'une période donnée. Elle s'exprime toujours avec la période (par exemple 3% en 5 ans). (Chap. 10)
- Indicateur robuste (roum. "indicator robust" ; angl. robust statistic) : indicateur peu influencé par les valeurs extrêmes ou par les écarts à la normalité, comme la médiane et l'écart interquartile. (Chap. 2)
- Indice de Youden (J ; roum. "indicele Youden" ; angl. Youden index) : Se + Sp - 1, compris entre 0 et 1. Le seuil qui maximise J est un critère fréquent de choix du seuil optimal, lorsque les coûts des deux types d'erreurs sont considérés comme égaux. (Chap. 12)
- Insu (aveugle, masquage ; roum. "orbire" ; angl. blinding) : dissimulation de l'intervention reçue au participant (simple insu), au participant et à l'investigateur (double insu), ou encore à la personne qui évalue les résultats et analyse les données. Il réduit le biais de performance et le biais d'évaluation. (Chap. 13)
- Intégrité de la recherche (roum. "integritatea cercetării" ; angl. research integrity) : respect de l'honnêteté à toutes les étapes de la recherche. Les manquements graves sont la fabrication des données, leur falsification et le plagiat ; s'y ajoutent des pratiques discutables, comme la présentation sélective des résultats significatifs. (Chap. 1)
- Intervalle de confiance (IC ; roum. "interval de încredere" ; angl. confidence interval, CI) : intervalle calculé à partir de l'échantillon qui, si l'étude était répétée de nombreuses fois, contiendrait la valeur réelle du paramètre dans un pourcentage donné des cas (habituellement 95%). Sa largeur indique la précision de l'estimation. (Chap. 4)
- Intervalle de confiance d'une différence (roum. "interval de încredere pentru o diferență" ; angl. confidence interval for a difference) : IC calculé pour la différence de deux moyennes ou de deux proportions. S'il ne contient pas 0, la différence est significative au seuil correspondant ; pour les rapports (RR, OR), la valeur de référence est 1. (Chap. 4)
- Intervalle de référence (valeurs normales ; roum. "interval de referință" ; angl. reference interval) : intervalle qui contient les 95% centraux des valeurs d'une population en bonne santé, calculé comme la moyenne ± 1,96 écart-type (données normales) ou entre les percentiles 2,5 et 97,5. Il décrit les individus, et non la précision de la moyenne. (Chap. 3)
- Intervalle de Wald pour une proportion (roum. "intervalul Wald pentru proporție" ; angl. Wald interval) : IC calculé comme $p_i \pm z \times \sqrt{\frac{p_i(1-p_i)}{n}}$. Il est simple, mais inexact pour les petits échantillons ou les proportions proches de 0 ou de 1, où il peut même donner des limites négatives. (Chap. 4)
- Intervalle de Wilson pour une proportion (roum. "intervalul Wilson pentru proporție" ; angl. Wilson score interval) : IC d'une proportion aux propriétés meilleures que celles de l'intervalle de Wald, recommandé surtout pour les petits échantillons et les proportions extrêmes. Il n'est pas symétrique autour de p. (Chap. 4)
- JASP (roum. "JASP" ; angl. JASP) : logiciel statistique gratuit, à interface graphique, utilisé lors des travaux pratiques pour la statistique descriptive, les tests t, l'ANOVA, les tests non paramétriques, la corrélation et la régression. (Chap. 2)

- Lecture critique d'un article (roum. "interpretarea critică a unui articol" ; angl. critical appraisal) : évaluation systématique de la validité, des résultats et de l'applicabilité d'une étude publiée : schéma d'étude adapté à la question, contrôle des biais, adéquation des tests, présentation des IC et de la taille de l'effet. (Chap. 14)
- Létalité (roum. "letalitate" ; angl. case fatality rate) : proportion des malades atteints d'une maladie donnée qui en décèdent au cours d'une période donnée. Elle mesure la gravité de la maladie, et non sa fréquence dans la population. (Chap. 10)
- Linéarité (roum. "liniaritate" ; angl. linearity) : condition selon laquelle la relation entre deux variables peut être décrite par une droite. Elle se vérifie sur le diagramme de dispersion et sur le graphique des résidus. (Chap. 9)
- Loi binomiale (roum. "distribuție binomială" ; angl. binomial distribution) : loi du nombre de "succès" en n essais indépendants, chacun ayant la même probabilité p (par exemple le nombre de patients diabétiques dans un échantillon de 20). Pour n grand, on l'approche par la loi normale. (Chap. 3)
- Loi de Fisher (loi F ; roum. "distribuția Fisher" ; angl. F distribution) : loi du rapport de deux variances, définie par deux nombres de degrés de liberté (numérateur et dénominateur). On l'utilise en ANOVA, en régression et dans les tests de comparaison de variances. (Chap. 6)
- Loi de Poisson (roum. "distribuție Poisson" ; angl. Poisson distribution) : loi du nombre d'événements rares survenant dans un intervalle de temps ou d'espace (nombre de nouveaux cas par mois dans un hôpital). Sa moyenne et sa variance sont égales. (Chap. 3)
- Loi de probabilité (roum. "distribuție de probabilitate" ; angl. probability distribution) : description de toutes les valeurs possibles d'une variable aléatoire et des probabilités avec lesquelles elles apparaissent. Pour les variables continues, elle se décrit par une courbe de densité, sous laquelle l'aire totale vaut 1. (Chap. 3)
- Loi de Student (loi t ; roum. "distribuția Student" ; angl. Student t distribution) : loi symétrique, semblable à la loi normale, mais à queues plus épaisses, utilisée lorsque l'écart-type de la population est inconnu. Sa forme dépend des degrés de liberté et se rapproche de la loi normale lorsque ceux-ci augmentent. (Chap. 4)
- Loi des grands nombres (roum. "legea numerelor mari" ; angl. law of large numbers) : à mesure que le nombre d'observations augmente, la fréquence relative d'un événement se rapproche de sa probabilité, et la moyenne de l'échantillon de la moyenne de la population. (Chap. 1)
- Loi log-normale (roum. "distribuție lognormală" ; angl. lognormal distribution) : loi d'une variable dont le logarithme suit une loi normale. Elle est asymétrique à droite et décrit bien de nombreuses concentrations biologiques. (Chap. 3)
- Loi normale (loi de Gauss ; roum. "distribuție normală" ; angl. normal distribution) : loi continue, symétrique, en forme de cloche, entièrement définie par la moyenne et l'écart-type. La moyenne, la médiane et le mode coïncident, et les queues s'étendent théoriquement à l'infini. (Chap. 3)
- Loi normale centrée réduite (roum. "distribuție normală standard" ; angl. standard normal distribution) : loi normale de moyenne 0 et d'écart-type 1, à laquelle on parvient par la transformation z. Ses aires sont tabulées (annexe E). (Chap. 3)
- Marge d'erreur (roum. "marja de eroare" ; angl. margin of error) : demi-largeur de l'intervalle de confiance, c'est-à-dire la distance entre l'estimation ponctuelle et chacune des bornes. C'est la grandeur fixée lors du calcul de la taille de l'échantillon. (Chap. 4)
- MedCalc (roum. "MedCalc" ; angl. MedCalc) : logiciel statistique orienté vers la recherche médicale, utilisé dans le cours en particulier pour les courbes ROC, les tableaux 2 x 2, les risques relatifs et le calcul de la taille de l'échantillon. (Chap. 12)
- Médecine fondée sur les preuves (EBM ; roum. "medicina bazată pe dovezi" ; angl. evidence-based medicine) : utilisation consciencieuse et explicite des meilleures preuves disponibles issues

de la recherche, intégrées à l'expérience clinique et aux valeurs du patient, dans les décisions concernant la prise en charge de chaque patient. (Chap. 11)

- Médiane (roum. "mediana" ; angl. median) : valeur qui divise la série ordonnée en deux moitiés égales en nombre d'observations. Elle n'est pas influencée par les valeurs extrêmes et on la préfère pour les distributions asymétriques et pour les variables ordinales. (Chap. 2)
- Méta-analyse (roum. "meta-analiză" ; angl. meta-analysis) : combinaison statistique des résultats de plusieurs études indépendantes qui répondent à la même question, afin d'obtenir une estimation globale plus précise de l'effet. (Chap. 11)
- Méthode des moindres carrés (roum. "metoda celor mai mici pătrate" ; angl. least squares method) : méthode par laquelle on choisit les coefficients de la droite de régression de sorte que la somme des carrés des résidus soit minimale. (Chap. 9)
- Mode (roum. "moda" ; angl. mode) : valeur ou catégorie la plus fréquente. C'est le seul indicateur de tendance centrale applicable aux variables nominales. (Chap. 2)
- Modification de l'effet (interaction ; roum. "modificare de efect" ; angl. effect modification) : situation dans laquelle l'ampleur de l'association entre l'exposition et la maladie diffère selon le niveau d'une troisième variable (par exemple, l'effet d'un traitement diffère entre les hommes et les femmes). (Chap. 10)
- Morbidité (roum. "morbiditate" ; angl. morbidity) : fréquence des maladies dans une population, exprimée par la prévalence ou l'incidence. (Chap. 10)
- Mortalité (roum. "mortalitate" ; angl. mortality) : fréquence des décès dans une population au cours d'une période, rapportée à l'ensemble de la population (par exemple, décès pour 100 000 habitants par an). (Chap. 10)
- Moyenne arithmétique (roum. "media aritmetică" ; angl. arithmetic mean) : somme des valeurs divisée par leur nombre. C'est l'indicateur de tendance centrale préféré pour les distributions symétriques, mais il est sensible aux valeurs extrêmes. (Chap. 2)
- Moyenne géométrique (roum. "media geometrică" ; angl. geometric mean) : racine n-ième du produit des n valeurs, égale à l'exponentielle de la moyenne des logarithmes. Elle convient aux variables positives présentant une asymétrie à droite, comme les triglycérides ou les titres d'anticorps. (Chap. 2)
- Moyenne pondérée (roum. "media ponderată" ; angl. weighted mean) : moyenne dans laquelle chaque valeur contribue proportionnellement à un poids, par exemple l'effectif du sous-groupe. On l'utilise pour combiner les moyennes de groupes de tailles différentes. (Chap. 2)
- Multicolinéarité (roum. "multicoliniaritate" ; angl. multicollinearity) : forte corrélation entre les prédicteurs d'un modèle de régression, qui rend les coefficients instables et difficiles à interpréter (par exemple le poids et l'IMC dans le même modèle). (Chap. 9)
- Nettoyage des données (roum. "curățarea datelor" ; angl. data cleaning) : vérification systématique du fichier avant l'analyse : catégories codées de façon non uniforme, valeurs impossibles, texte dans des colonnes numériques, valeurs manquantes, doublons, incohérences entre variables. Chaque correction se fait d'après le document source et est consignée dans un journal. (Chap. 2)
- Niveau de confiance (roum. "nivel de încredere" ; angl. confidence level) : pourcentage (1 - α) x 100 associé à un intervalle de confiance, habituellement 95% ou 99%. Un niveau plus élevé produit un intervalle plus large. (Chap. 4)
- Nombre de sujets à traiter (NST ; roum. "numărul necesar de tratat" ; angl. number needed to treat, NNT) : nombre de patients qu'il faut traiter pour qu'un patient supplémentaire en tire un bénéfice, égal à 1/RAR et arrondi à l'entier supérieur. Un NST faible indique un traitement efficace. (Chap. 10)
- Nombre de sujets à traiter pour nuire (NNH ; roum. "numărul necesar pentru a produce un efect advers" ; angl. number needed to harm) : nombre de patients qu'il faut traiter pour qu'un patient

supplémentaire subisse un effet indésirable, égal à l'inverse de l'augmentation absolue du risque d'effet indésirable. (Chap. 10)

- Nomogramme de Fagan (roum. "nomograma Fagan" ; angl. Fagan nomogram) : graphique à trois axes (probabilité pré-test, rapport de vraisemblance, probabilité post-test) sur lequel une ligne droite réalise, sans calcul, la conversion bayésienne. (Chap. 11)
- Normalité (roum. "normalitate" ; angl. normality) : propriété des données de suivre, approximativement, une loi normale. C'est une condition des tests paramétriques ; elle se vérifie graphiquement (histogramme, graphique Q-Q) et par des tests (Shapiro-Wilk). (Chap. 3)
- Ordonnée à l'origine (a ; roum. "ordonata la origine" ; angl. intercept) : valeur prédite de y lorsque x = 0. Elle n'a souvent pas d'interprétation clinique, car x = 0 se situe en dehors des données (un IMC de 0 n'existe pas). (Chap. 9)
- Paramètre (roum. "parametru" ; angl. parameter) : valeur numérique qui décrit la population (la moyenne μ, l'écart-type σ, la proportion π). En règle générale, il est inconnu et on l'estime à partir de l'échantillon. (Chap. 4)
- Pente de la droite de régression (coefficient de régression, b ; roum. "panta dreptei de regresie" ; angl. slope, regression coefficient) : variation moyenne de y lorsque x augmente d'une unité. Son signe coïncide avec celui du coefficient de corrélation. (Chap. 9)
- Percentile (roum. "percentilă" ; angl. percentile) : valeur au-dessous de laquelle se trouve un certain pourcentage des observations ; le 90e percentile laisse au-dessous de lui 90% des valeurs. Les courbes de croissance en pédiatrie sont construites sur des percentiles. (Chap. 2)
- Personnes-temps (roum. "persoane-timp" ; angl. person-time) : somme des temps de suivi de tous les participants jusqu'à la maladie, la perte de vue ou la fin de l'étude. C'est le dénominateur de la densité d'incidence. (Chap. 10)
- Perte de vue (roum. "pierdere din urmărire" ; angl. loss to follow-up) : situation dans laquelle le résultat d'un participant ne peut plus être mesuré (il a déménagé, il ne répond plus). Des pertes importantes ou inégales entre les bras peuvent introduire un biais. (Chap. 13)
- Phases des essais cliniques (roum. "fazele studiilor clinice" ; angl. clinical trial phases) : étapes du développement d'un médicament : phase I (sécurité, dose, peu de volontaires), phase II (efficacité préliminaire), phase III (comparaison randomisée sur de grands effectifs), phase IV (surveillance après l'autorisation de mise sur le marché). (Chap. 13)
- Placebo (roum. "placebo" ; angl. placebo) : intervention inactive, d'apparence identique au traitement actif, utilisée dans le groupe témoin pour permettre l'insu et pour séparer l'effet spécifique de l'effet placebo. (Chap. 13)
- Plan d'analyse statistique (PAS ; roum. "plan de analiză statistică" ; angl. statistical analysis plan, SAP) : document qui fixe à l'avance les méthodes statistiques pour chaque critère de jugement, le traitement des valeurs manquantes et les analyses en sous-groupes, protégeant l'étude des choix faits après avoir vu les données. (Chap. 13)
- Point influent (roum. "valoare influentă" ; angl. influential point) : observation dont la suppression modifie beaucoup le coefficient de corrélation ou la droite de régression, habituellement un point extrême sur l'axe des x. (Chap. 9)
- Population (roum. "populație" ; angl. population) : ensemble de tous les individus qui partagent au moins une caractéristique commune et sur lesquels nous voulons tirer des conclusions, délimité dans l'espace et dans le temps. Exemple : tous les adultes à risque cardiométabolique inscrits chez les médecins de famille d'une région. (Chap. 1)
- Population cible (roum. "populație țintă" ; angl. target population) : population à laquelle sont destinées les conclusions de l'étude, souvent plus large que la population dans laquelle l'échantillon a effectivement été tiré (la population source). L'écart entre les deux limite la validité externe. (Chap. 1)

- Précision (roum. "precizie" ; angl. precision) : degré de proximité entre les résultats de mesures ou d'estimations répétées ; en estimation, elle se reflète dans la largeur de l'intervalle de confiance. Une estimation peut être précise mais biaisée, s'il existe un biais. (Chap. 4)
- Prédiction (roum. "predicție" ; angl. prediction) : estimation de la valeur de la variable dépendante pour un nouvel individu, à partir du modèle. L'intervalle de prédiction pour un individu est beaucoup plus large que l'intervalle de confiance de la moyenne de y. (Chap. 9)
- Présentation sélective des résultats (roum. "raportare selectivă" ; angl. selective reporting) : présentation des seuls résultats significatifs ou favorables parmi toutes les analyses effectuées. Elle déforme la littérature et se prévient par l'enregistrement préalable des études. (Chap. 5)
- Prévalence (roum. "prevalență" ; angl. prevalence) : proportion des personnes d'une population qui ont la maladie à un moment donné (prévalence instantanée) ou au cours d'une période (prévalence de période). Elle inclut aussi bien les nouveaux cas que les cas anciens. (Chap. 10)
- Probabilité (roum. "probabilitate" ; angl. probability) : nombre compris entre 0 et 1 qui exprime à quel point un événement est possible ; 0 signifie impossible, 1 certain. On peut l'estimer par la fréquence relative de l'événement sur un grand nombre d'observations. (Chap. 1)
- Probabilité conditionnelle (roum. "probabilitate condiționată" ; angl. conditional probability) : probabilité d'un événement sachant qu'un autre événement s'est produit, notée $P(A \mid B) = \frac{P(A \text{ et } B)}{P(B)}$. La sensibilité, la spécificité et les valeurs prédictives sont des probabilités conditionnelles. (Chap. 1)
- Probabilité post-test (probabilité a posteriori ; roum. "probabilitate post-test" ; angl. post-test probability) : probabilité de la maladie une fois connu le résultat du test. Pour un test positif, elle est égale à la VPP correspondant à la probabilité pré-test. (Chap. 11)
- Probabilité pré-test (probabilité a priori ; roum. "probabilitate pre-test" ; angl. pre-test probability) : probabilité que le patient ait la maladie avant le résultat du test, estimée à partir de la prévalence et des données cliniques du patient. (Chap. 11)
- Protocole de l'étude (roum. "protocolul studiului" ; angl. study protocol) : document qui décrit en détail la question, le schéma d'étude, la population, les interventions, les critères de jugement, la taille de l'échantillon, la randomisation, la collecte des données, le plan d'analyse statistique et les aspects éthiques d'une étude. (Chap. 13)
- Pseudonymisation (roum. "pseudonimizare" ; angl. pseudonymisation) : remplacement des identifiants directs (nom, CNP, le code numérique personnel roumain) par un code, la clé de correspondance étant conservée séparément. Les données pseudonymisées restent des données à caractère personnel et doivent être protégées comme telles. (Chap. 1)
- Puissance du test (roum. "puterea testului" ; angl. statistical power) : probabilité de rejeter H0 lorsqu'elle est fausse, égale à 1 - β. Elle augmente avec la taille de l'effet, avec la taille de l'échantillon et avec le seuil α, et elle diminue lorsque la variabilité est grande ; on la planifie habituellement à 80% ou 90%. (Chap. 5)
- Quartiles (roum. "cuartile" ; angl. quartiles) : les trois valeurs (Q1, Q2 = la médiane, Q3) qui divisent la série ordonnée en quatre parties égales. Entre Q1 et Q3 se trouve la moitié centrale des observations. (Chap. 2)
- Question PICO (roum. "întrebarea PICO" ; angl. PICO question) : formulation structurée d'une question clinique : patient ou problème (P), intervention ou exposition (I), comparaison (C) et résultat (O). Elle facilite la recherche bibliographique. (Chap. 11)
- Randomisation (roum. "randomizare" ; angl. randomisation) : répartition des participants dans les bras de l'étude par un procédé aléatoire, de sorte que chacun ait une chance connue (habituellement égale) de recevoir chaque intervention. C'est la principale protection contre les facteurs de confusion. (Chap. 13)

- Randomisation par blocs (roum. "randomizare în blocuri" ; angl. block randomisation) : randomisation dans laquelle les attributions se font par petits blocs (par exemple de 4 ou de 6), chaque bloc contenant un nombre égal d'attributions pour chaque bras. Elle maintient des groupes numériquement équilibrés tout au long du recrutement. (Chap. 13)
- Randomisation stratifiée (roum. "randomizare stratificată" ; angl. stratified randomisation) : randomisation effectuée séparément dans des strates définies par des facteurs pronostiques importants (centre, sexe, sévérité), afin d'équilibrer ces facteurs entre les bras. (Chap. 13)
- Rang (roum. "rang" ; angl. rank) : position d'une valeur dans la série classée par ordre croissant. Les valeurs égales reçoivent le rang moyen des positions qu'elles occupent. (Chap. 7)
- Rapport de cotes (OR ; roum. "raportul cotelor" ; angl. odds ratio) : rapport entre la cote de la maladie chez les exposés et la cote de la maladie chez les non-exposés, ad/bc dans un tableau 2 x 2. C'est la seule mesure d'association valide dans les études cas-témoins, et il approche le RR lorsque la maladie est rare. (Chap. 10)
- Rapport de cotes diagnostique (DOR ; roum. "raportul cotelor de diagnostic" ; angl. diagnostic odds ratio) : rapport entre la cote d'un test positif chez les malades et la cote d'un test positif chez les non-malades, égal à ad/bc ou à RV+/RV-. Il résume la performance du test en un seul nombre. (Chap. 12)
- Rapport de prévalences (RP ; roum. "raportul prevalențelor" ; angl. prevalence ratio) : analogue du risque relatif dans les études transversales, calculé comme le rapport entre la prévalence de la maladie chez les exposés et chez les non-exposés. (Chap. 10)
- Rapport de vraisemblance (RV ; roum. "raport de verosimilitate" ; angl. likelihood ratio, LR) : nombre de fois où un résultat donné du test est plus probable chez un malade que chez un non-malade. $RV^{+} = \frac{Se}{1-Sp}$, $RV^{-} = \frac{1-Se}{Sp}$; un RV+ supérieur à 10 ou un RV- inférieur à 0,1 modifie fortement la probabilité de la maladie. (Chap. 11)
- Recensement (roum. "recensământ" ; angl. census) : étude de tous les individus d'une population. Il est exact, mais coûteux et rarement possible en recherche médicale, où l'on travaille presque toujours sur des échantillons. (Chap. 1)
- Recommandation CONSORT (roum. "ghidul CONSORT" ; angl. Consolidated Standards of Reporting Trials) : guide de rédaction des comptes rendus d'essais cliniques randomisés, comprenant une liste de contrôle et un diagramme de flux des participants (évalués, randomisés, attribués, suivis, analysés). (Chap. 13)
- Réduction absolue du risque (RAR ; roum. "reducerea absolută a riscului" ; angl. absolute risk reduction) : différence entre le risque du groupe témoin et le risque du groupe traité, pour un traitement bénéfique. Son inverse est le NST. (Chap. 10)
- Réduction relative du risque (RRR ; roum. "reducerea relativă a riscului" ; angl. relative risk reduction) : diminution du risque exprimée en proportion du risque du groupe témoin, 1 - RR. Elle paraît impressionnante même lorsque le bénéfice absolu est faible. (Chap. 10)
- Région de rejet (roum. "regiune de respingere" ; angl. rejection region) : ensemble des valeurs de la statistique de test pour lesquelles on rejette H0. Dans un test bilatéral, elle est divisée en deux queues, chacune d'aire α/2. (Chap. 5)
- Registre d'essais cliniques (roum. "registru de studii clinice" ; angl. trial registry) : base de données publique dans laquelle les études sont enregistrées avant l'inclusion du premier participant, avec des critères de jugement prédéfinis. Il prévient la présentation sélective des résultats et la publication des seuls résultats favorables. (Chap. 13)
- Règle 68-95-99,7 (roum. "regula 68-95-99,7" ; angl. empirical rule) : dans une distribution normale, environ 68,3% des valeurs se situent à au plus un écart-type de la moyenne, 95,4% à au plus deux et 99,7% à au plus trois. Exactement 95% se situent dans l'intervalle moyenne ± 1,96 écart-type. (Chap. 3)

- Règle d'addition (roum. "regula adunării" ; angl. addition rule) : $P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ et } B)$. Le terme soustrait évite de compter deux fois les individus qui présentent les deux caractéristiques. (Chap. 1)
- Règle de multiplication (roum. "regula înmulțirii" ; angl. multiplication rule) : $P(A \text{ et } B) = P(B) \times P(A \mid B)$; pour des événements indépendants, elle devient P(A) x P(B). (Chap. 1)
- Règlement (UE) 536/2014 (roum. "Regulamentul (UE) 536/2014" ; angl. EU Clinical Trials Regulation) : règlement européen qui encadre l'autorisation et la conduite des essais cliniques de médicaments à usage humain dans l'Union européenne. (Chap. 13)
- Règlement général sur la protection des données (RGPD ; roum. "Regulamentul general privind protecția datelor" ; angl. General Data Protection Regulation, GDPR) : Règlement (UE) 2016/679, qui fixe les conditions dans lesquelles les données à caractère personnel peuvent être traitées, y compris les données médicales utilisées en recherche. (Chap. 1)
- Régression linéaire multiple (roum. "regresie liniară multiplă" ; angl. multiple linear regression) : modèle avec une variable dépendante numérique et plusieurs prédicteurs ; chaque coefficient indique l'effet du prédicteur correspondant à valeurs constantes des autres (ajustement). (Chap. 9)
- Régression linéaire simple (roum. "regresie liniară simplă" ; angl. simple linear regression) : modèle qui décrit la variable dépendante y comme une fonction linéaire d'un prédicteur x, y = a + b x x, et qui permet de prédire les valeurs de y. (Chap. 9)
- Régression logistique (roum. "regresie logistică" ; angl. logistic regression) : modèle pour une variable dépendante dichotomique (malade ou sain), dont les coefficients, une fois exponentiés, donnent des rapports de cotes ajustés sur les autres prédicteurs. (Chap. 9)
- Reproductibilité (roum. "reproductibilitate" ; angl. reproducibility) : possibilité pour d'autres chercheurs d'obtenir les mêmes résultats à partir des mêmes données et des mêmes méthodes. Elle suppose des données documentées, des scripts ou des étapes d'analyse décrits en totalité et des versions de logiciel précisées. (Chap. 14)
- Résidu (roum. "reziduu" ; angl. residual) : différence entre la valeur observée et la valeur prédite par le modèle pour un individu. L'analyse des résidus vérifie la linéarité, l'homogénéité de la variance et la normalité. (Chap. 9)
- Revue systématique (roum. "revizie sistematică" ; angl. systematic review) : synthèse de toutes les études pertinentes pour une question précise, avec une recherche exhaustive, des critères de sélection explicites et l'évaluation de la qualité de chaque étude. (Chap. 11)
- Risque attribuable (roum. "risc atribuibil" ; angl. attributable risk) : excès de risque chez les exposés dû à l'exposition, égal à la différence de risques, en supposant une relation causale. (Chap. 10)
- Risque relatif (RR ; roum. "risc relativ" ; angl. relative risk, risk ratio) : rapport entre le risque (l'incidence) de la maladie chez les exposés et le risque chez les non-exposés. RR = 1 indique l'absence d'association, RR > 1 un facteur de risque, RR < 1 un facteur protecteur. Il se calcule dans les études de cohorte et dans les essais cliniques. (Chap. 10)
- Robustesse (roum. "robustețe" ; angl. robustness) : propriété d'un test de donner des résultats corrects même lorsque ses conditions sont modérément violées. Le test t est assez robuste aux légers écarts à la normalité sur les grands échantillons. (Chap. 8)
- Score z (score standardisé ; roum. "scor z" ; angl. z-score) : nombre d'écarts-types qui séparent une valeur de la moyenne, $z = \frac{x-m}{s}$. Il permet de comparer des valeurs mesurées sur des échelles différentes et de lire les aires dans la table de la loi normale. (Chap. 3)
- Sensibilité (Se ; roum. "sensibilitate" ; angl. sensitivity) : proportion des malades chez lesquels le test est positif, a/(a + b). Un test très sensible manque peu de malades ; s'il est négatif, il exclut la maladie (SnNOut). (Chap. 12)

- Série de cas (roum. "serie de cazuri" ; angl. case series) : description d'un groupe de patients atteints de la même maladie ou recevant le même traitement, sans groupe de comparaison. Elle peut générer des hypothèses, mais ne peut pas les tester. (Chap. 10)
- Seuil de signification (α ; roum. "prag de semnificație" ; angl. significance level) : probabilité maximale d'erreur de type I acceptée, fixée avant l'analyse, habituellement 0,05. Si $p < \alpha$, le résultat est déclaré statistiquement significatif. (Chap. 5)
- Signification clinique (roum. "semnificație clinică" ; angl. clinical significance) : importance pratique d'un effet pour le patient ou pour la décision médicale. On en juge d'après la taille de l'effet et l'intervalle de confiance, comparés à la différence minimale cliniquement importante. (Chap. 5)
- Signification statistique (roum. "semnificație statistică" ; angl. statistical significance) : situation dans laquelle $p < \alpha$, c'est-à-dire que les données sont peu compatibles avec H0. Elle ne garantit pas que l'effet soit grand, cliniquement important ou causal. (Chap. 5)
- SnNOut (roum. "SnNOut" ; angl. SnNOut) : moyen mnémotechnique : un test très sensible (Sn), négatif (N), exclut la maladie (Out, rule out). (Chap. 12)
- Spécificité (Sp ; roum. "specificitate" ; angl. specificity) : proportion des non-malades chez lesquels le test est négatif, d/(c + d). Un test très spécifique donne peu de résultats faussement positifs ; s'il est positif, il confirme la maladie (SpPIn). (Chap. 12)
- Sphéricité (roum. "sfericitate" ; angl. sphericity) : condition de l'ANOVA à mesures répétées selon laquelle les variances des différences entre toutes les paires de temps de mesure sont égales ; elle se vérifie par le test de Mauchly et, si elle n'est pas remplie, on applique une correction (Greenhouse-Geisser). (Chap. 6)
- SpPIn (roum. "SpPIn" ; angl. SpPIn) : moyen mnémotechnique : un test très spécifique (Sp), positif (P), confirme la maladie (In, rule in). (Chap. 12)
- Standard de référence (étalon-or ; roum. "standard de referință" ; angl. reference standard) : meilleure méthode disponible pour établir la présence de la maladie, par rapport à laquelle on évalue le test étudié (par exemple la biopsie pour la fibrose hépatique, l'HGPO pour le diabète). (Chap. 12)
- Standardisation des taux (roum. "standardizarea ratelor" ; angl. rate standardisation) : méthode de comparaison des taux entre des populations de structures par âge différentes, par référence à une population standard commune. (Chap. 10)
- STARD (roum. "STARD" ; angl. Standards for Reporting of Diagnostic Accuracy Studies) : guide des éléments à rapporter dans une étude d'exactitude diagnostique, y compris le diagramme de flux des patients et le tableau 2 x 2. (Chap. 12)
- Statistique (indicateur d'échantillon ; roum. "statistică (indicator de eșantion)" ; angl. statistic) : valeur calculée à partir des données de l'échantillon (la moyenne m, l'écart-type s, la proportion p), utilisée comme estimation du paramètre correspondant. Elle varie d'un échantillon à l'autre. (Chap. 4)
- Statistique bayésienne (roum. "statistică bayesiană" ; angl. Bayesian statistics) : approche dans laquelle les connaissances antérieures (la distribution a priori) sont mises à jour par les données, ce qui produit une distribution a posteriori du paramètre. Elle permet des affirmations directes du type "la probabilité que l'effet soit positif est de 95%". (Chap. 11)
- Statistique de test (roum. "statistica testului" ; angl. test statistic) : valeur calculée à partir des données (t, z, F, χ^2, U) qui mesure à quel point les données s'éloignent de ce que prédirait H0. On en tire, avec sa distribution sous H0, la valeur p. (Chap. 5)
- Statistique descriptive (roum. "statistică descriptivă" ; angl. descriptive statistics) : partie de la statistique qui résume les données d'un échantillon par des tableaux de fréquences, des indicateurs (moyenne, médiane, écart-type) et des graphiques, sans formuler d'affirmations sur la population. C'est la première étape de toute analyse. (Chap. 2)

- Statistique F (roum. "statistica F" ; angl. F statistic) : rapport entre le carré moyen intergroupes et le carré moyen intragroupe. Des valeurs très supérieures à 1 montrent que les moyennes diffèrent davantage que ne l'expliquerait la variabilité aléatoire. (Chap. 6)
- Statistique fréquentiste (roum. "statistică frecventistă" ; angl. frequentist statistics) : approche dans laquelle la probabilité est la fréquence à long terme d'un événement lors de répétitions hypothétiques de l'étude ; c'est sur elle que reposent les valeurs p et les intervalles de confiance de ce cours. (Chap. 11)
- Statistique inférentielle (roum. "statistică inferențială" ; angl. inferential statistics) : partie de la statistique qui généralise de l'échantillon à la population, par l'estimation des paramètres avec des intervalles de confiance et par les tests d'hypothèses. Ses conclusions sont probabilistes, et non certaines. (Chap. 1)
- STROBE (roum. "STROBE" ; angl. Strengthening the Reporting of Observational Studies in Epidemiology) : guide de rédaction des comptes rendus d'études observationnelles (de cohorte, cas-témoins, transversales). (Chap. 10)
- Système GRADE (roum. "sistemul GRADE" ; angl. GRADE approach) : méthode d'évaluation du niveau de certitude des preuves (élevé, modéré, faible, très faible) et de la force des recommandations des guides de pratique. (Chap. 11)
- Tableau 2 x 2 (roum. "tabel 2 x 2" ; angl. two-by-two table) : tableau de contingence à deux lignes et deux colonnes (les cases a, b, c, d), utilisé aussi bien dans l'analyse du risque (exposition x maladie) que dans l'évaluation des tests diagnostiques (maladie x résultat du test). (Chap. 10)
- Tableau de contingence (roum. "tabel de contingență" ; angl. contingency table) : tableau à double entrée qui croise les catégories de deux variables et donne la fréquence de chaque combinaison. Le tableau 2 x 2 en est le cas le plus simple. (Chap. 7)
- Tableau de fréquences (roum. "tabel de frecvențe" ; angl. frequency table) : tableau qui indique, pour chaque valeur ou catégorie, la fréquence absolue, le pourcentage, le pourcentage valide et, pour les variables ordonnées, le pourcentage cumulé. C'est l'outil le plus simple pour détecter les erreurs de codage. (Chap. 2)
- Taille de l'échantillon (roum. "dimensiunea eșantionului" ; angl. sample size) : nombre d'unités statistiques nécessaire pour estimer un paramètre avec la précision voulue ou pour détecter un effet donné avec une puissance fixée. Elle se calcule avant l'étude et s'arrondit à l'entier supérieur. (Chap. 4)
- Taille de l'effet (roum. "mărimea efectului" ; angl. effect size) : mesure de l'ampleur d'une différence ou d'une association, indépendante de la taille de l'échantillon (d de Cohen, êta carré, r, OR). Elle doit être présentée à côté de la valeur p. (Chap. 5)
- Taille de l'effet clinique (effet absolu ; roum. "mărime a efectului clinic" ; angl. absolute effect) : expression du résultat en unités faciles à comprendre pour la décision clinique (mmHg, points de pourcentage, NST), complémentaire des mesures relatives comme le RR. (Chap. 10)
- Taux (roum. "rata" ; angl. rate) : rapport qui comporte le temps au dénominateur (cas par personnes-années). Dans le langage courant, le terme est souvent employé improprement pour des proportions. (Chap. 10)
- Taux d'erreur de classement (roum. "rata erorii de clasificare" ; angl. misclassification rate) : proportion des personnes mal classées, $\frac{b+c}{N} = \frac{\mathrm{FN+FP}}{N}$, c'est-à-dire 1 - l'exactitude. (Chap. 12)
- Test bilatéral (roum. "test bilateral" ; angl. two-sided test) : test dans lequel l'hypothèse alternative admet des différences dans les deux sens ($\mu 1 \neq \mu 2$). C'est le choix par défaut en recherche médicale. (Chap. 5)
- Test de dépistage (roum. "test de screening" ; angl. screening test) : test appliqué à des personnes asymptomatiques pour identifier la maladie à un stade précoce. Il doit être très sensible, peu coûteux et acceptable ; les résultats positifs sont confirmés par un test spécifique. (Chap. 12)

- Test de Dunn (roum. "testul Dunn" ; angl. Dunn test) : test post-hoc pour les comparaisons deux à deux après un test de Kruskal-Wallis significatif, avec ajustement des valeurs p pour les comparaisons multiples. (Chap. 7)
- Test de Friedman (roum. "testul Friedman" ; angl. Friedman test) : alternative non paramétrique de l'ANOVA à mesures répétées, pour trois mesures ou plus chez les mêmes individus. Les rangs sont attribués à l'intérieur de chaque individu. (Chap. 7)
- Test de Kolmogorov-Smirnov (KS ; roum. "testul Kolmogorov-Smirnov" ; angl. Kolmogorov-Smirnov test) : test qui compare la distribution observée à une distribution théorique, utilisé aussi pour la normalité. Pour vérifier la normalité, il est moins puissant que le test de Shapiro-Wilk. (Chap. 3)
- Test de Kruskal-Wallis (roum. "testul Kruskal-Wallis" ; angl. Kruskal-Wallis test) : alternative non paramétrique de l'ANOVA à un facteur, pour trois groupes indépendants ou plus. Un résultat significatif est suivi de comparaisons post-hoc (par exemple le test de Dunn). (Chap. 7)
- Test de Levene (roum. "testul Levene" ; angl. Levene test) : test de l'égalité des variances dans deux groupes ou plus, avec H0 "les variances sont égales". Un $p < 0,05$ oriente vers le test de Welch ou vers l'ANOVA de Welch. (Chap. 6)
- Test de McNemar (roum. "testul McNemar" ; angl. McNemar test) : test pour deux variables dichotomiques mesurées chez les mêmes individus (par exemple, le résultat de deux tests diagnostiques chez les mêmes patients) ; il n'utilise que les paires discordantes. (Chap. 7)
- Test de Shapiro-Wilk (roum. "testul Shapiro-Wilk" ; angl. Shapiro-Wilk test) : test de normalité dans lequel H0 affirme que les données proviennent d'une loi normale. Un $p < 0,05$ indique un écart à la normalité ; un $p \geq 0,05$ ne prouve pas la normalité, mais seulement que l'écart n'a pas été détecté. (Chap. 3)
- Test de Tukey (roum. "testul Tukey" ; angl. Tukey HSD test) : test post-hoc qui compare toutes les paires de moyennes en maintenant l'erreur globale de type I à α ; il donne les différences avec des IC ajustés. (Chap. 6)
- Test des rangs signés de Wilcoxon (roum. "testul Wilcoxon al rangurilor cu semn" ; angl. Wilcoxon signed-rank test) : alternative non paramétrique du test t apparié ; il analyse les rangs des valeurs absolues des différences, en tenant compte de leur signe. (Chap. 7)
- Test diagnostique (roum. "test diagnostic" ; angl. diagnostic test) : toute procédure (analyse de laboratoire, imagerie, score clinique, question) utilisée pour établir si un patient a ou non une maladie donnée. (Chap. 12)
- Test du khi-deux d'ajustement (roum. "testul chi-pătrat de concordanță" ; angl. chi-square goodness-of-fit test) : test qui compare la distribution observée d'une seule variable catégorielle à une distribution théorique (par exemple, des proportions égales dans trois bras). (Chap. 7)
- Test du khi-deux d'association (test du χ^2 d'indépendance ; roum. "testul chi-pătrat de asociere" ; angl. chi-square test of independence) : test qui vérifie si deux variables catégorielles sont associées, en comparant les fréquences observées d'un tableau de contingence à celles attendues en l'absence d'association. (Chap. 7)
- Test exact de Fisher (roum. "testul exact Fisher" ; angl. Fisher exact test) : test pour les tableaux 2 x 2 (et plus grands) qui calcule exactement la probabilité de la configuration observée et des configurations plus extrêmes. On l'utilise lorsque les fréquences attendues sont faibles. (Chap. 7)
- Test non paramétrique (roum. "test neparametric" ; angl. nonparametric test) : test qui ne suppose pas une distribution particulière des données (habituellement la loi normale) et qui travaille souvent sur les rangs des valeurs. On l'utilise pour les variables ordinales, pour les distributions asymétriques avec de petits échantillons et en présence de valeurs aberrantes. (Chap. 7)
- Test paramétrique (roum. "test parametric" ; angl. parametric test) : test qui suppose une distribution particulière des données, en règle générale normale, et qui compare des paramètres

comme la moyenne (tests t, ANOVA, corrélation de Pearson). Lorsque ses conditions sont remplies, il est plus puissant que son équivalent non paramétrique. (Chap. 8)

- Test statistique (roum. "test statistic" ; angl. statistical test) : procédé par lequel on décide, avec un risque d'erreur connu, si les données sont compatibles avec l'hypothèse nulle. Le choix du test dépend de la question, du type des variables, du nombre de groupes et de leur indépendance. (Chap. 5)
- Test t apparié (roum. "testul t pereche" ; angl. paired t-test) : il compare les moyennes de deux séries de mesures liées (le même patient avant et après le traitement), en analysant la moyenne des différences individuelles. La condition principale est la normalité des différences. (Chap. 6)
- Test t de Welch (roum. "testul t Welch" ; angl. Welch t-test) : version du test t pour échantillons indépendants qui ne suppose pas l'égalité des variances et ajuste les degrés de liberté. C'est un choix sûr dans la plupart des situations. (Chap. 6)
- Test t pour échantillons indépendants (roum. "testul t pentru eşantioane independente" ; angl. independent samples t-test) : il compare les moyennes de deux groupes formés d'individus différents. La version classique (de Student) suppose des variances égales ; la version de Welch, non. (Chap. 6)
- Test t pour un échantillon (roum. "testul t pentru un eşantion" ; angl. one-sample t-test) : il compare la moyenne d'un échantillon à une valeur de référence fixée (par exemple une norme publiée), lorsque σ est inconnu. Il a n - 1 degrés de liberté. (Chap. 6)
- Test U de Mann-Whitney (test de la somme des rangs de Wilcoxon ; roum. "testul Mann-Whitney U" ; angl. Mann-Whitney U test) : alternative non paramétrique du test t pour échantillons indépendants ; il compare les distributions de deux groupes à partir des rangs. Le résultat se présente avec des médianes et des écarts interquartiles. (Chap. 7)
- Test unilatéral (roum. "test unilateral" ; angl. one-sided test) : test dans lequel l'hypothèse alternative n'admet qu'un seul sens ($\mu 1 > \mu 2$). Il ne se justifie que si une différence dans le sens opposé est impossible ou sans intérêt et si le sens a été fixé avant l'analyse. (Chap. 5)
- Test z (roum. "testul z" ; angl. z-test) : test fondé sur la loi normale, utilisé pour une moyenne lorsque l'écart-type de la population est connu ou que l'échantillon est grand, ainsi que pour les proportions. (Chap. 5)
- Tests en parallèle (roum. "teste în paralel" ; angl. parallel testing) : application simultanée de deux tests ; le patient est considéré comme positif si l'un ou l'autre des tests est positif. Cela augmente la sensibilité et diminue la spécificité. (Chap. 12)
- Tests en série (roum. "teste în serie" ; angl. serial testing) : application du second test uniquement aux personnes positives au premier ; le patient n'est considéré comme malade que si les deux tests sont positifs. Cela augmente la spécificité et diminue la sensibilité. (Chap. 12)
- Tests post-hoc (roum. "teste post-hoc" ; angl. post hoc tests) : comparaisons deux à deux effectuées après une ANOVA significative, avec contrôle de l'erreur de type I (Tukey, Bonferroni, Games-Howell). Ils montrent quels groupes diffèrent entre eux. (Chap. 6)
- Théorème central limite (roum. "teorema limită centrală" ; angl. central limit theorem) : la moyenne d'échantillons suffisamment grands suit une loi approximativement normale, quelle que soit la forme de la distribution de la variable dans la population. Il justifie l'utilisation des tests fondés sur la moyenne pour les grands échantillons. (Chap. 4)
- Théorème de Bayes (roum. "teorema lui Bayes" ; angl. Bayes' theorem) : relation qui permet de calculer la probabilité d'une cause (la maladie) connaissant l'effet observé (le test positif) : $P(B \mid T) = \frac{P(T|B) \times P(B)}{P(T)}$. En médecine, il transforme la probabilité pré-test en probabilité post-test. (Chap. 11)
- Transformation logarithmique (roum. "transformare logaritmică" ; angl. log transformation) : remplacement de chaque valeur par son logarithme, afin de réduire l'asymétrie positive et de

rapprocher la distribution de la loi normale. Les résultats sont ramenés sur l'échelle initiale par exponentiation, ce qui donne la moyenne géométrique. (Chap. 3)

- Unité statistique (roum. "unitate statistică" ; angl. statistical unit) : élément individuel sur lequel portent les observations (un patient, un œil, un prélèvement sanguin). Il est important de bien définir l'unité : les deux yeux d'un même patient ne sont pas des observations indépendantes. (Chap. 1)
- Valeur aberrante (roum. "valoare aberantă" ; angl. outlier) : observation qui s'écarte fortement du reste des données. Il peut s'agir d'une erreur de saisie (PAS 1540 mmHg) ou d'une valeur réelle extrême ; dans le premier cas, on la corrige d'après la source, dans le second, on la conserve et on analyse son influence. (Chap. 2)
- Valeur critique (roum. "valoare critică" ; angl. critical value) : valeur de la statistique de test qui sépare la région de non-rejet de la région de rejet de H0, pour un α donné. Elle se lit dans les tables statistiques (annexe E). (Chap. 5)
- Valeur de référence du test (μ0 ; roum. "valoare de referință a testului" ; angl. reference value, null value) : valeur du paramètre spécifiée par l'hypothèse nulle, par exemple la moyenne de 120 mmHg dans un test t sur un échantillon, 1 pour le RR et l'OR, ou 0 pour une différence. (Chap. 5)
- Valeur manquante (roum. "valoare lipsă" ; angl. missing value) : observation qui aurait dû être enregistrée, mais qui manque. Les valeurs manquantes se signalent pour chaque variable et ne se complètent pas à volonté, mais uniquement par une méthode d'imputation prévue dans le plan d'analyse. (Chap. 2)
- Valeur p (roum. "valoarea p" ; angl. p-value) : probabilité d'obtenir un résultat au moins aussi extrême que celui observé, si H0 était vraie. Ce n'est pas la probabilité que H0 soit vraie, et elle ne mesure ni la taille ni l'importance clinique de l'effet. (Chap. 5)
- Valeur prédictive négative (VPN ; roum. "valoarea predictivă negativă" ; angl. negative predictive value, NPV) : proportion des personnes sans la maladie parmi celles dont le test est négatif, d/(b + d). Elle augmente lorsque la prévalence diminue. (Chap. 12)
- Valeur prédictive positive (VPP ; roum. "valoarea predictivă pozitivă" ; angl. positive predictive value, PPV) : proportion des malades parmi les personnes dont le test est positif, a/(a + c). Elle dépend fortement de la prévalence : le même test a une VPP plus faible dans une population où la maladie est rare. (Chap. 12)
- Valeur seuil (cut-off ; roum. "valoare prag" ; angl. cut-off value) : valeur d'un test numérique (score, concentration) au-dessus de laquelle le résultat est considéré comme positif. Déplacer le seuil augmente la sensibilité aux dépens de la spécificité, ou inversement. (Chap. 12)
- Valeurs prédictives et prévalence (roum. "valoare predictivă și prevalență" ; angl. predictive values and prevalence) : relation selon laquelle, à sensibilité et spécificité constantes, la VPP augmente et la VPN diminue lorsque la prévalence de la maladie augmente. Elle explique pourquoi un test performant à l'hôpital donne de nombreux faux positifs dans le dépistage de la population générale. (Chap. 12)
- Validation des données à la saisie (roum. "validarea datelor la introducere" ; angl. data validation rules) : règles imposées dans le formulaire électronique (plages autorisées, champs obligatoires, listes d'options) qui empêchent la saisie de valeurs impossibles ou codées de façon non uniforme. (Chap. 13)
- Variabilité biologique (roum. "variabilitate biologică" ; angl. biological variability) : fait qu'une même caractéristique diffère d'un individu à l'autre et, chez un même individu, d'un moment à l'autre. C'est la raison pour laquelle la médecine a besoin de la statistique. (Chap. 1)
- Variabilité intergroupes (roum. "variabilitate între grupuri" ; angl. between-group variability) : part de la variabilité totale expliquée par les différences entre les moyennes des groupes (la somme des carrés intergroupes). (Chap. 6)

- Variabilité intragroupe (roum. "variabilitate în interiorul grupurilor" ; angl. within-group variability) : part de la variabilité totale due aux différences entre les individus d'un même groupe (la somme des carrés résiduelle). (Chap. 6)
- Variable (roum. "variabilă" ; angl. variable) : caractéristique qui peut prendre des valeurs différentes d'un individu à l'autre (âge, sexe, glycémie). Le type de la variable détermine les indicateurs descriptifs, les graphiques et les tests utilisables. (Chap. 2)
- Variable aléatoire (roum. "variabilă aleatoare" ; angl. random variable) : variable dont la valeur dépend du résultat d'un processus aléatoire, par exemple la glycémie d'un patient tiré au hasard. (Chap. 3)
- Variable continue (roum. "variabilă continuă" ; angl. continuous variable) : variable numérique qui peut prendre, en théorie, n'importe quelle valeur d'un intervalle (la glycémie, la taille) ; sa précision n'est limitée que par l'instrument de mesure. (Chap. 2)
- Variable dépendante (variable de résultat ; roum. "variabilă dependentă" ; angl. dependent variable, outcome) : variable dont nous voulons expliquer ou prédire le comportement (par exemple la PAS ou la présence d'un diabète). (Chap. 8)
- Variable dérivée (roum. "variabilă derivată" ; angl. derived variable) : variable calculée à partir d'autres variables (l'IMC à partir du poids et de la taille, la perte de poids à partir de deux mesures). Elle se recalcule au lieu d'être saisie à la main, et la comparaison avec la valeur enregistrée aide à découvrir les erreurs. (Chap. 2)
- Variable dichotomique (variable binaire ; roum. "variabilă dihotomică (binară)" ; angl. binary variable) : variable nominale à exactement deux catégories (malade ou sain, 0 ou 1). C'est le type de variable des tableaux 2 x 2 de risque et de diagnostic. (Chap. 2)
- Variable discrète (roum. "variabilă discretă" ; angl. discrete variable) : variable numérique qui ne prend que des valeurs séparées, habituellement des nombres entiers issus d'un comptage (nombre de consultations, nombre de grossesses). (Chap. 2)
- Variable indépendante (prédicteur ; roum. "variabilă independentă" ; angl. independent variable, predictor) : variable utilisée pour expliquer ou prédire la variable de résultat (par exemple l'IMC, l'âge, le bras de traitement). (Chap. 8)
- Variable indicatrice (variable muette, dummy ; roum. "variabilă indicator" ; angl. dummy variable) : variable 0/1 créée pour introduire une catégorie dans un modèle de régression ; une variable à k catégories nécessite k - 1 indicatrices, par rapport à une catégorie de référence. (Chap. 9)
- Variable nominale (roum. "variabilă nominală" ; angl. nominal variable) : variable catégorielle sans ordre naturel entre les catégories (le milieu urbain ou rural, le bras de traitement). Les codes numériques attribués aux catégories sont de simples étiquettes. (Chap. 2)
- Variable numérique (variable quantitative ; roum. "variabilă numerică (cantitativă)" ; angl. numerical variable) : variable exprimée par un nombre assorti d'une unité de mesure, pour laquelle les opérations arithmétiques ont un sens (la PAS en mmHg, le poids en kg). (Chap. 2)
- Variable ordinale (roum. "variabilă ordinală" ; angl. ordinal variable) : variable catégorielle dont les catégories ont un ordre naturel, mais dont les distances ne sont pas nécessairement égales (le stade d'une maladie, le niveau d'études, le score de satisfaction 1-5). Elle se décrit par la médiane et les quartiles. (Chap. 2)
- Variable qualitative (variable catégorielle ; roum. "variabilă calitativă (categorială)" ; angl. categorical variable) : variable dont les valeurs sont des catégories, et non des mesures (le sexe, le groupe sanguin, le statut tabagique). Elle se décrit par des fréquences et des pourcentages. (Chap. 2)
- Variance (roum. "varianță" ; angl. variance) : moyenne des carrés des écarts à la moyenne, calculée sur l'échantillon avec le dénominateur n - 1. Son unité est le carré de l'unité de la variable ; c'est pourquoi on présente l'écart-type dans les résultats. (Chap. 2)

- Vrai négatif (VN ; roum. "adevărat negativ" ; angl. true negative, TN) : personne non malade dont le test est négatif (case d). (Chap. 12)
- Vrai positif (VP ; roum. "adevărat pozitiv" ; angl. true positive, TP) : malade dont le test est positif (case a du tableau diagnostique). (Chap. 12)

Annexe C. Le jeu de données pédagogique mg2026 : scénario, dictionnaire des données et erreurs du fichier brut

Presque tous les exemples numériques de cet ouvrage, de la moyenne de la pression systolique du chapitre 2 à la courbe ROC du score FINDRISC du chapitre 12, sont calculés sur le même jeu de données, appelé mg2026. Nous avons choisi cette solution pour une raison simple : un étudiant qui reconnaît les variables d'un chapitre à l'autre peut se concentrer sur la méthode, sans réapprendre chaque fois le contexte clinique. La présente annexe rassemble en un seul endroit tout ce qu'il faut savoir sur les fichiers : le scénario, le dictionnaire des variables, les erreurs laissées intentionnellement dans la version brute et la manière de citer le jeu de données dans un exposé ou dans un devoir.

C.1 Le scénario

Le jeu de données décrit un essai randomisé pragmatique, simulé, mené dans 12 cabinets de médecine générale de l'ouest de la Roumanie. Au total, 212 adultes ont été évalués pour l'éligibilité ; 52 d'entre eux ont été exclus (31 ne remplissaient pas les critères, 21 ont refusé de participer) et 160 ont été randomisés (le diagramme de flux CONSORT est présenté au chapitre 13). Le principal critère d'inclusion était un risque cardiométabolique élevé : surpoids ou obésité, prédiabète ou suspicion de diabète non diagnostiqué.

Les participants ont été répartis au hasard dans trois bras :

- bras A (n = 53) : conseil standard sur le mode de vie, assuré par le médecin de famille ;
- bras B (n = 53) : le même conseil, plus une application mobile d'automesure (poids, pas, alimentation) ;
- bras C (n = 54) : conseil, application et une téléconsultation mensuelle avec le médecin.

Les mesures ont été réalisées à l'inclusion (temps 0), à 3 mois et à 6 mois. À l'inclusion, tous les patients ont passé une hyperglycémie provoquée par voie orale (HGPO), considérée ici comme le standard de référence pour le diagnostic du diabète, ce qui permet d'évaluer le score FINDRISC en tant que test de dépistage (chapitre 12). Le critère de jugement principal de l'étude est la diminution de la glycémie à jeun à 6 mois ; les critères secondaires sont la perte de poids (y compris la proportion de patients ayant perdu au moins 5% de leur poids), la pression artérielle systolique à 6 mois et la satisfaction à l'égard de l'intervention.

Le même jeu de données permet aussi des analyses observationnelles, transversales, sur les données de l'inclusion : l'association entre obésité et diabète, la corrélation entre l'IMC et la pression artérielle, la comparaison entre les sexes. Il s'agit d'un choix pédagogique délibéré. Dans une étude réelle, les analyses secondaires devraient être déclarées dans le protocole avant la collecte des données, et leurs résultats seraient présentés comme exploratoires.

C.2 Les fichiers

Il existe deux fichiers, disponibles sur la plateforme Moodle de la discipline [52] :

- **mg2026_curat.csv** : 160 lignes (une ligne par patient) et 28 colonnes, sans valeurs manquantes et sans erreurs de saisie. C'est le fichier sur lequel sont faites toutes les analyses du cours et des travaux pratiques.

- **mg2026_brut.csv** : 161 lignes et les mêmes 28 colonnes, avec 11 erreurs de saisie laissées intentionnellement et une ligne en double. C'est le support de la séance de travaux pratiques consacrée au nettoyage des données (chapitre 2 et premières séances de travaux pratiques).

Les deux fichiers utilisent la virgule comme séparateur de champs et le point comme séparateur décimal (convention CSV internationale). Lors de l'importation dans Google Sheets avec les paramètres régionaux de la Roumanie ou dans Excel en roumain, le logiciel peut mal interpréter les décimales ; dans JASP et dans MedCalc, l'importation fonctionne directement. Le codage des caractères est UTF-8, et les noms des variables sont écrits sans diacritiques, pour assurer la compatibilité avec tous les logiciels.

C.3 Le dictionnaire des données

Le tableau C.1 décrit chaque variable. La colonne "Domaine valide" donne la règle de validation que nous recommandons pour la vérification du fichier brut (valeurs cliniquement plausibles), suivie, entre parenthèses, de l'intervalle effectivement observé dans le fichier nettoyé. Le fichier nettoyé ne contient aucune valeur manquante ; la colonne "Manquantes" se rapporte au fichier brut.

Tableau C.1. Dictionnaire des variables du jeu de données mg2026.

Variable	Description	Type	Échelle	Unité / codage	Domaine valide (observé)	Manquantes dans le fichier brut
id	identifiant du patient	texte	nominale (identifiant)	P001-P160	unique sur chaque ligne (P001-P160)	0 ; un id apparaît deux fois
brat	bras de randomisation	catégorielle	nominale	A, B, C (voir C.1)	A, B, C (A 53, B 53, C 54)	0
sex	sexe	catégorielle	nominale dichotomique	F, M	F, M (F 86, M 74)	0
varsta	âge à l'inclusion	numérique discrète	rapport	années révolues	18-100 (35-79)	1
mediu	milieu de résidence	catégorielle	nominale dichotomique	urban, rural	urban, rural (95 / 65)	0
educatie	niveau d'études	catégorielle ordonnée	ordinale	1 collège (premier cycle du secondaire) ; 2 lycée ou école professionnelle ; 3 études universitaires ; 4 études postuniversitaires	1-4 (1-4)	0
fumat	statut tabagique	catégorielle	nominale (3 catégories)	nefumator, fost fumator, fumator	les trois libellés (74 / 33 / 53)	0
activitate	activité physique autoévaluée	catégorielle ordonnée	ordinale	1 très faible... 5 très intense	1-5 (1-5)	0
inaltime	taille (stature)	numérique continue	rapport	cm, entier	130-210 (145-192)	0
greutate0	poids à	numérique	rapport	kg, une	35-250 (58,1-	0

Variable	Description	Type	Échelle	Unité / codage	Domaine valide (observé)	Manquantes dans le fichier brut
	l'inclusion	continue		décimale	159,9)	
imc0	indice de masse corporelle à l'inclusion	numérique continue (dérivée)	rapport	kg/m², une décimale ; greutate0 / (inaltime/100)²	15-70 (23,5-48,3)	0
talie	tour de taille	numérique continue	rapport	cm, entier	50-180 (65-140)	0
tas0	pression artérielle systolique à l'inclusion	numérique	rapport	mmHg	70-260 (102-169)	0
tad0	pression artérielle diastolique à l'inclusion	numérique	rapport	mmHg	40-150 (58-103)	0
fc0	fréquence cardiaque à l'inclusion	numérique	rapport	battements/min	35-180 (51-102)	0
colesterol	cholestérol total	numérique	rapport	mg/dL	80-450 (119-306)	0
hdl	HDL-cholestérol	numérique	rapport	mg/dL	10-120 (20-78)	0
trigliceride	triglycérides sériques	numérique	rapport	mg/dL	20-1500 (44-442)	0
glicemie0	glycémie à jeun à l'inclusion	numérique	rapport	mg/dL, entier	40-400 (75-154)	0
glicemie3	glycémie à jeun à 3 mois	numérique	rapport	mg/dL	40-400 (49-144)	0
glicemie6	glycémie à jeun à 6 mois	numérique	rapport	mg/dL	40-400 (45-150)	1
hba1c0	hémoglobine glyquée à l'inclusion	numérique continue	rapport	%, une décimale	3,5-15 (4,5-7,0)	0
findrisc	score FINDRISC de risque de diabète	numérique discrète (score)	ordinale traitée comme numérique	points	0-26 (0-25)	0
diabet_ogtt	diabète confirmé par HGPO à l'inclusion	catégorielle	nominale dichotomique	0 non, 1 oui	0, 1 (29 avec diabète)	0
greutate6	poids à 6 mois	numérique continue	rapport	kg, une décimale	35-250 (50,7-155,6)	0
tas6	pression artérielle systolique à 6 mois	numérique	rapport	mmHg	70-260 (85-166)	0
scadere5	perte de poids d'au moins 5% à 6 mois	catégorielle (dérivée)	nominale dichotomique	0 non, 1 oui ; 1 si (greutate0 - greutate6)/greutate0 ≥ 0,05	0, 1 (38 avec 1)	0

Variable	Description	Type	Échelle	Unité / codage	Domaine valide (observé)	Manquantes dans le fichier brut
satisfactie	satisfaction à l'égard de l'intervention à 6 mois	catégorielle ordonnée	ordinale (Likert)	1 pas du tout satisfait... 5 très satisfait	1-5 (1-5)	0

Quelques remarques sur les échelles, car c'est précisément là qu'apparaissent les erreurs à l'examen. Les variables educatie, activitate et satisfactie sont codées par des chiffres, mais elles sont ordinales : leur moyenne a un sens limité, et pour les comparaisons on utilise la médiane, l'écart interquartile et les tests non paramétriques (chapitre 7). Le score FINDRISC est, à strictement parler, lui aussi ordinal, mais avec 27 valeurs possibles il est traité en pratique comme numérique ; nous le décrivons par la moyenne et par la médiane. La pression artérielle, la glycémie et le poids sont des variables de rapport (elles ont un zéro absolu), même si, pour la statistique de ce cours, la distinction entre échelle d'intervalle et échelle de rapport compte peu. Enfin, imc0 et scadere5 sont des variables dérivées, calculées à partir d'autres colonnes ; on ne les saisit pas à la main, on les recalcule, et la différence entre la valeur enregistrée et la valeur recalculée est un excellent outil de contrôle des erreurs (voir C.4).

C.4 Les erreurs du fichier brut et leur correction

Le fichier mg2026_brut.csv imite ce qui se passe dans tout cabinet ou service où les données sont recopiées à la main des dossiers dans une feuille de calcul. Les erreurs sont de quatre sortes : codage non uniforme des catégories, valeurs impossibles (habituellement un chiffre en trop, une virgule déplacée ou une unité de mesure erronée), texte saisi dans une colonne numérique et valeurs manquantes. S'y ajoute une ligne copiée deux fois. Chaque erreur est une déformation de la valeur vraie ; une fois détectée, elle peut donc être corrigée ; toutefois, la correction se confirme toujours dans le document source (dossier du patient, cahier d'observation, compte rendu de laboratoire), dont le fichier nettoyé joue ici le rôle.

Tableau C.2. Erreurs de mg2026_brut.csv, mode de détection et valeur correcte.

N°	Patient (ligne de données)	Variable	Valeur dans le fichier brut	Mode de détection	Nature de l'erreur	Valeur correcte (fichier nettoyé)
1	P005 (5)	sex	f	tableau de fréquences de sex : une troisième catégorie apparaît	lettre minuscule au lieu d'une majuscule	F
2	P021 (21)	sex	Feminin	même tableau de fréquences : une quatrième catégorie	mot entier au lieu du code	F
3	P034 (34)	tas0	1540	minimum/maximum, box-plot ; la valeur dépasse la règle 70-260	un zéro tapé en trop	154
4	P042 (42)	varsta	cellule vide	nombre de valeurs valides (160 au lieu de	valeur manquante	70 (récupérée dans le dossier)

N°	Patient (ligne de données)	Variable	Valeur dans le fichier brut	Mode de détection	Nature de l'erreur	Valeur correcte (fichier nettoyé)
				161) ; COUNTBLANK		
5	P057 (57)	inaltime	1.72	minimum < 130 ; valeur en mètres, et non en centimètres	unité de mesure erronée	172
6	P064 (64)	glicemie0	$\frac{87\text{mg}}{\text{dl}}$	la colonne est importée comme texte ; ISNUMBER donne FALSE	unité écrite dans la cellule	87
7	P084 (84)	fumat	"Fumator " (majuscule et espace final)	tableau de fréquences : une quatrième catégorie ; LEN indique un caractère de trop	majuscule et espace invisible	fumator
8	P089 (89)	hba1c0	5,2	texte avec virgule dans une colonne à point décimal ; la colonne devient du texte	séparateur décimal erroné	5,2 (écrit 5.2 dans le CSV)
9	P102 (102)	greutate0	7.26	minimum < 35 kg ; l'IMC recalculé serait de 2,7 kg/m²	virgule décimale déplacée	72,6
10	P121 (121)	mediu	Urban	tableau de fréquences : "Urban" séparé de "urban"	majuscule	urban
11	P134 (134)	glicemie6	cellule vide	nombre de valeurs valides pour glicemie6	valeur manquante	135 (récupérée dans le dossier)
12	P091 (91 et 161)	toutes	ligne identique répétée à la fin	COUNTIF sur id > 1 ; dans pandas, duplicated()	ligne en double	une seule ligne P091

Quelques commentaires sur le tableau, dans l'ordre où un étudiant attentif les découvrirait.

La première étape est toujours le tableau de fréquences de chaque variable catégorielle. Dans le fichier brut, le sexe a quatre catégories (F 84, M 75, f 1, Feminin 1), le tabagisme en a quatre (nefumator 75, fumator 52, fost fumator 33, "Fumator " 1), et le milieu de résidence trois (urban 95, rural 65, Urban 1). Les totaux dépassent de 1 les valeurs du fichier nettoyé à cause de la ligne en double. Les logiciels traitent "f" et "F" comme deux valeurs différentes, de même que "Urban" et "urban", et un espace invisible à la fin du mot suffit pour que "Fumator " soit compté à part. L'uniformisation des libellés se fait automatiquement (minuscules ou majuscules selon le

dictionnaire, suppression des espaces avec la fonction TRIM), puis le résultat est comparé au dossier.

La deuxième étape est la statistique descriptive avec le minimum et le maximum de chaque variable numérique. Trois erreurs sautent immédiatement aux yeux : une PAS de 1540 mmHg, une taille de 1,72 et un poids de 7,26 kg. La valeur de 1540 mmHg n'est pas une curiosité inoffensive. Si nous la laissons dans le fichier, la PAS moyenne calculée sur les 161 lignes devient 139,70 mmHg au lieu de 131,13, et l'écart-type passe de 13,88 à 111,89 mmHg ; la médiane reste à 131, ce qui illustre parfaitement la robustesse de la médiane face aux valeurs aberrantes (chapitre 2). Pour la taille et le poids, la vérification la plus élégante utilise la variable dérivée imc0, qui est restée correcte dans le fichier brut. Pour P057, l'IMC enregistré est de 26,3 kg/m² et le poids de 77,7 kg ; la taille compatible avec ces deux valeurs est $\sqrt{\frac{77,7}{26,3}} = 1,72m$, ce qui confirme que la taille a été écrite en mètres et que la valeur correcte est 172 cm. Pour P102, un IMC de 27,0 et une taille de 164 cm donnent un poids de 27,0 x 1,64² = 72,6 kg : la virgule décimale a été déplacée d'une position. La vérification croisée entre variables transforme un soupçon en certitude et elle est plus sûre que toute supposition.

La troisième étape porte sur le type des colonnes. Lorsqu'un seul texte apparaît dans une colonne numérique ("87 mg/dl" ou "5,2"), le programme d'importation déclare toute la colonne comme texte, et la moyenne ne peut plus être calculée du tout. Le symptôme est facile à reconnaître : JASP affiche la variable avec l'icône de variable nominale, et dans Google Sheets la fonction ISNUMBER renvoie FALSE. La correction (suppression de l'unité ou, selon le cas, remplacement de la virgule par le point) se fait dans la cellule concernée, après quoi la colonne est réimportée comme numérique.

Les valeurs manquantes (l'âge de P042 et la glycémie à 6 mois de P134) ne se complètent pas par la moyenne ou par une valeur "probable" sans un plan d'analyse qui prévoie explicitement la méthode d'imputation. Dans la pratique d'une étude, on cherche d'abord la valeur dans le dossier ; si elle n'existe pas, on indique combien de valeurs manquent pour chaque variable, et l'analyse se fait sur les cas disponibles ou, dans les essais cliniques, selon la stratégie prévue dans le protocole (chapitre 13). Dans notre scénario, la source contient les deux valeurs, qui ont donc été récupérées.

La ligne en double (P091, qui apparaît à la ligne 91 puis de nouveau à la ligne 161) gonfle la taille de l'échantillon à 161 et donne un poids double à un seul patient. On la détecte en vérifiant l'unicité de l'identifiant : dans Google Sheets, =COUNTIF(A:A; A2) > 1 ; dans JASP, en triant par id ; en Python, df.duplicated(). L'une des copies est supprimée, après avoir confirmé que les deux lignes sont identiques dans toutes les colonnes.

> **La règle qui évite l'erreur**
>
> Corrigez les données uniquement d'après le document source, et non d'après votre intuition, même lorsque la correction paraît évidente. Gardez le fichier brut intact, travaillez sur une copie et notez chaque modification dans un journal de nettoyage (la ligne, la variable, l'ancienne valeur, la nouvelle valeur, le motif). Un exposé qui dit "nous avons corrigé 11 erreurs et supprimé un doublon, conformément au journal figurant en annexe" est vérifiable ; un exposé qui dit seulement "les données ont été nettoyées" ne l'est pas.

C.5 Comment le jeu de données a été généré

Les données sont entièrement simulées. Les patients sont fictifs, il n'existe aucun lien avec des personnes réelles, et le jeu de données ne provient d'aucun cabinet ni d'aucun hôpital. Par conséquent, la question de la protection des données personnelles au sens du Règlement général

sur la protection des données [32] ne se pose pas, et les fichiers peuvent être utilisés librement à des fins pédagogiques.

La simulation a été réalisée en Python, avec le générateur de nombres aléatoires de la bibliothèque numpy (default_rng) initialisé avec une graine fixe (2026). La graine fixe rend le jeu de données reproductible : en exécutant de nouveau le script, on obtient exactement les mêmes 160 lignes. Les variables n'ont pas été générées indépendamment, mais au moyen de relations cliniquement plausibles : la pression systolique augmente avec l'âge, avec l'IMC et chez les fumeurs ; l'HbA1c dépend de la glycémie ; le diabète confirmé par HGPO et le score FINDRISC dépendent d'une même variable latente de risque (âge, IMC, tour de taille, activité physique) ; l'effet de l'intervention sur la glycémie et sur le poids augmente du bras A au bras C. La distribution des triglycérides a été générée selon une loi log-normale, afin de disposer d'un exemple de variable asymétrique qui se normalise par transformation logarithmique (chapitre 3). Les tailles d'effet ont été choisies de sorte que le jeu de données contienne, à des fins pédagogiques, aussi bien des résultats clairement significatifs (ANOVA par bras, corrélation IMC-PAS) que des résultats non significatifs (PAS selon le sexe) ou situés à la limite (test de Wilcoxon pour le poids dans le bras A, p = 0,052).

La conséquence pratique est importante et doit être dite honnêtement : les chiffres de mg2026 ne représentent l'estimation d'aucune réalité épidémiologique. La prévalence de 18,1% du diabète, l'effet de l'application mobile ou l'aire sous la courbe ROC du score FINDRISC (0,852) sont des propriétés de la simulation, et non des résultats qui peuvent être cités comme preuves cliniques. Pour des preuves réelles, les chapitres de l'ouvrage s'appuient sur les études publiées des collègues de la discipline et sur la littérature spécialisée.

C.6 Comment citer le jeu de données dans les exposés et les devoirs

Lorsque vous utilisez mg2026 dans un exposé, dans un devoir ou à l'examen pratique, précisez trois choses : que les données sont simulées, quel fichier vous avez utilisé (nettoyé ou brut) et quels traitements vous avez effectués avant l'analyse. Voici un modèle de phrase pour la section des méthodes :

"L'analyse a utilisé le jeu de données pédagogique mg2026 (données simulées, 160 patients fictifs, fichier mg2026_curat.csv), mis à disposition sur la plateforme Moodle de la Discipline d'informatique médicale et de biostatistique, Université de médecine et de pharmacie "Victor Babeș" de Timișoara. Le traitement statistique a été réalisé avec JASP 0.98.1, avec un seuil de signification α = 0,05."

Modèle d'entrée dans la liste bibliographique :

Muntean C. Le jeu de données pédagogique mg2026 (données simulées pour la discipline Biostatistique, Médecine, première année). In : Muntean C. Biostatistică aplicată în medicina generală. Orizont 2026 [Biostatistique appliquée en médecine générale. Horizon 2026]. Timișoara : Éditions "Victor Babeș" ; 2026. Fichiers mg2026_curat.csv et mg2026_brut.csv, disponibles sur moodle.umft.ro.

Les logiciels se citent avec la version utilisée : JASP 0.98.1 [47], MedCalc 24.0 [50], Epi Info 7.2.7 [51]. Si vous avez travaillé sur le fichier brut, ajoutez en annexe de l'exposé le journal de nettoyage (le tableau C.2 en est un modèle) et indiquez le nombre final de lignes analysées. Ne présentez jamais les résultats obtenus sur mg2026 comme s'ils provenaient de patients réels, pas même dans un exercice de rédaction d'un résumé de congrès : l'habitude de déclarer correctement la provenance des données se prend dès la première année.

À retenir

- mg2026 est un jeu de données simulé, reproductible (graine 2026), comprenant 160 patients fictifs répartis en trois bras randomisés (A 53, B 53, C 54) et des mesures à 0, 3 et 6 mois.
- Le fichier nettoyé (160 lignes) sert aux analyses ; le fichier brut (161 lignes) contient 11 erreurs et un doublon, pour l'exercice de nettoyage.
- Ordre des vérifications : fréquences pour les catégories, minimum et maximum pour les variables numériques, type des colonnes, valeurs manquantes, unicité de l'identifiant, cohérence des variables dérivées (IMC).
- Chaque erreur est une déformation de la valeur vraie, mais la correction se confirme dans le document source ; la cohérence avec les variables dérivées (IMC) est la meilleure vérification.
- Dans tout travail, les données sont déclarées comme simulées et les traitements sont documentés.

Annexe D. Fiche de formules

La fiche rassemble les formules utilisées dans le cours, classées selon le chapitre où elles apparaissent. Pour chaque formule, vous trouverez la signification des symboles, les conditions dans lesquelles elle peut être utilisée et un exemple numérique d'une ligne, le plus souvent sur le jeu de données pédagogique mg2026 (annexe C). Les formules sont écrites en texte simple : x signifie multiplication, √ racine carrée, Σ somme de tous les termes, ln logarithme naturel, et e^x la fonction exponentielle. Les résultats intermédiaires ne sont arrondis qu'à la fin, pour ne pas accumuler d'erreurs d'arrondi ; à l'examen, gardez au moins trois décimales dans les calculs intermédiaires.

Notations générales, valables pour toute la fiche : n est la taille de l'échantillon, N l'effectif total (ou celui de la population, selon le contexte), xi une valeur individuelle, m la moyenne de l'échantillon, μ la moyenne de la population, s l'écart-type de l'échantillon, σ l'écart-type de la population, p une proportion observée, α le seuil de signification, β la probabilité de l'erreur de type II.

Pour les tableaux 2 x 2, on utilise deux conventions, l'une pour l'analyse du risque et l'autre pour les tests diagnostiques. Nous les rappelons ici, car la moitié des erreurs à l'examen viennent de leur confusion.

Tableau D.1. Le tableau 2 x 2 dans l'analyse du risque (chapitre 10).

	Maladie présente	**Maladie absente**	**Total**
Exposés	a	b	a + b
Non-exposés	c	d	c + d
Total	a + c	b + d	N

Tableau D.2. Le tableau 2 x 2 d'un test diagnostique (chapitre 12), avec la maladie en lignes, comme dans les travaux pratiques de la discipline.

	Test positif	**Test négatif**	**Total**
Malades (B+)	a = vrais positifs (VP)	b = faux négatifs (FN)	a + b
Non-malades (B-)	c = faux positifs (FP)	d = vrais négatifs (VN)	c + d
Total	a + c	b + d	N

D.1 Chapitre 1. Probabilités

Probabilité classique. P(A) = nombre de cas favorables / nombre de cas possibles. La valeur est comprise entre 0 et 1. Condition : tous les cas sont également probables (par exemple, le tirage au hasard d'un dossier). Exemple : la probabilité qu'un patient tiré au hasard dans mg2026 soit fumeur est de 53/160 = 0,331.

Probabilité de l'événement contraire. $P(\text{non} - A) = 1 - P(A)$. Exemple : $P\big($ non-fumeur ou ancien fumeur $\big) = 1 - 0{,}331 = 0{,}669$.

Règle d'addition. $P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ et } B)$. Si A et B s'excluent mutuellement, P(A et B) = 0 et le terme disparaît. Exemple : $P(\text{ femme ou fumeur }) = \frac{86}{160} + \frac{53}{160} - \frac{30}{160} = \frac{109}{160} = 0{,}681$.

Probabilité conditionnelle. $P(A \mid B) = \frac{P(A \text{ et } B)}{P(B)}$, qui se lit "probabilité de A sachant que B s'est produit". Exemple : $P(\text{ diabète } \mid \text{obèse}) = \frac{\frac{18}{160}}{\frac{54}{160}} = \frac{18}{54} = 0{,}333$.

Règle de multiplication. $P(A \text{ et } B) = P(B) \times P(A \mid B)$. Si A et B sont indépendants, P(A | B) = P(A) et $P(A \text{ et } B) = P(A) \times P(B)$. Exemple : sous l'hypothèse d'indépendance, P(femme et fumeur) serait de 0,5375 x 0,331 = 0,178 ; la valeur observée est de 30/160 = 0,1875, très proche, donc le sexe et le tabagisme semblent indépendants dans ce jeu de données.

Fréquence relative comme estimation de la probabilité. $f = \frac{k}{n}$, où k est le nombre d'apparitions de l'événement en n observations. Exemple : 29 patients avec un diabète confirmé sur 160, f = 0,181.

D.2 Chapitre 2. Description des données

Fréquences absolue, relative et cumulée. $f_i = \frac{n_i}{n}$; fréquence en pourcentage = 100 x ni / n ; fréquence cumulée à la catégorie k = somme des fréquences des catégories 1...k (uniquement pour les variables ordinales ou numériques groupées). Lorsqu'il existe des valeurs manquantes, la fréquence "valide" se calcule sur le nombre de réponses valides, et non sur le total. Exemple : satisfaction 4 ou 5 : (55 + 28)/160 = 51,9% ; fréquence cumulée jusqu'au score 3 inclus : (4 + 19 + 54)/160 = 48,1%.

Moyenne arithmétique. $m = \frac{\sum x_i}{n}$. Conditions : variable numérique ; elle est sensible aux valeurs extrêmes, on la préfère donc pour les distributions approximativement symétriques. Exemple : PAS chez cinq patients 120, 135, 128, 142, 130 mmHg, $m = \frac{655}{5} = 131$ mmHg.

Moyenne pondérée. mp $= \frac{\sum(w_i \times x_i)}{\sum w_i}$, où wi est le poids (habituellement l'effectif du sous-groupe). On l'utilise lorsqu'on combine les moyennes de groupes de tailles différentes. Exemple : diminution moyenne de la glycémie dans les trois bras, $\frac{53\times3,0+53\times11,1+54\times16,3}{160} = 10,17$ mg/dL (la moyenne calculée directement sur tous les patients est de 10,19 ; la différence provient de l'arrondi des moyennes de groupe).

Moyenne géométrique. mg $= (x_1 \times x_2 \times ... \times \text{xn})^{\frac{1}{n}} = e$^(moyenne de ln x). Conditions : toutes les valeurs strictement positives ; elle convient aux variables asymétriques à droite, log-normales (triglycérides, titres d'anticorps, charges virales). Exemple : pour 100, 150 et 225, mg $= (100 \times 150 \times 225)^{\frac{1}{3}} = 150$; pour les triglycérides de mg2026, la moyenne des logarithmes est de 4,970 et mg $= e^{4,970} = 144,0$ mg/dL (la moyenne arithmétique est de 161,18).

Position de la médiane. On classe les valeurs par ordre croissant ; la médiane se trouve à la position (n + 1)/2. Si n est pair, la médiane est la moyenne des deux valeurs centrales (positions n/2 et n/2 + 1). Exemple : pour n = 160, la position est 80,5, donc la médiane de la PAS est la moyenne des valeurs de rangs 80 et 81 : (131 + 131)/2 = 131 mmHg.

Positions des quartiles (méthode simple). Q1 à la position (n + 1)/4, Q3 à la position 3(n + 1)/4, avec interpolation entre les valeurs voisines. Les logiciels utilisent des méthodes d'interpolation légèrement différentes, ils peuvent donc donner des résultats qui diffèrent d'une décimale. Exemple : dans mg2026, Q1 de la PAS = 122 et Q3 = 140 mmHg.

Étendue (amplitude de variation). A = xmax - xmin. Exemple : PAS 169 - 102 = 67 mmHg.

Écart interquartile. EIQ = Q3 - Q1 ; il contient les 50% centraux des valeurs. Exemple : PAS 140 - 122 = 18 mmHg ; IMC $31,4 - 26,2 = \frac{5,2 \text{ kg}}{m^2}$.

Limites pour les valeurs aberrantes (règle de Tukey, utilisée pour le box-plot). Limite inférieure = Q1 - 1,5 x EIQ ; limite supérieure = Q3 + 1,5 x EIQ. Exemple : IMC, $31{,}4 + 1{,}5 \times 5{,}2 = \frac{39{,}2\ \text{kg}}{m^2}$; la valeur maximale de 48,3 apparaît comme un point isolé sur le box-plot.

Écart absolu moyen. EMA $= \frac{\sum|x_i - m|}{n}$. Exemple : pour 120, 135, 128, 142, 130 (m = 131), les écarts absolus sont 11, 4, 3, 11, 1, et EMA $= \frac{30}{5} = 6$ mmHg.

Variance de l'échantillon. $s^2 = \frac{\sum(x_i - m)^2}{n-1}$. Le dénominateur n - 1 (les degrés de liberté) corrige la tendance à sous-estimer la variance de la population. Exemple : pour les mêmes cinq valeurs, $\sum(x_i - m)^2 = 121 + 16 + 9 + 121 + 1 = 268$, $s^2 = \frac{268}{4} = 67$ mmHg^2.

Écart-type. $s = \sqrt{s^2}$. Il a la même unité de mesure que les données. Exemple : $s = \sqrt{67} =$ 8,19 mmHg ; dans mg2026, s(PAS) = 13,88 mmHg.

Coefficient de variation. CV = 100 x s / m (%). Conditions : variable de rapport, à moyenne positive ; il permet de comparer la variabilité de variables d'unités différentes. À titre indicatif, un CV inférieur à 10% indique une variabilité faible, un CV de 10-30% une variabilité modérée et un CV supérieur à 30% une variabilité élevée. Exemple : PAS, $\frac{100 \times 13{,}88}{131{,}13} = 10{,}6\%$; triglycérides, 47,3%.

Erreur standard de la moyenne. ES $= \frac{s}{\sqrt{n}}$. Elle mesure la précision de la moyenne en tant qu'estimation de μ, et non la dispersion des patients. Exemple : PAS, 13,88/√160 = 13,88/12,649 = 1,097 mmHg.

D.3 Chapitre 3. Loi normale, score z, transformations

Score z (standardisation). $z = \frac{x-m}{s}$ (sur l'échantillon) ou $z = \frac{x-\mu}{\sigma}$ (sur la population). Il indique à combien d'écarts-types de la moyenne se trouve une valeur. Condition pour lire les aires dans la table de la loi normale : distribution approximativement normale. Exemple : PAS 160 mmHg, $z = \frac{160-131{,}13}{13{,}88} = 2{,}08$; d'après la table E.1, Φ(2,08) = 0,9812, donc environ 1,9% des valeurs dépasseraient 160 mmHg.

Transformation inverse. x = m + z x s. Exemple : le 97,5e percentile de la PAS (z = 1,96) est de 131,13 + 1,96 x 13,88 = 158,3 mmHg.

Intervalles pour les valeurs individuelles (règle 68-95-99,7). Pour une distribution normale, environ 68,3% des valeurs se situent dans $m \pm s$, 95,4% dans $m \pm 2s$, 99,7% dans $m \pm 3s$, et exactement 95% dans $m \pm 1{,}96s$. Exemple (PAS) : $m \pm s = 117{,}3 - 145{,}0$; $m \pm 1{,}96s = 103{,}9 - 158{,}3$ (contient 93,8% des observations réelles) ; $m \pm 2s = 103{,}4 - 158{,}9$; $m \pm 3s = 89{,}5 - 172{,}8$ mmHg.

Intervalles pour la moyenne de la population (à partir de l'erreur standard). μ se situe, avec les probabilités correspondantes, dans $m \pm \text{ES}(68{,}3\%)$, $m \pm 2\text{ES}(95{,}4\%)$, $m \pm 3\text{ES}(99{,}7\%)$ ou $m \pm 1{,}96\text{ES}(95\%)$. Condition : n suffisamment grand (théorème central limite) ou données normales. Exemple (PAS, ES = 1,097) : $m \pm \text{ES} = 130{,}03 - 132{,}23$; $m \pm 2\text{ES} = 128{,}94 - 133{,}32$; $m \pm 3\text{ES} = 127{,}84 - 134{,}42$ mmHg.

Transformation logarithmique. y = ln(x) ; après l'analyse sur l'échelle logarithmique, les résultats sont ramenés sur l'échelle initiale avec $x = e^y$. Condition : x > 0 ; utile en cas d'asymétrie positive. Exemple : l'asymétrie des triglycérides passe de 0,95 à -0,28 après

transformation logarithmique, et l'IC 95% de la moyenne géométrique est $e^{4,970 \pm \frac{1,975 \times 0,488}{\sqrt{160}}} = 133,44 - 155,42$ mg/dL.

Asymétrie et aplatissement (interprétation). Une asymétrie (skewness) proche de 0 indique la symétrie ; positive, une queue vers la droite. Un excès d'aplatissement (kurtosis moins 3) proche de 0 indique une forme proche de la forme normale. Règle indicative : des valeurs absolues supérieures à 1 suggèrent un écart important. Exemple : IMC, asymétrie 1,42 et excès d'aplatissement 3,68, donc distribution non normale, ce que confirme le test de Shapiro-Wilk ($p < 0,001$).

D.4 Chapitre 4. Estimation, intervalles de confiance, taille de l'échantillon

IC d'une moyenne avec la loi normale (σ connu ou n grand). $IC(1-\alpha) = m \pm \frac{z\left(1-\frac{\alpha}{2}\right) \times \sigma}{\sqrt{n}}$; z = 1,645 (90%), 1,960 (95%), 2,576 (99%). Exemple : PAS, $131,13 \pm 1,96 \times 1,097 = 128,98 - 133,28$ mmHg.

IC d'une moyenne avec la loi de Student (σ inconnu, cas habituel). $IC(1-\alpha) = m \pm \frac{t(\alpha;\, n-1) \times s}{\sqrt{n}}$. Condition : données approximativement normales ou n grand. Exemple : PAS, t(0,05 ; 159) = 1,975, $IC\,95\% = 131,13 \pm 1,975 \times 1,097 = 128,96 - 133,30$ mmHg ; IC 99% = 128,27-133,99 mmHg.

Largeur de l'IC et effet de n. Largeur = $\frac{2 \times t \times s}{\sqrt{n}}$; si n est multiplié par 4, la largeur diminue approximativement de moitié. Exemple : l'IC de la moyenne de la PAS a une largeur de 27,2 mmHg pour n = 10, de 10,6 pour n = 40 et de 4,3 pour n = 160.

Erreur standard d'une proportion. $ES(p_i) = \sqrt{\frac{p_i \times (1-p_i)}{n}}$. Exemple : diabète 29/160, p_i = 0,181, $ES = \sqrt{\frac{0,181 \times 0,819}{160}} = 0,0305$.

IC d'une proportion, méthode de Wald. $IC\,95\% = p_i \pm 1,96 \times \sqrt{\frac{p_i(1-p_i)}{n}}$. Conditions : n x $p_i \geq 5$ et n x (1 - p_i) ≥ 5 (certains auteurs exigent 10) ; pour les petites proportions ou les petits échantillons, elle donne des intervalles trop étroits ou des limites négatives. Exemple : $0,181 \pm 1,96 \times 0,0305 = 0,122 - 0,241$, c'est-à-dire 12,2-24,1%.

IC d'une proportion, méthode de Wilson (recommandée). Centre = (p_i + z^2/(2n)) / (1 + z^2/n) ; demi-largeur = $\left[\frac{z}{1+\frac{z^2}{n}}\right] \times \sqrt{\frac{p_i(1-p_i)}{n} + \frac{z^2}{4n^2}}$; IC = centre ± demi-largeur. Elle fonctionne bien même pour les proportions proches de 0 ou de 1. Exemple : pour 29/160 et z = 1,96, centre = 0,1887, demi-largeur = 0,0595, IC 95% = 12,9-24,8%.

IC de la moyenne géométrique. $IC = e^{ml \pm \frac{t \times sl}{\sqrt{n}}}$, où ml et sl sont la moyenne et l'écart-type de ln x. Exemple : triglycérides, 133,44-155,42 mg/dL (voir D.3).

Taille de l'échantillon pour l'estimation d'une moyenne. $n = \left(\frac{z\left(1-\frac{\alpha}{2}\right) \times \sigma}{E}\right)^2$, où E est la marge d'erreur souhaitée (demi-largeur de l'IC). On arrondit toujours à l'entier supérieur. Exemple : σ = 14 mmHg, $E = \pm 3$ mmHg, $n = \left(\frac{1,96 \times 14}{3}\right)^2 = 83,7$, donc 84 patients.

Taille de l'échantillon pour l'estimation d'une proportion. $n = \frac{z^2 \times p_i(1-p_i)}{E^2}$. Si p_i est inconnu, on prend $p_i = 0,5$ (l'option la plus prudente). Exemple : $p_i = 0,20$, E = 0,05, $n = \frac{1,96^2 \times 0,2 \times 0,8}{0,0025} = 245,9$, donc 246.

Effectif par groupe pour la comparaison de deux moyennes (test t pour échantillons indépendants). n par groupe = 2 x (z(1-α/2) + z(1-β))² / d², où $d = \frac{\mu_1 - \mu_2}{\sigma}$ est la taille de l'effet ; z(1-β) = 0,842 pour une puissance de 80% et 1,282 pour une puissance de 90%. Exemple : d = 0,5, α = 0,05, puissance 80% : $\frac{2 \times (1,96+0,842)^2}{0,25} = 62,8$, c'est-à-dire 63 par groupe ; les logiciels qui utilisent la loi t (G*Power) donnent 64 par groupe, et pour d = 0,3 ils donnent 176.

Effectif par groupe pour la comparaison de deux proportions. n par groupe = (z(1-α/2) + z(1-β))² x [p1(1 - p1) + p2(1 - p2)] / (p1 - p2)². Exemple : p1 = 0,30, p2 = 0,11, α = 0,05, puissance 80% : $\frac{7,849 \times (0,21+0,0979)}{0,0361} = 66,9$, donc 67 par groupe.

Correction pour les abandons. n ajusté $= \frac{n}{1-r}$, où r est la proportion estimée de perdus de vue. Exemple : 64 par groupe et 15% d'abandons, 64/0,85 = 75,3, donc 76 par groupe.

D.5 Chapitre 5. Bases des tests d'hypothèses

Erreurs et puissance. α = P(on rejette H0 | H0 vraie), l'erreur de type I ; β = P(on ne rejette pas H0 | H0 fausse), l'erreur de type II ; puissance = 1 - β ; niveau de confiance = 1 - α. Exemple : une étude planifiée avec α = 0,05 et une puissance de 80% accepte un risque de 5% de résultat faussement positif et de 20% de manquer un effet réel de l'ampleur supposée.

Règle de décision avec p. Si p < α, on rejette H0 (différence statistiquement significative) ; si p ≥ α, on ne rejette pas H0 (aucune différence n'a été démontrée). Exemple : PAS femmes vs hommes, p = 0,81 > 0,05, donc on ne rejette pas H0.

Règle de décision avec la valeur critique. On rejette H0 si |statistique calculée| ≥ valeur critique de la table (annexe E) pour α et les degrés de liberté correspondants. Exemple : t = 11,14 avec ddl = 159 > t critique 1,975, donc p < 0,05.

Test z pour une moyenne (σ connu). $z = \frac{m - \mu_0}{\frac{\sigma}{\sqrt{n}}}$. Conditions : σ connu à partir de la population ou n grand, données approximativement normales. Exemple : PAS moyenne de 131,13 comparée à une valeur de référence μ0 = 128 mmHg, avec σ = 14 : $z = \frac{3,13}{\frac{14}{\sqrt{160}}} = 2,83$, p bilatéral = 0,005.

Test t pour un échantillon. $t = \frac{m - \mu_0}{\frac{s}{\sqrt{n}}}$, avec ddl = n - 1. Conditions : données approximativement normales (ou n grand), observations indépendantes. Exemple : la même comparaison avec s = 13,88 : $t = \frac{3,13}{1,097} = 2,85$, ddl = 159, p = 0,005.

Test z pour une proportion. $z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$. Conditions : n x p0 ≥ 5 et n x (1 - p0) ≥ 5. Exemple : diabète 0,181 comparé à une prévalence supposée p0 = 0,12 : $z = \frac{0,061}{\sqrt{\frac{0,12 \times 0,88}{160}}} = \frac{0,061}{0,0257} = 2,38$, p bilatéral = 0,017.

Correction de Bonferroni pour comparaisons multiples. α' = α / k, où k est le nombre de comparaisons. Exemple : trois comparaisons deux à deux entre les bras A, B, C, α' = 0,05/3 = 0,0167.

D.6 Chapitre 6. Tests paramétriques : t et ANOVA

Test t apparié. d = différence pour chaque patient ; md = moyenne des différences, sd = leur écart-type ; $t = \frac{\text{md}}{\frac{\text{sd}}{\sqrt{n}}}$, ddl = n - 1. Conditions : mesures appariées (même patient, deux temps de mesure ou deux méthodes), différences approximativement normales. Exemple : glycémie à 0 contre 6 mois, md = 10,19, sd = 11,57, n = 160 : $t = \frac{10{,}19}{\frac{11{,}57}{12{,}649}} = 11{,}14$, ddl = 159, p < 0,001.

IC de la moyenne des différences appariées. $\text{IC}\,95\% = \text{md} \pm \frac{t(0{,}05;\, n-1) \times \text{sd}}{\sqrt{n}}$. Exemple : $10{,}19 \pm 1{,}975 \times 0{,}915 = 8{,}38 - 11{,}99$ mg/dL.

Variance commune (pooled) pour deux groupes indépendants. $\text{sp}^2 = \frac{(n_1-1)\times {s_1}^2 + (n_2-1)\times {s_2}^2}{n_1+n_2-2}$. Exemple : PAS, hommes (130,85 ; 13,25 ; 74) et femmes (131,37 ; 14,47 ; 86) : $\text{sp}^2 = \frac{73\times 175{,}56 + 85\times 209{,}38}{158} = 193{,}76$, sp = 13,92 mmHg.

Test t pour échantillons indépendants classique (de Student, variances égales). $t = \frac{m_1 - m_2}{\text{sp}\times\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, ddl = n1 + n2 - 2. Conditions : groupes indépendants, données approximativement normales dans chaque groupe, variances égales (test de Levene non significatif). Exemple : $t = -\frac{0{,}52}{13{,}92\times\sqrt{\frac{1}{74}+\frac{1}{86}}} = -0{,}24$, ddl = 158, p = 0,81.

Test t de Welch (variances inégales). $t = \frac{m_1 - m_2}{\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}}$; $\text{ddl} = \frac{\left(\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}\right)^2}{\frac{\left(\frac{{s_1}^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{{s_2}^2}{n_2}\right)^2}{n_2-1}}$. On peut l'utiliser dans tous les cas, et pas seulement lorsque les variances sont inégales ; JASP le propose comme option. Exemple : PAS selon le sexe, $t = -\frac{0{,}52}{2{,}193} = -0{,}24$, ddl = 157,4, p = 0,81.

IC de la différence de deux moyennes indépendantes. $\text{IC}\,95\% = (m_1 - m_2) \pm t \times \text{ES}(m_1 - m_2)$, avec $\text{ES} = \text{sp} \times \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ ou l'ES de Welch. Exemple : différence H - F de la PAS, -0,5 mmHg, IC 95% de -4,9 à 3,8 mmHg (l'intervalle contient 0).

Taille d'effet d de Cohen (groupes indépendants). $d = \frac{m_1 - m_2}{\text{sp}}$. Repères indicatifs [76] : 0,2 faible, 0,5 moyen, 0,8 fort. Exemple : PAS H vs F, $d = -\frac{0{,}52}{13{,}92} = -0{,}04$ (effet négligeable).

Taille d'effet pour données appariées. $\text{dz} = \frac{\text{md}}{\text{sd}}$. Exemple : glycémie 0 vs 6 mois, $\text{dz} = \frac{10{,}19}{11{,}57} = 0{,}88$.

ANOVA à un facteur : les sommes des carrés. Moyenne générale mG ; mj et nj, moyenne et effectif du groupe j ; k, nombre de groupes ; N, effectif total. SST (totale) = $\sum(x_i - m_G)^2$SSB (intergroupes) = $\sum n_j \times \left(m_j - m_G\right)^2$SSW (intragroupe, résiduelle) = $\sum\left(x_i - m_j\right)^2 = \sum\left(n_j - 1\right) \times {s_j}^2$SST = SSB + SSW. Exemple : diminution de la glycémie à 6 mois selon les bras, SSB = 4 806,9, SSW = 16 459,5, SST = 21 266,4.

ANOVA à un facteur : le tableau et la statistique F. ddl intergroupes = k - 1 ; ddl résiduels = N - k ; $\text{MSB} = \frac{\text{SSB}}{k-1}$; $\text{MSW} = \frac{\text{SSW}}{N-k}$; $F = \frac{\text{MSB}}{\text{MSW}}$. Conditions : groupes indépendants, normalité dans chaque groupe (ou des résidus), variances homogènes (Levene). Exemple : $\text{MSB} = \frac{4806{,}9}{2} = 2403{,}4$; $\text{MSW} = \frac{16459{,}5}{157} = 104{,}84$; F(2 ; 157) = 22,93, p < 0,001.

Tableau D.3. Structure du tableau d'ANOVA, avec les valeurs de l'exemple.

Source de variation	**SS**	**ddl**	$\mathrm{MS} = \frac{\mathrm{SS}}{\mathrm{ddl}}$	**F**
Intergroupes	4806,9	2	2403,4	22,93
Intragroupe (résiduelle)	16459,5	157	104,84	
Total	21266,4	159		

Êta carré. $\eta^2 = \frac{\mathrm{SSB}}{\mathrm{SST}}$; proportion de la variabilité totale expliquée par l'appartenance au groupe. Repères indicatifs : 0,01 faible, 0,06 moyen, 0,14 fort. Exemple : 4 806,9/21 266,4 = 0,226.

D.7 Chapitre 7. Tests non paramétriques et test du khi-deux

Fréquences attendues. E = (total de la ligne x total de la colonne) / N. Exemple : dans le tableau obésité x diabète, $E(\text{ obèses avec diabète }) = \frac{54 \times 29}{160} = 9{,}79$; les autres : 44,21 ; 19,21 ; 86,79.

Statistique du khi-deux (de Pearson). $\chi^2 = \frac{\sum(O-E)^2}{E}$, sommée sur toutes les cases, avec ddl = (r - 1) x (c - 1). Conditions : observations indépendantes, fréquences absolues (et non pourcentages) ; au plus 20% des cases avec E < 5 et aucune case avec E < 1 ; sinon, test exact de Fisher (pour un 2 x 2) ou regroupement des catégories. Exemple : obésité x diabète, χ^2 = 12,70, ddl = 1, p < 0,001.

Formule rapide pour le tableau 2 x 2. $\chi^2 = \frac{N \times (\mathrm{ad} - \mathrm{bc})^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)}$. Exemple : $\frac{160 \times (18 \times 95 - 36 \times 11)^2}{54 \times 106 \times 29 \times 131} = 12{,}70$.

Correction de continuité de Yates (2 x 2). $\chi^2 Y = \frac{N \times \left(|\mathrm{ad} - \mathrm{bc}| - \frac{N}{2}\right)^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)}$. Exemple : 11,20, p = 0,0008.

Degrés de liberté. ddl = (nombre de lignes - 1) x (nombre de colonnes - 1). Exemple : bras (3) x perte ≥5% (2), ddl = 2 x 1 = 2 ; χ^2 = 6,77, p = 0,034.

Test U de Mann-Whitney. On classe toutes les valeurs ensemble et on attribue des rangs (en cas d'ex aequo, le rang moyen). R1 = somme des rangs du groupe 1. $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$; U2 = n1 x n2 - U1 ; U = min(U1, U2) pour la table. Conditions : deux groupes indépendants, variable au moins ordinale. Exemple : groupe A {12, 15, 18}, groupe B {14, 20, 22, 25} ; les rangs de A sont 1, 3, 4, R1 = 8, U1 = 8 - 6 = 2, U2 = 12 - 2 = 10, U = 2.

Approximation normale de U (grands échantillons). $z = \frac{U - \frac{\mathrm{n1n2}}{2}}{\sqrt{\frac{n_1 \times n_2 \times (n_1 + n_2 + 1)}{12}}}$. Exemple : satisfaction bras C vs A, U = 2 265,5, n1 = 54, n2 = 53 : $z = \frac{2265{,}5 - 1431}{\sqrt{\frac{2862 \times 108}{12}}} = \frac{834{,}5}{160{,}5} = 5{,}2$, p < 0,001 (sans correction pour les ex aequo).

Corrélation bisériale de rangs (taille d'effet du test de Mann-Whitney). r = 2 x U1/(n1 x n2) - 1, où U1 correspond au groupe aux rangs les plus élevés. Exemple : 2 x 2 265,5/2 862 - 1 = 0,58.

Test de Wilcoxon pour données appariées (des rangs signés). On calcule les différences, on élimine les différences nulles, on classe les valeurs absolues et on attribue des rangs ; W+ = somme des rangs des différences positives, W- = somme des rangs des différences négatives, et

$W^+ + W^- = \frac{n(n+1)}{2}$. La statistique de test est min(W+, W-). Exemple : pour 6 différences non nulles, $W^+ + W^- = \frac{6\times7}{2} = 21$.

Test de Kruskal-Wallis. $H = \left[\frac{12}{N(N+1)}\right] \times \sum\left(\frac{{R_j}^2}{n_j}\right) - 3(N+1)$, avec ddl = k - 1 (approximation du khi-deux). Conditions : k ≥ 3 groupes indépendants, variable au moins ordinale. Exemple : diminution de la glycémie selon les bras, H = 36,63, ddl = 2, p < 0,001.

Test de Friedman. $\chi^2 F = \left[\frac{12}{n\times k\times(k+1)}\right] \times \sum {R_j}^2 - 3n(k+1)$, où Rj est la somme des rangs pour le temps de mesure j (les rangs sont attribués à l'intérieur de chaque patient), ddl = k - 1. Exemple : glycémie à 0, 3 et 6 mois, χ^2F = 74,67, ddl = 2, p < 0,001.

Coefficient de corrélation de Spearman. $\text{rho} = 1 - \frac{6\times\sum d^2}{n(n^2-1)}$, où d est la différence entre le rang de x et le rang de y chez un même individu (formule exacte en l'absence d'ex aequo). Exemple : les rangs x 1, 2, 3, 4, 5 et y 2, 1, 4, 3, 5 donnent d = -1, 1, -1, 1, 0, $\sum d^2 = 4$, $\text{rho} = 1 - \frac{24}{120} = 0{,}80$.

D.8 Chapitre 8. Choix du test (repères numériques)

Vérification de la normalité. Test de Shapiro-Wilk : p ≥ 0,05 signifie qu'aucun écart à la normalité n'a été démontré (on peut utiliser les tests paramétriques) ; p < 0,05 oriente vers une transformation ou vers des tests non paramétriques. Exemple : PAS, p = 0,303 (paramétrique) ; IMC, p < 0,001 (non paramétrique ou transformation).

Vérification de l'homogénéité des variances. Test de Levene : p ≥ 0,05 permet le test t classique ou l'ANOVA classique ; p < 0,05 impose Welch. Exemple : PAS selon le sexe, Levene p = 0,529.

Règle des fréquences attendues. Khi-deux si toutes les E ≥ 5 (ou au plus 20% des cases inférieures à 5) ; Fisher dans le cas contraire. Exemple : bras x perte ≥5%, la plus petite E = 12,6, donc khi-deux.

D.9 Chapitre 9. Corrélation et régression

Coefficient de corrélation de Pearson. $r = \frac{\sum(x_i-m_x)(y_i-m_y)}{\sqrt{\left[\sum(x_i-m_x)^2\times\sum(y_i-m_y)^2\right]}}$, avec des valeurs comprises entre -1 et 1. Conditions : les deux variables numériques, relation linéaire, distribution approximativement normale, pas de valeurs extrêmes influentes. Exemple : IMC et PAS dans mg2026, r = 0,405.

Test t pour le coefficient de corrélation. $t = \frac{r\times\sqrt{n-2}}{\sqrt{1-r^2}}$, ddl = n - 2 ; H0 : ρ = 0. Exemple : $\frac{0{,}405\times\sqrt{158}}{\sqrt{1-0{,}164}} = 5{,}57$, p < 0,001.

Coefficient de détermination. $R^2 = r^2$ (en régression simple) ; proportion de la variabilité de y expliquée par x. Exemple : 0,405² = 0,164, c'est-à-dire que l'IMC explique 16,4% de la variabilité de la PAS.

Pente de la droite de régression (méthode des moindres carrés). $b = \frac{\sum(x_i-m_x)(y_i-m_y)}{\sum(x_i-m_x)^2} = \frac{r\times s_y}{s_x}$. Signification : variation moyenne de y lorsque x augmente d'une unité. Exemple : $b = \frac{0{,}405\times13{,}88}{3{,}91} = 1{,}44$ (le logiciel, avec des données non arrondies, donne 1,435 mmHg par kg/m²).

Ordonnée à l'origine. a = my - b x mx. Exemple : a = 131,13 - 1,435 x 29,04 = 89,46 (logiciel : 89,47 mmHg).

Équation de régression et prédiction. ŷ = a + b x x. Condition : la prédiction n'a de sens que dans l'intervalle des valeurs de x observées. Exemple : PAS = 89,47 + 1,435 x IMC ; pour un IMC de 30, ŷ = 132,5 mmHg ; pour un IMC de 35, ŷ = 139,7 mmHg.

Résidu. ei = yi - ŷi. Exemple : un patient avec un IMC de 30 et une PAS de 140 a un résidu de 140 - 132,5 = 7,5 mmHg.

Erreur standard de l'estimation (SEE). $\text{SEE} = \sqrt{\left[\frac{\sum {e_i}^2}{n-2}\right]} \approx s_y \times \sqrt{1-r^2} \times \sqrt{\frac{n-1}{n-2}}$. Exprimée dans les unités de y, elle indique à quelle distance de la droite se trouvent, en moyenne, les points. Exemple : $13{,}88 \times \sqrt{0{,}836} \times \sqrt{\frac{159}{158}} = 12{,}73$ mmHg.

IC de la pente. $b \pm t(0{,}05;\ n-2) \times \text{ES}(b)$. Exemple : 1,435, IC 95% 0,925-1,944 ; l'intervalle ne contient pas 0, donc la pente est significative.

Coefficient de détermination ajusté (régression multiple). $R^2_{\text{aj}} = 1 - \frac{(1-R^2)\times(n-1)}{n-p-1}$, où p est le nombre de prédicteurs (variables du modèle, y compris les variables indicatrices). Exemple : PAS ~ IMC + âge + sexe + tabagisme (5 termes), $R^2 = 0{,}308$, $R^2_{\text{aj}} = 1 - \frac{0{,}692\times 159}{154} = 0{,}286$.

Régression logistique : du coefficient à l'OR. $\text{OR} = e^b$; $\text{IC}\,95\% = e^{b \pm 1{,}96 \times \text{ES}(b)}$. Exemple : pour l'âge, OR = 1,11 par an, c'est-à-dire que la cote de diabète augmente d'environ 11% pour chaque année supplémentaire, à obésité et sexe constants.

D.10 Chapitre 10. Épidémiologie et analyse du risque

Prévalence. P = nombre de cas existants à un moment donné / population à ce moment. Exemple : diabète confirmé à l'inclusion, 29/160 = 18,1%.

Incidence cumulée (risque). IC = nombre de nouveaux cas au cours d'une période / population à risque au début de la période. Exemple : si 12 personnes sur 400 sans diabète développent la maladie en 5 ans, IC = 3,0% en 5 ans.

Densité d'incidence (taux). DI = nombre de nouveaux cas / somme des personnes-temps de suivi. Exemple : 12 nouveaux cas pour 1 850 personnes-années, DI = 6,5 pour 1 000 personnes-années.

Relation approximative prévalence-incidence (maladie stable, rare). P ≈ DI x D, où D est la durée moyenne de la maladie. Exemple : DI = 2 pour 1 000 par an et durée de 10 ans, P ≈ 2%.

Risque relatif (RR) ou rapport de prévalences (RP). $\text{RR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$. Il se calcule à partir des études de cohorte et des essais cliniques (RR), ou des études transversales (RP). RR = 1 signifie l'absence d'association, RR > 1 un facteur de risque, RR < 1 un facteur protecteur. Exemple : obésité et diabète, (18/54)/(11/106) = 0,333/0,104 = 3,21.

IC du RR sur l'échelle logarithmique. $\text{ES}[\ln(\text{RR})] = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$; $\text{IC}\,95\% = e^{\ln \text{RR} \pm 1{,}96 \times \text{ES}}$. Exemple : ln 3,21 = 1,167, ES = 0,344, $\text{IC}\,95\% = e^{1{,}167 \pm 0{,}675} = 1{,}64 - 6{,}31$.

Rapport de cotes (odds ratio, OR). $\mathrm{OR} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \times d}{b \times c}$. Seule mesure d'association valide dans les études cas-témoins ; il approche le RR lorsque la maladie est rare (moins d'environ 10%). Exemple : $\frac{18 \times 95}{36 \times 11} = \frac{1710}{396} = 4{,}32$.

IC de l'OR sur l'échelle logarithmique (méthode de Woolf). $\mathrm{ES}[\ln(\mathrm{OR})] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$; $\mathrm{IC}\,95\% = e^{\ln \mathrm{OR} \pm 1{,}96 \times \mathrm{ES}}$. Exemple : ln 4,32 = 1,463, ES = 0,430, IC 95% = 1,86-10,03. Si l'IC ne contient pas 1, l'association est significative au seuil de 5%.

Différence de risques (risque attribuable chez les exposés, réduction absolue du risque). $\mathrm{DR} = p_1 - p_0 = \frac{a}{a+b} - \frac{c}{c+d}$. $\mathrm{IC}\,95\% = \mathrm{DR} \pm 1{,}96 \times \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}}$. Exemple : perte ≥5% dans le bras C par rapport au bras A, 16/54 - 6/53 = 0,296 - 0,113 = 0,183, IC 95% 0,034-0,332, soit 18,3 points de pourcentage (3,4-33,2).

Nombre de sujets à traiter (NST). $\mathrm{NNT} = \frac{1}{|\mathrm{DR}|}$, arrondi à l'entier supérieur ; les bornes de l'IC s'obtiennent en inversant les bornes de l'IC de la DR. Pour un effet nocif, la même formule donne le nombre de sujets à traiter pour nuire (NNH). Exemple : 1/0,183 = 5,5, donc 6 patients traités dans le bras C au lieu du bras A pour qu'un patient de plus perde au moins 5% de son poids ; IC 95% 3,0-29,1.

Réduction relative du risque. $\mathrm{RRR} = \frac{p_0 - p_1}{p_0} = 1 - \mathrm{RR}$ (pour un traitement qui diminue le risque). Exemple : si la mortalité passe de 10% à 7%, RR = 0,7, RRR = 30%, DR = 3 points de pourcentage, NST = 34.

Fraction attribuable chez les exposés. FAe = (RR - 1) / RR. Exemple : (3,21 - 1)/3,21 = 0,689 ; 68,9% des cas de diabète chez les obèses seraient attribuables à l'obésité, si la relation était causale et non confondue.

Fraction attribuable dans la population (formule de Levin). FAP = pe x (RR - 1) / [1 + pe x (RR - 1)], où pe est la proportion d'exposés dans la population ; de façon équivalente, $\mathrm{FAP} = \frac{\mathrm{Pt} - P_0}{\mathrm{Pt}}$. Exemple : pe = 54/160 = 0,3375, $\mathrm{FAP} = \frac{0{,}3375 \times 2{,}21}{1 + 0{,}3375 \times 2{,}21} = 0{,}427$; vérification : (0,1813 - 0,1038)/0,1813 = 0,427.

Test d'hypothèse pour le RR et l'OR. H0 : RR = 1 (ou OR = 1) ; on utilise le test du khi-deux sur le tableau 2 x 2 (D.7) ou le test exact de Fisher. Exemple : χ^2 = 12,70, p < 0,001, ce qui concorde avec l'IC de l'OR qui exclut 1.

D.11 Chapitre 11. Bayes et médecine fondée sur les preuves

Théorème de Bayes. $P(B \mid T) = \frac{P(T|B) \times P(B)}{P(T)}$, avec $P(T) = P(T \mid B) \times P(B) + P(T \mid \text{non} - B) \times P(\text{non} - B)$. Exemple : voir la formule de la VPP ci-dessous.

Probabilité post-test d'un test positif (VPP à partir de la Se, de la Sp et de la prévalence). $\mathrm{VPP} = \frac{\mathrm{Se} \times P}{\mathrm{Se} \times P + (1 - \mathrm{Sp}) \times (1 - P)}$. Exemple : prévalence 10%, Se 80%, Sp 90% : 0,08/(0,08 + 0,09) = 0,471 (47%).

VPN à partir de la Se, de la Sp et de la prévalence. $\mathrm{VPN} = \frac{\mathrm{Sp} \times (1 - P)}{\mathrm{Sp} \times (1 - P) + (1 - \mathrm{Se}) \times P}$. Exemple : mêmes valeurs, 0,81/(0,81 + 0,02) = 0,976.

Conversion probabilité-cote (odds). cote = p / (1 - p) ; p = cote / (1 + cote). Exemple : p = 0,181, cote = 0,181/0,819 = 0,221 ; une cote de 0,666 correspond à $p = \frac{0{,}666}{1{,}666} = 0{,}40$.

Rapport de vraisemblance positif. $\mathrm{RV}^{+} = \frac{\mathrm{Se}}{1-\mathrm{Sp}}$. Interprétation indicative : au-dessus de 10, forte augmentation de la probabilité de la maladie ; 5-10, augmentation modérée ; 2-5, faible ; 1, le test n'apporte aucune information. Exemple : FINDRISC ≥ 12, 0,759/0,252 = 3,01.

Rapport de vraisemblance négatif. $\mathrm{RV}^{-} = \frac{1-\mathrm{Se}}{\mathrm{Sp}}$. Au-dessous de 0,1, forte diminution de la probabilité de la maladie. Exemple : 0,241/0,748 = 0,32.

Cote post-test. cote post-test = cote pré-test x RV ; puis probabilité post-test = cote/(1 + cote). C'est la forme pratique du théorème de Bayes, sur laquelle repose le nomogramme de Fagan [122]. Exemple : probabilité pré-test 18,1% (cote 0,221), test positif avec un RV+ de 3,01 : cote post-test 0,666, probabilité post-test 40,0%, identique à la VPP du tableau.

D.12 Chapitre 12. Tests diagnostiques et courbe ROC

Toutes les formules ci-dessous utilisent la notation du tableau D.2. L'exemple commun est le score FINDRISC avec le seuil ≥ 12 par rapport au diabète confirmé par HGPO : a = 22, b = 7, c = 33, d = 98, N = 160.

Prévalence (dans l'échantillon de l'étude). $P = \frac{a+b}{N}$. Exemple : 29/160 = 18,1%.

Sensibilité. $\mathrm{Se} = \frac{a}{a+b} = P(T^{+} \mid B^{+})$. Exemple : 22/29 = 75,9%.

Spécificité. $\mathrm{Sp} = \frac{d}{c+d} = P(T^{-} \mid B^{-})$. Exemple : 98/131 = 74,8%.

Valeur prédictive positive. $\mathrm{VPP} = \frac{a}{a+c} = P(B^{+} \mid T^{+})$. Elle dépend de la prévalence. Exemple : 22/55 = 40,0%.

Valeur prédictive négative. $\mathrm{VPN} = \frac{d}{b+d} = P(B^{-} \mid T^{-})$. Exemple : 98/105 = 93,3%.

Exactitude (proportion des classements corrects). $\mathrm{Ex} = \frac{a+d}{N}$. Exemple : (22 + 98)/160 = 75,0%.

Taux d'erreur de classement. $\mathrm{TE} = \frac{b+c}{N} = \frac{\mathrm{FN+FP}}{N} = 1 - \mathrm{Ex}$. Attention : dans certains supports circule la formule erronée (b + d)/N ; la formule correcte est (b + c)/N. Exemple : (7 + 33)/160 = 25,0%.

Proportion des faux positifs et des faux négatifs. Taux de faux positifs = $\frac{c}{c+d} = 1 - \mathrm{Sp}$; taux de faux négatifs = $\frac{b}{a+b} = 1 - \mathrm{Se}$. Exemple : 33/131 = 25,2% ; 7/29 = 24,1%.

Indice de Youden. J = Se + Sp - 1, entre 0 (test inutile) et 1 (test parfait). Le seuil optimal selon ce critère est celui qui maximise J [134]. Exemple : au seuil FINDRISC ≥ 11, J = 0,931 + 0,664 - 1 = 0,595 ; au seuil ≥ 12, J = 0,759 + 0,748 - 1 = 0,507.

Rapport de cotes diagnostique (DOR). $\mathrm{DOR} = \frac{\mathrm{RV}^{+}}{\mathrm{RV}^{-}} = \frac{a \times d}{b \times c}$. Exemple : $\frac{22 \times 98}{7 \times 33} = 9{,}33$.

Aire sous la courbe ROC (AUC), interprétation. AUC = probabilité qu'un malade choisi au hasard ait un score plus élevé qu'un non-malade choisi au hasard ; 0,5 = test inutile, 1 = test parfait. Repères utilisés dans le cours : 0,91-1 excellent, 0,81-0,90 très bon, 0,71-0,80 bon, 0,61-0,70 satisfaisant, au-dessous de 0,60 faible. Exemple : FINDRISC, AUC = 0,852 (IC 95% 0,762-0,943).

Reconstitution du tableau 2 x 2 à partir des indicateurs. Si l'on connaît N, la prévalence P, la Se et la Sp : a + b = P x N ; a = Se x (a + b) ; b = (a + b) - a ; c + d = N - (a + b) ; d = Sp x (c + d) ; c = (c + d) - d. Si l'on connaît la VPP et a : $a + c = \frac{a}{\mathrm{VPP}}$. Si l'on connaît la VPN et b : $b + d =$

$\frac{b}{1-\text{VPN}}$. Exemple : N = 1 000, P = 10%, Se = 80%, Sp = 90% donnent a = 80, b = 20, c = 90, d = 810, et $\text{VPP} = \frac{80}{170} = 47{,}1\%$.

Coefficient kappa de Cohen (accord entre deux évaluateurs). κ = (po - pe) / (1 - pe), où po est la proportion d'accord observé et pe = Σ(proportion marginale de l'évaluateur 1 x proportion marginale de l'évaluateur 2), somme sur les catégories. Interprétation selon Landis et Koch [103] : 0,21-0,40 faible, 0,41-0,60 modéré, 0,61-0,80 important, au-dessus de 0,80 presque parfait. Exemple : deux médecins lisent 100 ECG ; tous deux concluent "pathologique" pour 20, tous deux "normal" pour 60, désaccord pour 20 (10 + 10). po = 0,80 ; chaque médecin a déclaré pathologiques 30% des tracés, donc pe = 0,3 x 0,3 + 0,7 x 0,7 = 0,58 ; κ = 0,22/0,42 = 0,52 (accord modéré).

D.13 Chapitre 13. Essais cliniques (formules de planification)

Les formules de planification (taille de l'échantillon pour deux moyennes ou deux proportions et correction pour les abandons) se trouvent à la section D.4. Dans le compte rendu d'un essai clinique, on utilise, pour un critère de jugement principal binaire, le RR, la DR et le NST (D.10), et pour un critère numérique la différence des moyennes avec son IC 95% (D.6).

Taux d'abandon et taux de réponse. Taux d'abandon = perdus de vue / randomisés ; taux de participation = randomisés / évalués pour l'éligibilité. Exemple : dans le scénario mg2026, 160 randomisés sur 212 évalués, taux de participation = 75,5%.

Randomisation par blocs (nombre de séquences possibles). Pour un bloc de 4 avec 2 attributions A et 2 attributions B, il existe $\frac{4!}{2!\times 2!} = 6$ séquences possibles (AABB, ABAB, ABBA, BAAB, BABA, BBAA).

À retenir

- Dans le tableau diagnostique avec la maladie en lignes : $\text{Se} = \frac{a}{a+b}$, $\text{Sp} = \frac{d}{c+d}$, $\text{VPP} = \frac{a}{a+c}$, $\text{VPN} = \frac{d}{b+d}$, taux d'erreur = (b + c)/N.
- Dans le tableau de risque avec l'exposition en lignes : $\text{RR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$, $\text{OR} = \frac{\text{ad}}{\text{bc}}$; l'IC se calcule sur l'échelle logarithmique, il est donc asymétrique autour de l'estimation.
- $\text{ES} = \frac{s}{\sqrt{n}}$ décrit la précision de la moyenne ; s décrit la dispersion des patients. L'intervalle $m \pm 1{,}96s$ concerne les valeurs individuelles, $m \pm 1{,}96\text{ES}$ concerne la moyenne.
- $\text{NNT} = \frac{1}{\text{DR}}$, arrondi à l'entier supérieur ; le RR et l'OR ne disent rien de l'ampleur absolue du bénéfice.
- Les tailles d'échantillon s'arrondissent toujours à l'entier supérieur et se corrigent pour les abandons.

Annexe E. Tables statistiques

Les tables de cette annexe ont été calculées avec la bibliothèque scipy (Python), au moyen des fonctions de répartition inverses des lois normale, de Student, du khi-deux et de Fisher, et n'ont pas été recopiées d'autres manuels. Les logiciels statistiques (JASP, MedCalc, Epi Info) donnent directement la valeur p, de sorte que dans la pratique quotidienne les tables sont rarement nécessaires. À l'examen et en séance de travaux dirigés, elles restent toutefois utiles : elles vous obligent à comprendre ce que signifie une valeur critique, pourquoi un test bilatéral exige un seuil plus élevé qu'un test unilatéral et comment la valeur critique de t diminue à mesure que le nombre de degrés de liberté augmente. Toutes les décimales sont écrites avec une virgule.

La règle générale de lecture est la même pour les tableaux 3-7 : calculez la statistique de test à partir des données, trouvez dans la table la valeur critique pour les degrés de liberté et le seuil α choisi, puis comparez. Si la valeur calculée (en valeur absolue) est supérieure ou égale à la valeur critique, $p \leq \alpha$ et on rejette H0. Si elle est inférieure, on ne rejette pas H0.

E.1 Loi normale centrée réduite : fonction de répartition Φ(z)

La table donne Φ(z), c'est-à-dire l'aire sous la courbe normale centrée réduite à gauche de la valeur z (la probabilité $P(Z \leq z)$). La ligne donne la partie entière et la première décimale de z, et la colonne donne la deuxième décimale. À partir de Φ(z), on obtient les autres aires dont vous avez besoin :

- aire entre 0 et z : Φ(z) - 0,5 ;
- aire de la queue droite ($P(Z > z)$) : 1 - Φ(z) ;
- aire entre -z et +z : 2 x Φ(z) - 1 ;
- valeur p bilatérale pour un score z observé : 2 x (1 - Φ(|z|)) ;
- pour z négatif : Φ(-z) = 1 - Φ(z), car la courbe est symétrique.

Exemple. Dans le jeu de données mg2026, le score z d'une pression systolique de 160 mmHg est de 2,08. À la ligne 2,0, colonne 0,08, nous lisons Φ(2,08) = 0,9812. Par conséquent, environ 1 - 0,9812 = 0,0188, soit 1,9% des patients, devraient avoir une PAS supérieure à 160 mmHg si la distribution était exactement normale, et l'aire comprise entre la moyenne et 160 mmHg est de 0,9812 - 0,5 = 0,4812.

Tableau E.1. Fonction de répartition Φ(z) de la loi normale centrée réduite, z = 0,00-3,49.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

E.2 Valeurs critiques de la loi normale centrée réduite

La valeur critique z est le point au-delà duquel se trouve une aire égale à α (test unilatéral) ou deux aires de α/2 chacune (test bilatéral). Les mêmes valeurs servent à construire les intervalles de confiance : pour un IC à 95%, on utilise le z bilatéral pour α = 0,05, c'est-à-dire 1,960 ; pour 99%, 2,576 ; pour 90%, 1,645.

Exemple. Un test z bilatéral donne z = 2,30. La valeur critique pour α = 0,05 en bilatéral est 1,960, et pour α = 0,01 en bilatéral, 2,576. Comme 1,960 < 2,30 < 2,576, il en résulte 0,01 < p < 0,05 (la valeur exacte, d'après la table E.1, est 2 x (1 - 0,9893) = 0,0214).

Tableau E.2. Valeurs critiques z pour les seuils usuels.

α	z critique unilatéral (z1-α)	z critique bilatéral (z1-α/2)	Niveau de l'IC bilatéral (1 - α)
0,200	0,842	1,282	80,0%
0,100	1,282	1,645	90,0%
0,050	1,645	1,960	95,0%
0,025	1,960	2,241	97,5%
0,020	2,054	2,326	98,0%
0,010	2,326	2,576	99,0%
0,005	2,576	2,807	99,5%
0,001	3,090	3,291	99,9%

E.3 Valeurs critiques de la loi de Student (t)

La table donne la valeur t pour laquelle l'aire des deux queues prises ensemble est égale à α (test bilatéral). Pour un test unilatéral au seuil α, lisez la colonne 2α (par exemple, le test unilatéral à α = 0,05 utilise la colonne 0,10). Les degrés de liberté (ddl) sont n - 1 pour le test t sur un échantillon et pour le test t apparié, n1 + n2 - 2 pour le test t pour échantillons indépendants classique et n - 2 pour le test du coefficient de corrélation. La ligne ∞ coïncide avec les valeurs

critiques z du tableau E.2. Si le nombre de degrés de liberté ne figure pas dans la table, utilisez la ligne immédiatement inférieure (le résultat est conservateur).

Exemple. Pour l'IC 95% de la moyenne de la PAS sur un sous-échantillon de n = 10 patients, ddl = 9 et t(0,05 ; 9) = 2,262. Avec une moyenne de 130,0 et un écart-type s, $IC = 130{,}0 \pm \frac{2{,}262 \times s}{\sqrt{10}}$. Pour n = 160 (ddl = 159), on utilise la ligne 120, t = 1,980, très proche de 1,96.

Tableau E.3. Valeurs critiques t bilatérales.

ddl	$\alpha = 0{,}20$	$\alpha = 0{,}10$	$\alpha = 0{,}05$	$\alpha = 0{,}02$	$\alpha = 0{,}01$	$\alpha = 0{,}001$
1	3,078	6,314	12,706	31,821	63,657	636,619
2	1,886	2,920	4,303	6,965	9,925	31,599
3	1,638	2,353	3,182	4,541	5,841	12,924
4	1,533	2,132	2,776	3,747	4,604	8,610
5	1,476	2,015	2,571	3,365	4,032	6,869
6	1,440	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,408
8	1,397	1,860	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,250	4,781
10	1,372	1,812	2,228	2,764	3,169	4,587
11	1,363	1,796	2,201	2,718	3,106	4,437
12	1,356	1,782	2,179	2,681	3,055	4,318
13	1,350	1,771	2,160	2,650	3,012	4,221
14	1,345	1,761	2,145	2,624	2,977	4,140
15	1,341	1,753	2,131	2,602	2,947	4,073
16	1,337	1,746	2,120	2,583	2,921	4,015
17	1,333	1,740	2,110	2,567	2,898	3,965
18	1,330	1,734	2,101	2,552	2,878	3,922
19	1,328	1,729	2,093	2,539	2,861	3,883
20	1,325	1,725	2,086	2,528	2,845	3,850
21	1,323	1,721	2,080	2,518	2,831	3,819
22	1,321	1,717	2,074	2,508	2,819	3,792
23	1,319	1,714	2,069	2,500	2,807	3,768
24	1,318	1,711	2,064	2,492	2,797	3,745
25	1,316	1,708	2,060	2,485	2,787	3,725
26	1,315	1,706	2,056	2,479	2,779	3,707
27	1,314	1,703	2,052	2,473	2,771	3,690
28	1,313	1,701	2,048	2,467	2,763	3,674
29	1,311	1,699	2,045	2,462	2,756	3,659
30	1,310	1,697	2,042	2,457	2,750	3,646
40	1,303	1,684	2,021	2,423	2,704	3,551
50	1,299	1,676	2,009	2,403	2,678	3,496
60	1,296	1,671	2,000	2,390	2,660	3,460
80	1,292	1,664	1,990	2,374	2,639	3,416
100	1,290	1,660	1,984	2,364	2,626	3,390
120	1,289	1,658	1,980	2,358	2,617	3,373
∞	1,282	1,645	1,960	2,326	2,576	3,291

E.4 Valeurs critiques de la loi du khi-deux (χ^2)

Le test du khi-deux d'association est toujours unilatéral, sur la queue droite : les grandes valeurs de χ^2 indiquent un écart important des fréquences observées par rapport aux fréquences attendues. La table donne la valeur de χ^2 qui laisse à droite une aire égale à α. Pour un tableau de contingence à r lignes et c colonnes, ddl = (r - 1) x (c - 1) ; un tableau 2 x 2 a ddl = 1.

Exemple. Dans mg2026, le tableau bras x perte de poids d'au moins 5% a 3 lignes et 2 colonnes, donc ddl = 2, et le χ^2 calculé = 6,77. La valeur critique pour $\alpha = 0,05$ et ddl = 2 est 5,991, et pour $\alpha = 0,01$, 9,210. Par conséquent, $0,01 < p < 0,05$ (le logiciel donne p = 0,034) et on rejette H0 au seuil de 5%.

Tableau E.4. Valeurs critiques de χ^2 (aire à droite = α).

ddl	$\alpha = 0,10$	$\alpha = 0,05$	$\alpha = 0,01$	$\alpha = 0,001$
1	2,706	3,841	6,635	10,828
2	4,605	5,991	9,210	13,816
3	6,251	7,815	11,345	16,266
4	7,779	9,488	13,277	18,467
5	9,236	11,070	15,086	20,515
6	10,645	12,592	16,812	22,458
7	12,017	14,067	18,475	24,322
8	13,362	15,507	20,090	26,124
9	14,684	16,919	21,666	27,877
10	15,987	18,307	23,209	29,588
11	17,275	19,675	24,725	31,264
12	18,549	21,026	26,217	32,909
13	19,812	22,362	27,688	34,528
14	21,064	23,685	29,141	36,123
15	22,307	24,996	30,578	37,697
16	23,542	26,296	32,000	39,252
17	24,769	27,587	33,409	40,790
18	25,989	28,869	34,805	42,312
19	27,204	30,144	36,191	43,820
20	28,412	31,410	37,566	45,315
21	29,615	32,671	38,932	46,797
22	30,813	33,924	40,289	48,268
23	32,007	35,172	41,638	49,728
24	33,196	36,415	42,980	51,179
25	34,382	37,652	44,314	52,620
26	35,563	38,885	45,642	54,052
27	36,741	40,113	46,963	55,476
28	37,916	41,337	48,278	56,892
29	39,087	42,557	49,588	58,301
30	40,256	43,773	50,892	59,703

E.5 Valeurs critiques de la loi de Fisher (F) pour α = 0,05

La loi F a deux nombres de degrés de liberté. Dans l'ANOVA à un facteur, ddl1 (numérateur) = k - 1, où k est le nombre de groupes, et ddl2 (dénominateur) = N - k, où N est le nombre total d'observations. Les colonnes de la table correspondent à ddl1, les lignes à ddl2. Si le F calculé est supérieur ou égal à la valeur de la table, $p \leq 0,05$. La même table sert, avec ddl1 = nombre de prédicteurs et ddl2 = n - p - 1, au test F global d'une régression multiple.

Exemple. ANOVA de la diminution de la glycémie dans les trois bras de mg2026 : k = 3, N = 160, donc ddl1 = 2 et ddl2 = 157. La ligne 157 n'existe pas dans la table ; nous utilisons la ligne 120 (choix conservateur) : F critique = 3,072. La valeur calculée, F = 22,93, dépasse largement le seuil, donc $p < 0,05$ (le logiciel donne $p < 0,001$).

Tableau E.5. Valeurs critiques de F pour α = 0,05 (ddl1 en colonnes, ddl2 en lignes).

ddl2 \ ddl1	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

ddl2 / ddl1	1	2	3	4	5	6	7	8	9	10
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91
200	3,89	3,04	2,65	2,42	2,26	2,14	2,06	1,98	1,93	1,88
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83

E.6 Valeurs critiques du coefficient de corrélation de Pearson (r)

Pour décider si un coefficient de corrélation observé diffère significativement de zéro, comparez |r| à la valeur critique correspondant à la taille de l'échantillon n. Les valeurs proviennent de la relation entre r et le test t, avec ddl = n - 2 :

$$r_{critique} = t_{critique} / \sqrt{(ddl + t_{critique}^2)}$$

La table est bilatérale (H1 : $\rho \neq 0$). Observez à quel point le seuil diminue à mesure que n augmente : avec 10 patients, il vous faut $|r| \geq 0{,}632$ pour avoir $p < 0{,}05$, alors qu'avec 100 patients $|r| \geq 0{,}197$ suffit. D'où l'avertissement du chapitre 9 : sur les grands échantillons, des corrélations faibles, sans importance clinique, deviennent "significatives".

Exemple. La corrélation entre l'IMC et la PAS dans mg2026 est r = 0,405, avec n = 160. Même le seuil pour n = 100 à $\alpha = 0{,}01$ (0,256) est dépassé, donc $p < 0{,}01$. Pour n = 160, le seuil est encore plus bas.

Tableau E.6. Valeurs critiques bilatérales du r de Pearson.

n	ddl $= n - 2$	$\alpha = 0{,}05$	$\alpha = 0{,}01$
5	3	0,878	0,959
6	4	0,811	0,917
7	5	0,754	0,875
8	6	0,707	0,834
9	7	0,666	0,798
10	8	0,632	0,765
11	9	0,602	0,735
12	10	0,576	0,708
13	11	0,553	0,684
14	12	0,532	0,661
15	13	0,514	0,641
16	14	0,497	0,623
17	15	0,482	0,606
18	16	0,468	0,590
19	17	0,456	0,575
20	18	0,444	0,561
21	19	0,433	0,549
22	20	0,423	0,537
23	21	0,413	0,526
24	22	0,404	0,515
25	23	0,396	0,505
26	24	0,388	0,496
27	25	0,381	0,487
28	26	0,374	0,479
29	27	0,367	0,471
30	28	0,361	0,463
35	33	0,334	0,430
40	38	0,312	0,403
45	43	0,294	0,380
50	48	0,279	0,361
55	53	0,266	0,345
60	58	0,254	0,330
65	63	0,244	0,317
70	68	0,235	0,306
75	73	0,227	0,296
80	78	0,220	0,286
85	83	0,213	0,278
90	88	0,207	0,270
95	93	0,202	0,263
100	98	0,197	0,256

E.7 Valeurs critiques approximatives du coefficient de corrélation de Spearman (rho)

Pour les très petits échantillons, la distribution de rho sous H0 n'est pas continue et les valeurs critiques s'obtiennent par l'énumération de toutes les permutations des rangs. Dans la table ci-

dessous, les valeurs pour n = 5-10 sont exactes (calculées par l'énumération complète des n! permutations), et celles pour n = 11-30 sont approchées par la formule fondée sur la loi t à n - 2 degrés de liberté, la même que pour le r de Pearson. Le tiret signifie que, pour ce n, aucune valeur de rho ne peut atteindre le seuil requis. L'approximation t sous-estime légèrement le seuil exact pour n compris entre 11 et 15 (d'environ 0,01-0,02) ; un rho situé juste à la limite doit donc être jugé d'après la valeur p exacte. La table est bilatérale ; règle de décision : |rho| ≥ valeur critique signifie p ≤ α. Pour la présentation des résultats, utilisez la valeur p exacte donnée par le logiciel.

Exemple. Un médecin de famille met en corrélation, chez 12 patients, le score d'activité physique (1-5) et le nombre de consultations par an, et obtient rho = -0,61. La valeur critique pour n = 12 et α = 0,05 est 0,576 ; comme |-0,61| ≥ 0,576, la corrélation négative est significative au seuil de 5%, mais pas à 1% (le seuil est de 0,708).

Tableau E.7. Valeurs critiques bilatérales du rho de Spearman (exactes pour n ≤ 10, approximatives pour n ≥ 11).

n	$\alpha = 0{,}05$	$\alpha = 0{,}01$	Méthode
5	1,000	-	exacte
6	0,886	1,000	exacte
7	0,786	0,929	exacte
8	0,738	0,881	exacte
9	0,700	0,833	exacte
10	0,648	0,794	exacte
11	0,602	0,735	approximation t
12	0,576	0,708	approximation t
13	0,553	0,684	approximation t
14	0,532	0,661	approximation t
15	0,514	0,641	approximation t
16	0,497	0,623	approximation t
17	0,482	0,606	approximation t
18	0,468	0,590	approximation t
19	0,456	0,575	approximation t
20	0,444	0,561	approximation t
21	0,433	0,549	approximation t
22	0,423	0,537	approximation t
23	0,413	0,526	approximation t
24	0,404	0,515	approximation t
25	0,396	0,505	approximation t
26	0,388	0,496	approximation t
27	0,381	0,487	approximation t
28	0,374	0,479	approximation t
29	0,367	0,471	approximation t
30	0,361	0,463	approximation t

E.8 Table de 1 000 chiffres aléatoires

Les chiffres ont été générés avec le générateur numpy (default_rng) initialisé avec la graine 2026, de sorte que la table peut être reproduite à tout moment. Ils sont disposés sur 25 lignes de 40 chiffres chacune, groupés par cinq pour faciliter la lecture. La table s'utilise pour les exercices d'échantillonnage aléatoire simple et de randomisation des chapitres 1, 4 et 13.

Mode d'emploi. (1) Numérotez les unités de la population, par exemple les dossiers des patients de 001 à N. (2) Choisissez un point de départ sans regarder les chiffres (par exemple, fermez les yeux et posez le crayon sur la table) et un sens de lecture (de gauche à droite, puis la

ligne suivante). (3) Lisez des groupes d'autant de chiffres que N en comporte (trois chiffres pour N = 160). (4) Gardez les nombres compris entre 001 et N, ignorez ceux qui sont en dehors de l'intervalle et ceux qui se répètent, jusqu'à obtenir la taille d'échantillon voulue.

Exemple. Nous voulons tirer 5 patients parmi les 160 du jeu de données mg2026, en partant du début de la ligne 01 et en lisant les chiffres par trois. Nous ne retenons que les nombres 001-160, dans l'ordre où ils apparaissent, sans répétition. Le résultat pour ce point de départ est donné sous la table, afin que vous puissiez vérifier.

Tableau E.8. 1 000 chiffres aléatoires (graine 2026).

Ligne	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
01	81063	40363	87797	18602	19792	66715	68644	31212
02	85042	68044	93114	53462	92989	75623	49755	44903
03	46430	21742	54456	11795	46033	47484	06759	47007
04	85737	43019	49302	37773	35641	75998	96211	92002
05	92160	40030	86945	64558	39112	92756	06057	99275
06	78417	61417	30265	83693	06635	68142	30409	54295
07	96987	37383	01462	82435	19965	74802	46779	89366
08	66541	10339	22016	36172	09330	52266	11272	41167
09	85241	24723	00595	05164	32134	56909	50882	13892
10	47172	81972	86476	02643	51950	47556	04495	18293
11	87760	28747	67504	28828	32252	26460	02323	94571
12	14641	69584	43551	91538	60028	85564	53427	67334
13	67098	92451	30143	08122	59013	10926	87030	94945
14	09852	44726	82752	08805	00095	97714	55106	19394
15	47814	17618	98872	93577	34433	92417	84012	12557
16	97438	07541	23055	39400	31889	54868	11037	55010
17	38782	39765	17337	41838	43883	35657	84889	48999
18	39498	11627	75755	25167	71771	06440	26623	43235
19	56213	78739	53584	03942	83309	83429	91282	38286
20	10159	41014	43622	95551	57032	59306	36123	85676
21	82625	39050	36236	76037	82381	93643	30979	59035
22	40148	62882	92537	00918	66700	00941	31970	36424
23	47546	80472	81077	29606	39208	23633	41589	75280
24	78458	90542	81342	66013	84467	16883	78189	88248
25	02231	19480	09902	46801	05552	79698	07652	33131

Vérification pour l'exemple ci-dessus : en lisant depuis le début de la table par groupes de trois chiffres, les cinq premiers nombres distincts compris entre 001 et 160 sont 036, 021, 121, 042, 034. Si vous êtes partis d'un autre point de la table, votre échantillon sera différent, et tout aussi valide.

Annexe F. QCM d'entraînement pour l'examen

L'examen théorique de biostatistique est un QCM de 50 questions, corrigé au format du concours de résidanat roumain. Cette annexe vous propose trois tests complets, construits sur le même modèle, afin que vous puissiez vous entraîner dans des conditions aussi proches que possible des conditions réelles. Chaque test couvre l'ensemble des 14 chapitres du cours, de façon à peu près proportionnelle à leur volume, et mêle des questions de théorie, d'interprétation et de calcul.

Comment les questions sont construites et comment elles sont notées

Les questions sont de deux types, signalés au début de l'énoncé :

- **QRU (question à réponse unique)** : parmi les cinq propositions (a-e), une seule est correcte.
- **QRM (question à réponses multiples)** : parmi les cinq propositions, au moins deux et au plus quatre sont correctes.

Les règles exactes de notation (combien de points vaut chaque type de question et comment est traitée une réponse QRM partielle) sont fixées par la méthodologie de l'université pour les examens corrigés au format du concours de résidanat et sont annoncées par la discipline avant la session ; je vous prie de les consulter à cet endroit, et non dans ce livre. Le principe à retenir est simple : une réponse est considérée comme correcte lorsqu'elle correspond exactement à la grille. Pour une QRU, cela signifie la proposition correcte et elle seule. Pour une QRM, la stratégie sûre consiste à cocher toutes les propositions correctes et aucune proposition fausse ; une proposition fausse cochée "par sécurité" ou une proposition correcte omise peut vous coûter les points de la question.

Quelques conseils pratiques, tirés de l'expérience de la correction de ces examens :

- Lisez l'énoncé jusqu'au bout. La différence entre "intervalle pour les valeurs individuelles" et "intervalle de confiance de la moyenne", ou entre "pourcentage valide" et "pourcentage du total", tient souvent à un seul mot.
- Pour les calculs, écrivez au brouillon le tableau 2 x 2 complet (a, b, c, d et les totaux) avant de calculer le moindre indicateur. La plupart des erreurs viennent de la lecture de la mauvaise case.
- Vérifiez l'ordre de grandeur du résultat : une probabilité ne peut pas dépasser 1, une sensibilité de 90% ne peut pas donner une VPP de 150%, une taille d'échantillon s'arrondit à l'entier supérieur.
- Pour une QRM, jugez chaque proposition séparément, comme s'il s'agissait d'une question de type vrai/faux.
- Pour les calculs, vous n'avez besoin que d'opérations simples et, parfois, des tables de l'annexe E.

Faites chaque test en une seule séance, sans interruption et sans regarder les réponses. Notez votre temps. Ce n'est qu'ensuite que vous vérifiez la grille de la section "Réponses et explications", à la fin de l'annexe, et que vous lisez les explications des questions auxquelles vous avez mal répondu. Les explications sont données pour toutes les questions de calcul et pour celles qui contiennent un piège ; pour les questions de simple définition, relisez la section correspondante du chapitre ou le glossaire (annexe B). Les formules sont rassemblées dans l'annexe D.

Les nombres signalés comme provenant du jeu de données mg2026 sont ceux qui ont été calculés dans les chapitres du cours sur le jeu de données pédagogique simulé (annexe C) ; les autres données des énoncés ont été construites pour l'exercice.

Test 1

1. (QRU) Laquelle des opérations suivantes est un exemple d'inférence statistique ?

a) Le calcul de la PAS moyenne des 160 patients du jeu de données mg2026

b) La construction de l'histogramme de la glycémie à jeun des patients de l'échantillon

c) Le dénombrement des fumeurs de l'échantillon

d) L'estimation, à partir de l'échantillon, de l'intervalle dans lequel se situe la PAS moyenne de l'ensemble des adultes à risque cardiométabolique de la région

e) L'établissement du tableau de fréquences du niveau d'études

2. (QRM) À propos des méthodes d'échantillonnage, les affirmations correctes sont :

a) Dans l'échantillonnage aléatoire simple, chaque individu de la population a la même probabilité d'être sélectionné

b) Un échantillon de convenance devient représentatif s'il comporte plus de 1 000 personnes

c) L'échantillonnage stratifié garantit la présence de toutes les strates de la population dans l'échantillon

d) L'erreur systématique de sélection diminue à mesure que la taille de l'échantillon augmente

e) Dans l'échantillonnage en grappes (clusters), on tire au hasard des groupes entiers, par exemple des cabinets

3. (QRU) Dans un cabinet de médecine générale, 30% des patients sont fumeurs, 20% ont une hypertension artérielle et 8% sont des fumeurs hypertendus. La probabilité qu'un patient choisi au hasard soit fumeur ou hypertendu est de :

a) 0,50

b) 0,42

c) 0,06

d) 0,58

e) 0,38

4. (QRU) La classe fonctionnelle NYHA de l'insuffisance cardiaque (I, II, III, IV) est une variable :

a) nominale

b) de rapport

c) continue

d) d'intervalle

e) ordinale

5. (QRM) Les durées d'hospitalisation (en jours) de cinq patients sont 4, 7, 7, 9, 13. Les affirmations correctes sont :

a) La moyenne est de 8 jours

b) La médiane est de 7 jours

c) Le mode est de 7 jours

d) L'étendue est de 13 jours

e) Si la valeur 13 était remplacée par 43, la médiane augmenterait

6. (QRU) Une variable a une moyenne de 80 et un écart-type de 12. Le coefficient de variation est de :

a) 6,7%

b) 12%

c) 15%

d) 66,7%

e) 0,15%

7. (QRM) À propos du box-plot, les affirmations correctes sont :

a) La ligne à l'intérieur de la boîte représente la moyenne arithmétique

b) Les bords de la boîte correspondent au premier et au troisième quartile

c) La hauteur de la boîte est l'écart interquartile

d) Les points représentés séparément, au-delà des moustaches, sont des valeurs aberrantes selon la règle 1,5 x EIQ

e) C'est le graphique recommandé pour les variables nominales

8. (QRU) Dans une population, le cholestérol total suit une loi normale de moyenne 200 mg/dL et d'écart-type 40 mg/dL. Le pourcentage approximatif des personnes dont le cholestérol dépasse 260 mg/dL est de :

a) 6,7%

b) 2,3%

c) 13,4%

d) 15,9%

e) 43,3%

9. (QRM) À propos de la loi normale, les affirmations correctes sont :

a) Elle est symétrique par rapport à la moyenne

b) La moyenne, la médiane et le mode coïncident

c) Environ 95,4% des valeurs se situent dans l'intervalle $m \pm 2s$

d) L'aire totale sous la courbe dépend de l'écart-type

e) Un test de Shapiro-Wilk avec p = 0,30 prouve que les données sont normales

10. (QRU) Dans le jeu de données mg2026, les triglycérides ont une asymétrie de 0,95 et un test de Shapiro-Wilk à p < 0,001. La façon la plus appropriée de décrire la tendance centrale et la dispersion, sans transformation, est :

a) la moyenne ± l'écart-type

b) la médiane et l'écart interquartile

c) le mode et l'étendue

d) la moyenne et l'erreur standard

e) uniquement le coefficient de variation

11. (QRU) Dans le jeu de données mg2026, la PAS a une moyenne de 131,13 mmHg et un écart-type de 13,88 mmHg. Le score z d'une valeur de 160 mmHg est de :

a) 0,21

b) 1,96

c) 2,08

d) 2,88

e) 28,87

12. (QRU) Sur un échantillon de 100 patients, la moyenne d'une variable est de 150 et l'écart-type de 20. L'intervalle de confiance à 95% de la moyenne de la population (avec z = 1,96) est :

a) 110,8-189,2

b) 148,0-152,0

c) 149,6-150,4

d) 146,1-153,9

e) 130,0-170,0

13. (QRM) À propos de l'écart-type (ÉT) et de l'erreur standard de la moyenne (ES), les affirmations correctes sont :

a) $\text{ES} = \frac{\text{ÉT}}{\sqrt{n}}$

b) L'ÉT décrit la dispersion des patients, et l'ES la précision de la moyenne

c) Si la taille de l'échantillon est multipliée par 4, l'ES diminue approximativement de moitié

d) Si la taille de l'échantillon augmente, l'ÉT diminue proportionnellement à $\sqrt{n}$

e) L'intervalle $m \pm 1{,}96\text{ES}$ contient environ 95% des patients

14. (QRU) Dans un échantillon de 200 femmes enceintes, 40 présentent une anémie. L'intervalle de confiance à 95% de la prévalence de l'anémie, calculé par la méthode de Wald, est approximativement :

a) 18,0-22,0%

b) 14,5-25,5%

c) 10,0-30,0%

d) 16,0-24,0%

e) 19,4-20,6%

15. (QRU) Quelle taille d'échantillon faut-il pour estimer une prévalence totalement inconnue avec une marge d'erreur de ±5% et un niveau de confiance de 95% ?

a) 100

b) 196

c) 246

d) 385

e) 1537

16. (QRU) La valeur p d'un test statistique représente :

a) la probabilité d'obtenir un résultat au moins aussi extrême que celui observé, si l'hypothèse nulle était vraie

b) la probabilité que l'hypothèse nulle soit vraie

c) la probabilité que l'hypothèse alternative soit vraie

d) la probabilité de l'erreur de type II

e) la taille de l'effet du traitement

17. (QRM) À propos des erreurs statistiques et de la puissance du test, les affirmations correctes sont :

a) L'erreur de type I consiste à rejeter l'hypothèse nulle alors qu'elle est vraie

b) β est la probabilité de l'erreur de type II

c) La puissance du test est égale à 1 - α

d) La puissance augmente avec la taille de l'échantillon

e) Abaisser le seuil α de 0,05 à 0,01, à taille d'échantillon égale, augmente la puissance

18. (QRU) Dans mg2026, la différence de PAS entre hommes et femmes est de -0,5 mmHg (IC 95% de -4,9 à 3,8 mmHg), p = 0,81. La conclusion correcte est :

a) il a été démontré que la PAS moyenne est égale dans les deux sexes

b) on accepte l'hypothèse nulle

c) on ne rejette pas l'hypothèse nulle ; les données sont compatibles aussi bien avec une absence de différence qu'avec des différences de quelques mmHg dans un sens ou dans l'autre

d) les hommes ont une PAS significativement plus basse

e) la puissance de l'étude est de 81%

19. (QRM) Pour un test z bilatéral au seuil $\alpha = 0,05$, les affirmations correctes sont :

a) La valeur critique est 1,96

b) La valeur critique est 1,645

c) Un z calculé de 2,30 correspond à un $p < 0,05$

d) Un z calculé de 2,30 correspond à un $p < 0,01$

e) Si l'IC 95% de la différence ne contient pas 0, alors $p < 0,05$

20. (QRU) Chez 16 patients, on mesure la PAS avant et après un traitement. La moyenne des différences est de 4 mmHg et l'écart-type des différences de 8 mmHg. La statistique t du test t apparié est :

a) 0,5

b) 1,0

c) 8,0

d) 16,0

e) 2,0

21. (QRM) Pour la situation de la question 20 (t apparié, n = 16), les affirmations correctes sont :

a) Le nombre de degrés de liberté est 15

b) La valeur critique bilatérale de t pour $\alpha = 0,05$ est 2,131

c) On rejette l'hypothèse nulle au seuil bilatéral de 0,05

d) La valeur p bilatérale est comprise entre 0,05 et 0,10

e) Dans un test unilatéral, avec un sens fixé avant l'étude et $\alpha = 0,05$, le résultat serait significatif

22. (QRU) Pour comparer la diminution moyenne de la glycémie entre trois bras indépendants d'une étude, avec des données normales et des variances homogènes, on utilise :

a) le test t apparié

b) trois tests t pour échantillons indépendants, sans aucune correction

c) l'ANOVA à un facteur

d) le test du khi-deux

e) le test de Friedman

23. (QRM) Dans mg2026, l'ANOVA de la diminution de la glycémie à 6 mois selon les bras donne $F(2 ; 157) = 22,93$, $p < 0,001$, $\eta^2 = 0,226$, et le test de Tukey donne pour la différence C-B $p = 0,024$. Les affirmations correctes sont :

a) L'ANOVA seule montre quels bras diffèrent entre eux

b) Les degrés de liberté correspondent à 3 groupes et 160 patients

c) Environ 22,6% de la variabilité de la diminution de la glycémie est expliquée par le bras

d) La différence entre les bras C et B est significative au seuil de 0,05

e) Le résultat impose de refaire l'analyse avec le test de Kruskal-Wallis

24. (QRU) L'alternative non paramétrique du test t apparié est :

a) le test de Mann-Whitney

b) le test des rangs signés de Wilcoxon

c) le test de Kruskal-Wallis

d) le test de Friedman

e) le test du khi-deux

25. (QRU) Un tableau de contingence a 3 lignes et 4 colonnes. Le nombre de degrés de liberté du test du khi-deux est :

a) 6

b) 7

c) 12

d) 5

e) 2

26. (QRU) Dans un tableau de contingence avec N = 200, le total d'une ligne est 60 et le total d'une colonne est 50. La fréquence attendue dans la case située à leur intersection est :

a) 10

b) 12

c) 15

d) 25

e) 30

27. (QRM) À propos du test du khi-deux d'association, les affirmations correctes sont :

a) Il se calcule sur les fréquences absolues

b) Il peut être appliqué directement aux pourcentages du tableau

c) Pour un tableau 2 x 2 avec des fréquences attendues inférieures à 5, on préfère le test exact de Fisher

d) Une valeur élevée de χ^2 indique un écart important des fréquences observées par rapport aux fréquences attendues

e) La valeur de χ^2 indique directement le sens et l'ampleur de l'association

28. (QRU) La satisfaction à l'égard de l'intervention (échelle de Likert 1-5) est comparée entre le bras C et le bras A. Le test approprié est :

a) le test t pour échantillons indépendants

b) le test des rangs signés de Wilcoxon

c) le test t apparié

d) le test de Mann-Whitney

e) le coefficient de corrélation de Pearson

29. (QRM) Lesquelles des situations suivantes exigent un test pour données appariées (dépendantes) ?

a) La glycémie des mêmes patients à l'inclusion et à 6 mois

b) La PAS des hommes comparée à celle des femmes

c) L'acuité visuelle de l'œil droit comparée à celle de l'œil gauche, chez les mêmes patients

d) Le résultat (positif ou négatif) de deux tests diagnostiques appliqués aux mêmes patients

e) L'IMC des patients de trois cabinets différents

30. (QRU) Un étudiant compare une variable numérique entre deux groupes indépendants de 25 patients chacun ; le test de Shapiro-Wilk donne p = 0,003 dans l'un des groupes, et l'histogramme montre une forte asymétrie. Le choix le plus approprié est :

a) le test t classique, sans réserve

b) le test de Mann-Whitney

c) le test du khi-deux

d) le coefficient de Pearson

e) l'ANOVA à mesures répétées

31. (QRU) Entre deux variables, le coefficient de corrélation de Pearson est r = 0,6. Le coefficient de détermination est :

a) 0,06

b) 0,36

c) 0,60

d) 0,77

e) 1,20

32. (QRM) Dans mg2026, la droite de régression de la PAS en fonction de l'IMC est PAS = 89,47 + 1,435 x IMC (r = 0,405). Les affirmations correctes sont :

a) Pour un IMC de 30 kg/m², la PAS prédite est d'environ 132,5 mmHg

b) La PAS augmente en moyenne de 1,435 mmHg pour chaque kg/m² supplémentaire

c) 89,47 mmHg est la PAS réelle d'un patient dont l'IMC est égal à 0

d) L'IMC explique environ 16,4% de la variabilité de la PAS

e) La valeur de r démontre que l'obésité provoque l'augmentation de la PAS

33. (QRU) Le coefficient de corrélation de Spearman est préféré au coefficient de Pearson lorsque :

a) les deux variables sont normales et la relation est linéaire

b) la taille de l'échantillon dépasse 1 000

c) les deux variables sont nominales

d) on souhaite une équation de prédiction

e) au moins l'une des variables est ordinale ou la relation est monotone, mais non linéaire

34. (QRM) Dans une étude portant sur 210 patients, on obtient r = 0,15, p = 0,03. Les affirmations correctes sont :

a) La corrélation est statistiquement significative au seuil de 0,05

b) L'intensité de l'association est très faible, bien que le résultat soit significatif

c) La variable x explique environ 2,25% de la variabilité de y

d) Le résultat démontre une relation causale

e) Un coefficient faible peut devenir significatif sur un grand échantillon

35. (QRU) Dans une étude de cohorte, 30 exposés sur 100 et 10 non-exposés sur 100 développent la maladie. Le risque relatif est de :

a) 2,0

b) 3,0

c) 3,86

d) 0,33

e) 20

36. (QRM) Pour les données de la question 35, les affirmations correctes sont :

a) L'OR est d'environ 3,86

b) La différence de risques est de 20 points de pourcentage

c) Le nombre de sujets nécessaire pour produire un cas supplémentaire (NNH) est de 5

d) L'OR est inférieur au RR

e) La fraction attribuable chez les exposés est d'environ 66,7%

37. (QRU) La mesure d'association qui peut être calculée correctement dans une étude cas-témoins est :

a) le rapport de cotes (OR)

b) le risque relatif (RR)

c) l'incidence cumulée

d) la prévalence de la maladie dans la population

e) le nombre de sujets à traiter

38. (QRU) Dans un essai randomisé, l'événement défavorable survient chez 20% des patients du groupe témoin et chez 15% des patients traités. Le nombre de sujets à traiter (NST) est de :

a) 5

b) 15

c) 20

d) 25

e) 33

39. (QRM) À propos du facteur de confusion, les affirmations correctes sont :

a) Il est associé à l'exposition

b) Il est associé au résultat (la maladie)

c) C'est un maillon intermédiaire sur la voie causale entre l'exposition et la maladie

d) Son influence peut être contrôlée par la randomisation

e) Il est éliminé automatiquement par l'augmentation de la taille de l'échantillon

40. (QRU) Une maladie a une prévalence de 2%. Un test a une sensibilité de 95% et une spécificité de 90%. La probabilité qu'une personne dont le test est positif ait la maladie est d'environ :

a) 95%

b) 90%

c) 2%

d) 16%

e) 50%

41. (QRM) La probabilité pré-test d'une maladie est de 10%, et le test a un rapport de vraisemblance positif RV+ = 9. Les affirmations correctes sont :

a) La cote pré-test est d'environ 0,11

b) Après un test positif, la cote post-test est de 1,0

c) La probabilité post-test après un test positif est de 50%

d) La probabilité post-test après un test positif est de 90%

e) Un RV+ de 9 correspond, par exemple, à un test de Se 90% et de Sp 90%

42. (QRU) Dans la formulation d'une question clinique selon le modèle PICO, la lettre O désigne :

a) l'observation

b) le résultat recherché (outcome)

c) l'objectif de l'étude

d) la cote (odds)

e) l'insu

43. (QRM) Un test est appliqué à 50 malades et à 150 non-malades : 45 malades ont un test positif, et 130 non-malades ont un test négatif. Les affirmations correctes sont :

a) La sensibilité est de 90%

b) La spécificité est d'environ 86,7%

c) La valeur prédictive positive est d'environ 69,2%

d) Le taux d'erreur de classement est (5 + 130)/200

e) L'exactitude du test est de 75%

44. (QRU) Pour le test de la question 43, le taux d'erreur de classement est de :

a) 2,5%

b) 10%

c) 13,3%

d) 67,5%

e) 12,5%

45. (QRU) Dans une population de 200 personnes, la prévalence de la maladie est de 30%. Le test a une sensibilité de 80% et une spécificité de 90%. La valeur prédictive positive est d'environ :

a) 30,0%

b) 77,4%

c) 80,0%

d) 90,0%

e) 91,3%

46. (QRM) À propos de la courbe ROC, les affirmations correctes sont :

a) L'axe vertical représente la sensibilité

b) L'axe horizontal représente 1 - la spécificité

c) Une aire sous la courbe de 0,5 correspond à un test sans valeur discriminante

d) Une aire de 0,852 signifie que 85,2% des patients sont correctement classés quel que soit le seuil

e) Le seuil qui maximise l'indice de Youden est toujours celui dont la sensibilité est maximale

47. (QRU) Un test a une sensibilité de 85% et une spécificité de 75%. L'indice de Youden est de :

a) 0,10

b) 0,40

c) 0,60

d) 0,64

e) 1,60

48. (QRU) Dans un essai contrôlé randomisé de supériorité, l'analyse principale recommandée est :

a) l'analyse en intention de traiter

b) l'analyse per protocole

c) l'analyse limitée aux patients qui ont terminé l'étude

d) l'analyse d'un sous-groupe choisi après avoir vu les données

e) l'analyse intermédiaire

49. (QRM) À propos des essais contrôlés randomisés, les affirmations correctes sont :

a) La randomisation tend à équilibrer entre les bras même les facteurs de confusion inconnus

b) La dissimulation de l'allocation et l'insu sont la même chose

c) Dans une étude en double insu, ni le patient ni l'investigateur ne connaissent le traitement reçu

d) Le critère de jugement principal peut être choisi après l'analyse des données

e) La recommandation CONSORT exige un diagramme de flux des participants

50. (QRU) Un article rapporte, pour une étude transversale portant sur 80 patients, "OR = 2,1 ; p = 0,04", sans autres détails. L'information manquante la plus importante pour l'interprétation est :

a) le nom du logiciel statistique

b) la valeur exacte du seuil α

c) l'âge moyen des patients

d) l'intervalle de confiance de l'OR

e) le nombre d'auteurs

Test 2

1. (QRU) La pression artérielle systolique moyenne de l'ensemble des adultes de Roumanie, valeur inconnue que nous cherchons à estimer, est :

a) un paramètre

b) une statistique d'échantillon

c) un échantillon

d) une variable qualitative

e) une erreur standard

2. (QRM) À propos de la protection des données et de l'éthique de la recherche, les affirmations correctes sont :

a) Les données pseudonymisées restent des données à caractère personnel

b) Le consentement éclairé peut être retiré à tout moment par le participant

c) L'anonymisation consiste à remplacer le nom par un code, en conservant une table de correspondance

d) Les données de santé font partie des catégories particulières de données protégées par le RGPD

e) Un jeu de données simulé, comportant des patients fictifs, relève du RGPD comme tout jeu de données clinique

3. (QRU) Les événements A et B sont indépendants, avec P(A) = 0,3 et P(B) = 0,4. La probabilité que les deux se produisent est de :

a) 0,70

b) 0,12

c) 0,10

d) 0,58

e) 0,75

4. (QRU) Le nombre de consultations par an chez six patients est 3, 5, 8, 10, 12, 20. La médiane est de :

a) 8

b) 9,67

c) 9

d) 10

e) 11,5

5. (QRM) Un questionnaire a été rempli par 200 personnes. À la question "Fumez-vous ?", 20 personnes n'ont pas répondu et 54 ont répondu "oui". Les affirmations correctes sont :

a) Le pourcentage valide de fumeurs est de 30%

b) Le pourcentage de fumeurs rapporté au total des répondants est de 27%

c) Le pourcentage valide se calcule en rapportant à 200

d) Le rapport doit préciser le nombre de valeurs manquantes

e) Les réponses manquantes sont remplacées par "non", pour ne pas perdre de données

6. (QRU) Pour une variable, Q1 = 20 et Q3 = 30. D'après la règle de Tukey utilisée pour le box-plot, la limite supérieure au-delà de laquelle une valeur est considérée comme aberrante est :

a) 35

b) 40

c) 50

d) 45

e) 60

7. (QRU) Le graphique approprié pour montrer la forme de la distribution de la glycémie à jeun chez 160 patients est :

a) le diagramme circulaire

b) le diagramme en barres

c) le diagramme de dispersion

d) le forest plot

e) l'histogramme

8. (QRU) Chez les femmes en bonne santé d'une population, l'hémoglobine suit une loi normale de moyenne 14 g/dL et d'écart-type 1,2 g/dL. Le pourcentage approximatif des femmes dont l'hémoglobine est inférieure à 11,6 g/dL est de :

a) 0,13%

b) 2,3%

c) 5%

d) 15,9%

e) 47,7%

9. (QRM) Une variable suit une loi normale de moyenne 120 et d'écart-type 10. Les affirmations correctes sont :

a) Environ 95% des individus ont des valeurs comprises entre 100,4 et 139,6

b) Environ 68,3% des individus ont des valeurs comprises entre 110 et 130

c) Environ 99,7% des individus ont des valeurs comprises entre 90 et 150

d) L'intervalle 100,4-139,6 est l'intervalle de confiance à 95% de la moyenne de la population

e) Environ 2,5% des individus ont des valeurs inférieures à 100,4

10. (QRU) Après l'analyse d'une variable sur l'échelle logarithmique, la moyenne des logarithmes ramenée sur l'échelle initiale par exponentiation est :

a) la moyenne arithmétique

b) la médiane

c) la moyenne géométrique

d) le mode

e) l'écart-type

11. (QRU) Le théorème central limite affirme que :

a) les moyennes d'échantillons suffisamment grands suivent une loi approximativement normale, quelle que soit la distribution de la variable dans la population

b) toutes les variables biologiques suivent une loi normale

c) la médiane et la moyenne sont toujours égales dans les grands échantillons

d) l'écart-type diminue lorsque la taille de l'échantillon augmente

e) tout test devient significatif si l'échantillon dépasse 30 personnes

12. (QRU) Sur 16 patients, la moyenne d'une variable est de 50 et l'écart-type de 8. Sachant que t(0,05 ; 15) = 2,131, l'intervalle de confiance à 95% de la moyenne de la population est :

a) 46,1-53,9

b) 45,7-54,3

c) 34,0-66,0

d) 48,0-52,0

e) 42,0-58,0

13. (QRU) Quelle taille d'échantillon faut-il pour estimer la moyenne d'une variable avec $\sigma = 10$, une marge d'erreur de ±2 unités et un niveau de confiance de 95% ?

a) 10

b) 49

c) 96

d) 97

e) 385

14. (QRM) Dans mg2026, l'intervalle de confiance à 95% de la PAS moyenne est 128,96-133,30 mmHg. Les affirmations correctes sont :

a) Si l'étude était répétée de nombreuses fois, environ 95% des intervalles construits de la même façon contiendraient la moyenne réelle

b) 95% des patients ont une PAS comprise entre 128,96 et 133,30 mmHg

c) L'intervalle de confiance à 99%, calculé sur les mêmes données, est plus large

d) Avec un échantillon 4 fois plus grand, l'intervalle serait environ deux fois plus étroit

e) L'intervalle est centré sur la moyenne de l'échantillon

15. (QRU) Dans une étude pilote, 2 patients sur 25 présentent une réaction indésirable. L'intervalle de confiance de Wald de la proportion a une limite inférieure négative. L'attitude correcte est :

a) on rapporte la limite négative telle qu'elle a été obtenue

b) on remplace la limite négative par 0 et on conserve la limite supérieure de Wald

c) on renonce à l'intervalle et on ne rapporte que la proportion

d) on augmente artificiellement le nombre de cas

e) on utilise l'intervalle de Wilson, adapté aux petits échantillons et aux proportions extrêmes

16. (QRM) À propos des hypothèses statistiques, les affirmations correctes sont :

a) Pour la comparaison de deux moyennes, H0 est $\mu 1 = \mu 2$

b) L'hypothèse alternative bilatérale est $\mu 1 \neq \mu 2$

c) Lorsque $p > 0,05$, l'hypothèse nulle est acceptée comme démontrée

d) Pour un risque relatif, H0 est $RR = 1$

e) Pour un coefficient de corrélation, H0 est $\rho = 1$

17. (QRU) On effectue 5 comparaisons sur les mêmes données et on applique la correction de Bonferroni pour un α global de 0,05. Le seuil utilisé pour chaque comparaison est :

a) 0,05

b) 0,025

c) 0,01

d) 0,005

e) 0,25

18. (QRU) Dans une étude portant sur 20 000 patients, un médicament diminue la PAS de 1,2 mmHg par rapport au placebo (IC 95% 0,8-1,6 mmHg), $p < 0,001$. L'interprétation la plus appropriée est :

a) l'effet est statistiquement significatif, mais probablement sans importance clinique

b) l'effet est cliniquement important, car p est très petit

c) le résultat n'est pas statistiquement significatif

d) l'intervalle de confiance contient 0

e) l'étude a eu une puissance insuffisante

19. (QRM) La puissance d'un test statistique augmente lorsque :

a) la taille de l'échantillon augmente

b) la taille de l'effet recherché augmente

c) la variabilité (l'écart-type) de la variable augmente

d) le seuil α passe de 0,01 à 0,05

e) un schéma apparié est remplacé par un schéma à groupes indépendants, pour le même nombre de mesures

20. (QRU) On compare, avec le test t classique pour échantillons indépendants, deux groupes de 25 et 30 patients. Le nombre de degrés de liberté est :

a) 55

b) 54

c) 53

d) 29

e) 24

21. (QRM) Dans mg2026, la glycémie à 0 et à 6 mois est comparée par le test t apparié : moyenne des différences 10,19 mg/dL, écart-type des différences 11,57, n = 160, t = 11,14, IC 95% 8,38-11,99. Les affirmations correctes sont :

a) L'erreur standard de la moyenne des différences est d'environ 0,915

b) Le nombre de degrés de liberté est 159

c) L'intervalle de confiance ne contient pas 0, en accord avec $p < 0,05$

d) Le test suppose l'égalité des variances de la glycémie aux deux temps de mesure

e) Une taille d'effet dz = 0,88 indique un effet fort

22. (QRU) Deux groupes indépendants à données normales ont des variances différentes (test de Levene p = 0,002). Le test recommandé pour comparer les moyennes est :

a) le test t apparié

b) le test du khi-deux

c) le test de Friedman

d) le test t de Welch

e) le test de McNemar

23. (QRU) Dans une ANOVA à un facteur, la somme des carrés intergroupes est de 300 et la somme des carrés totale de 1 200. L'êta carré est de :

a) 0,25

b) 0,33

c) 0,75

d) 4,00

e) 0,0025

24. (QRU) Deux groupes ont des moyennes de 136 et 130 mmHg, et l'écart-type commun (pooled) est de 12 mmHg. La taille d'effet d de Cohen et son interprétation sont :

a) 2,0, effet fort

b) 0,5, effet faible

c) 72, effet très fort

d) 0,05, effet négligeable

e) 0,5, effet moyen

25. (QRM) Dans un tableau 2 x 2, a = 20, b = 30, c = 10, d = 40 (N = 100). Sachant que les valeurs critiques de χ^2 pour ddl = 1 sont 3,841 ($\alpha = 0,05$) et 6,635 ($\alpha = 0,01$), les affirmations correctes sont :

a) χ^2 est d'environ 4,76

b) Le nombre de degrés de liberté est 1

c) $p < 0,05$

d) $p < 0,01$

e) La fréquence attendue dans la case a est de 15

26. (QRU) Deux tests de dépistage (A et B) sont appliqués aux mêmes 100 patients, et le résultat de chacun est positif ou négatif. Pour comparer les proportions de résultats positifs, on utilise :

a) le test du khi-deux d'indépendance

b) le test de McNemar

c) le test de Mann-Whitney

d) le test t pour échantillons indépendants

e) le test de Kruskal-Wallis

27. (QRU) Le groupe A a les valeurs 12, 15, 18, et le groupe B les valeurs 14, 20, 22, 25. La statistique U du test de Mann-Whitney (la plus petite de U1 et U2) est :

a) 2

b) 8

c) 10

d) 12

e) 20

28. (QRM) À propos des tests de Kruskal-Wallis et de Friedman, les affirmations correctes sont :

a) Le test de Kruskal-Wallis est l'alternative non paramétrique de l'ANOVA à un facteur

b) Le test de Friedman s'utilise pour des mesures répétées chez les mêmes individus

c) Un test de Kruskal-Wallis significatif montre exactement quels groupes diffèrent entre eux

d) Le test de Kruskal-Wallis repose sur les rangs

e) Le test de Friedman s'applique à des groupes indépendants

29. (QRU) Pour étudier l'association entre le tabagisme (oui/non) et la BPCO (oui/non) chez 300 patients, avec toutes les fréquences attendues supérieures à 5, on utilise :

a) le coefficient de Pearson

b) le test t pour échantillons indépendants

c) l'ANOVA

d) le test du khi-deux

e) le test de Wilcoxon

30. (QRM) Les associations correctes entre situation et test sont :

a) trois temps de mesure, mêmes patients, données normales : ANOVA à mesures répétées

b) deux groupes indépendants, variable ordinale : test de Mann-Whitney

c) deux variables nominales : test du khi-deux

d) deux variables numériques normales, relation linéaire : le coefficient de Spearman est obligatoire

e) trois groupes indépendants, données normales : test de Friedman

31. (QRU) Chez 6 patients, la somme des carrés des différences de rang entre deux variables est $\sum d^2 = 14$. Le coefficient de Spearman est :

a) 0,4

b) 0,6

c) 0,8

d) 0,93

e) -0,6

32. (QRM) Sur 27 patients, on obtient r = 0,5 entre deux variables. La valeur critique de t pour ddl = 25 et α = 0,05 en bilatéral est 2,060. Les affirmations correctes sont :

a) Le nombre de degrés de liberté est 25

b) La statistique t est d'environ 2,89

c) La corrélation est significative au seuil de 0,05

d) Le coefficient de détermination est 0,25

e) La corrélation est parfaite

33. (QRU) Dans mg2026, PAS = 108,3 + 0,406 x âge. La PAS prédite pour un patient de 60 ans est d'environ :

a) 132,7 mmHg

b) 24,4 mmHg

c) 108,7 mmHg

d) 168,3 mmHg

e) 141,8 mmHg

34. (QRU) Dans la régression multiple PAS ~ IMC + âge + sexe + tabagisme de mg2026, le coefficient de l'IMC est de 1,389 (p < 0,001). L'interprétation correcte est :

a) l'IMC explique 1,389% de la variabilité de la PAS

b) la corrélation entre l'IMC et la PAS est de 1,389

c) la PAS augmente en moyenne de 1,389 mmHg pour chaque kg/m² supplémentaire, à âge, sexe et statut tabagique constants

d) un patient dont l'IMC est de 0 a une PAS de 1,389 mmHg

e) l'IMC augmente de 1,389 kg/m² pour chaque mmHg de PAS

35. (QRU) Dans une étude cas-témoins, 40 cas sur 100 et 20 témoins sur 100 ont été exposés. Le rapport de cotes est d'environ :

a) 2,0

b) 1,5

c) 0,375

d) 20

e) 2,67

36. (QRM) Une étude rapporte un RR = 1,8 (IC 95% 0,9-3,6). Les affirmations correctes sont :

a) L'association n'est pas statistiquement significative au seuil de 0,05

b) Les données sont compatibles avec une absence d'effet

c) Les données sont compatibles aussi avec un doublement du risque

d) Il a été démontré que l'exposition n'a aucun effet

e) La valeur p est inférieure à 0,05

37. (QRU) Dans une cohorte de 2 000 personnes sans diabète suivies pendant 2 ans, 50 nouveaux cas de diabète apparaissent. L'incidence cumulée est de :

a) 2,5% en 2 ans

b) 1,25% en 2 ans

c) 25% en 2 ans

d) 50 pour 1 000 personnes-années

e) 2,5 pour 1 000 personnes-années

38. (QRU) On choisit 120 patients atteints d'un cancer du poumon et 240 personnes sans cancer et l'on compare la fréquence du tabagisme dans le passé. Le type d'étude est :

a) étude de cohorte prospective

b) essai contrôlé randomisé

c) étude écologique

d) étude cas-témoins

e) série de cas

39. (QRM) Dans mg2026 (analyse transversale à l'inclusion), 18 obèses sur 54 et 11 non-obèses sur 106 sont diabétiques. Les affirmations correctes sont :

a) La prévalence du diabète chez les obèses est de 33,3%

b) Le rapport de prévalences est d'environ 3,21

c) L'OR est d'environ 4,32

d) L'étude démontre que l'obésité a précédé l'apparition du diabète

e) L'OR est supérieur au rapport de prévalences parce que la maladie n'est pas rare dans cet échantillon

40. (QRU) Un test a un rapport de vraisemblance positif RV+ = 8. La probabilité pré-test de la maladie est de 20%. La probabilité post-test après un résultat positif est d'environ :

a) 20%

b) 66,7%

c) 80%

d) 88,9%

e) 160%

41. (QRM) À propos de la médecine fondée sur les preuves, les affirmations correctes sont :

a) Les revues systématiques avec méta-analyse d'essais randomisés de qualité se trouvent au sommet de la hiérarchie des preuves

b) Une série de cas peut générer des hypothèses, mais ne peut pas les tester

c) L'opinion d'expert est la forme de preuve la plus solide

d) La médecine fondée sur les preuves intègre les preuves issues de la recherche à l'expérience clinique et aux valeurs du patient

e) Une grande étude observationnelle est toujours supérieure à un essai randomisé

42. (QRU) Le même test diagnostique (sensibilité et spécificité inchangées) est utilisé dans une population où la prévalence de la maladie est plus faible. Qu'advient-il des valeurs prédictives ?

a) la VPP augmente, la VPN diminue

b) les deux restent inchangées

c) la VPP diminue, la VPN augmente

d) les deux diminuent

e) les deux augmentent

43. (QRM) On sait d'un test diagnostique qu'il a donné 40 résultats vrais positifs, VPP = 0,8, 10 résultats faux négatifs et VPN = 0,9. Les affirmations correctes sont :

a) Le nombre de résultats faux positifs est 10

b) Le nombre de résultats vrais négatifs est 90

c) La sensibilité est de 80%

d) La spécificité est de 90%

e) Le nombre total de personnes testées est 200

44. (QRU) Deux radiologues classent les mêmes radiographies ; l'accord observé est de 0,85, et l'accord attendu par hasard de 0,50. Le coefficient kappa est de :

a) 0,35

b) 0,50

c) 0,70

d) 0,85

e) 1,70

45. (QRU) Un test à très haute sensibilité donne un résultat négatif chez un patient. L'interprétation correcte (règle SnNOut) est :

a) la maladie est pratiquement exclue

b) la maladie est confirmée

c) le résultat n'apporte aucune information

d) il faut calculer la spécificité pour décider

e) le patient a la maladie à un stade précoce

46. (QRU) Pour le score FINDRISC de mg2026, le seuil ≥ 12 a une Se de 75,9% et une Sp de 74,8%, et le seuil ≥ 16 une Se de 41,4% et une Sp de 94,7%. Relever le seuil de 12 à 16 a pour effet :

a) l'augmentation de la sensibilité et de la spécificité

b) la diminution des deux

c) l'augmentation de la sensibilité et la diminution de la spécificité

d) la diminution de la sensibilité et l'augmentation de la spécificité

e) la modification de l'aire sous la courbe ROC

47. (QRM) À propos des tests de dépistage et de la combinaison de tests, les affirmations correctes sont :

a) Pour un test de dépistage, on préfère une sensibilité élevée

b) Les résultats positifs au dépistage doivent être confirmés par un test spécifique

c) Dans la population générale, la VPP d'un test est en règle générale plus faible que dans une consultation spécialisée

d) L'utilisation de tests en série augmente la sensibilité

e) L'utilisation de tests en parallèle augmente la sensibilité

48. (QRU) Le calcul montre qu'il faut 64 patients par groupe. On estime le taux d'abandon à 20%. Combien de patients faut-il recruter par groupe ?

a) 51

b) 77

c) 80

d) 84

e) 128

49. (QRM) À propos du protocole de l'étude et des formulaires de collecte des données, les affirmations correctes sont :

a) Le protocole est rédigé et approuvé avant la collecte des données

b) Les règles de validation des réponses dans un formulaire électronique réduisent les erreurs de saisie

c) Un formulaire bien conçu demande la PAS sous forme de texte libre, avec l'unité de mesure écrite par l'utilisateur

d) Le plan d'analyse statistique est fixé avant l'analyse des données

e) Les essais cliniques sont enregistrés dans les registres publics après la publication des résultats

50. (QRU) JASP affiche pour un test "p <.001", et un étudiant écrit dans son exposé "p = 0,000". La formulation correcte dans le rapport est :

a) $p = 0$

b) $p = 0{,}000$

c) p est nul, donc H0 est fausse avec certitude

d) $p = 0{,}001$

e) p < 0,001

Test 3

1. (QRU) La première étape d'une étude statistique est :

a) l'analyse statistique des données

b) la collecte des données

c) la formulation de la question de recherche et la planification de l'étude

d) le nettoyage du fichier de données

e) le choix du logiciel statistique

2. (QRM) À propos des erreurs en recherche, les affirmations correctes sont :

a) L'erreur aléatoire diminue lorsque la taille de l'échantillon augmente

b) L'erreur systématique (le biais) diminue lorsque la taille de l'échantillon augmente

c) Le biais de mémorisation est caractéristique des études cas-témoins

d) Un échantillon composé de volontaires peut introduire un biais de sélection

e) L'intervalle de confiance quantifie l'ampleur du biais

3. (QRM) Pour deux événements, P(A) = 0,3, P(B) = 0,4 et P(A et B) = 0,12. Les affirmations correctes sont :

a) A et B sont indépendants

b) $P(A \text{ ou } B) = 0{,}58$

c) $P(A \mid B) = 0{,}3$

d) A et B sont incompatibles

e) $P(\text{non} - A) = 0{,}7$

4. (QRU) Les titres d'anticorps de trois patients sont 10, 100 et 1 000. La moyenne géométrique est :

a) 370

b) 100

c) 1000

d) 30

e) 10

5. (QRU) La variable A a une moyenne de 50 et un écart-type de 5 ; la variable B a une moyenne de 200 et un écart-type de 30. Quelle affirmation est correcte ?

a) A a la variabilité relative la plus grande, parce que sa moyenne est plus petite

b) Les deux variables ont la même variabilité relative

c) La variabilité relative ne peut pas être comparée

d) B a la variabilité absolue la plus petite

e) B a la variabilité relative la plus grande (CV 15% contre 10%)

6. (QRM) Les associations correctes entre la variable et son type sont :

a) La glycémie (mg/dL) : numérique continue

b) Le nombre de grossesses : numérique discrète

c) Le groupe sanguin : ordinale

d) Le score de satisfaction 1-5 : ordinale

e) La température en degrés Celsius : échelle de rapport

7. (QRU) L'indicateur de tendance centrale le moins influencé par les valeurs extrêmes est :

a) la médiane

b) la moyenne arithmétique

c) la moyenne pondérée

d) l'étendue

e) l'écart-type

8. (QRU) Sur le graphique Q-Q d'une variable, les points se disposent approximativement sur la droite. Cela suggère que :

a) la variable est ordinale

b) il existe de nombreuses valeurs aberrantes

c) la distribution est fortement asymétrique

d) la distribution est proche de la loi normale

e) les variables sont corrélées

9. (QRM) Dans mg2026, l'IMC a une moyenne de 29,04, une médiane de 28,25, une asymétrie de 1,42 et un test de Shapiro-Wilk à $p < 0,001$. Les affirmations correctes sont :

a) La distribution est asymétrique à droite

b) La moyenne dépasse la médiane, en accord avec l'asymétrie positive

c) Le test de Shapiro-Wilk indique un écart à la normalité

d) L'intervalle $m \pm 1,96s$ contient exactement 95% des patients

e) Pour comparer l'IMC entre des groupes, on peut préférer les tests non paramétriques

10. (QRU) La valeur du score z qui laisse à sa droite 2,5% de l'aire sous la courbe normale centrée réduite est :

a) 1,645

b) 1,96

c) 2,33

d) 2,576

e) 3,00

11. (QRU) Le poids d'un enfant se situe au 90e percentile pour l'âge et le sexe. Cela signifie que :

a) l'enfant pèse 90% de plus que le poids moyen

b) l'enfant est obèse

c) 10% des enfants ont un poids inférieur au sien

d) son poids est 90 fois supérieur à la moyenne

e) environ 90% des enfants de même âge et de même sexe ont un poids inférieur

12. (QRU) Dans un échantillon de 60 patients, 18 ont une hypertension non contrôlée. L'intervalle de confiance à 95% de la proportion (méthode de Wald) est approximativement :

a) 18,4-41,6%

b) 24,1-35,9%

c) 28,0-32,0%

d) 10,0-50,0%

e) 0,18-0,42%

13. (QRM) La taille d'échantillon nécessaire pour estimer un paramètre augmente lorsque :

a) la marge d'erreur souhaitée diminue

b) le niveau de confiance passe de 95% à 99%

c) la variabilité de la variable (σ) est plus grande

d) la proportion estimée se rapproche de 0,5

e) la population dans laquelle on tire l'échantillon passe de 1 million à 2 millions de personnes

14. (QRU) La taille de l'échantillon passe de 40 à 160 patients, et l'écart-type reste approximativement le même. La largeur de l'intervalle de confiance de la moyenne devient approximativement :

a) quatre fois plus petite

b) inchangée

c) deux fois plus petite

d) deux fois plus grande

e) quatre fois plus grande

15. (QRU) Combien de patients faut-il pour estimer une prévalence supposée de 10%, avec une marge d'erreur de ±3% et un niveau de confiance de 95% ?

a) 100

b) 196

c) 400

d) 385

e) 1000

16. (QRU) Sur 64 patients, la moyenne d'une variable est de 128, l'écart-type de 16, et la valeur de référence est de 120. La statistique t du test t pour un échantillon est :

a) 0,5

b) 2

c) 4

d) 8

e) 32

17. (QRM) À propos de la valeur p, les affirmations correctes sont :

a) p = 0,03 signifie que la probabilité que H0 soit vraie est de 3%

b) Une valeur p très petite peut aussi apparaître pour un effet clinique négligeable, si l'échantillon est grand

c) Un p > 0,05 ne prouve pas l'absence d'effet

d) Pour un même effet, la valeur p dépend de la taille de l'échantillon

e) Les valeurs p = 0,049 et p = 0,051 représentent des preuves radicalement différentes

18. (QRU) Un test unilatéral est justifié lorsque :

a) le sens de l'effet a été fixé avant l'étude, et un effet dans le sens opposé est impossible ou sans intérêt

b) le test bilatéral a donné p = 0,08

c) le chercheur souhaite un résultat significatif

d) l'échantillon est petit

e) les données ne sont pas normales

19. (QRU) On effectue 20 tests indépendants, et pour tous H0 est vraie. À α = 0,05, la probabilité d'obtenir au moins un résultat faussement significatif est d'environ :

a) 0,05

b) 0,20

c) 0,36

d) 0,64

e) 1,00

20. (QRU) Dans une ANOVA avec trois groupes et 60 patients, le carré moyen intergroupes est de 150 et le carré moyen résiduel de 30. La statistique F est :

a) 0,2

b) 5

c) 4,5

d) 120

e) 180

21. (QRM) Pour l'ANOVA de la question 20, sachant que le F critique pour ddl (2 ; 57) et α = 0,05 est d'environ 3,16, les affirmations correctes sont :

a) Le nombre de degrés de liberté intergroupes est 2

b) Le nombre de degrés de liberté résiduels est 57

c) p < 0,05

d) Il a été démontré que les trois moyennes diffèrent toutes entre elles

e) L'étape suivante est un test post-hoc, par exemple celui de Tukey

22. (QRU) Le test de Tukey, appliqué après une ANOVA significative, a pour rôle de :

a) vérifier la normalité des données

b) vérifier l'égalité des variances

c) remplacer l'ANOVA lorsque les données sont ordinales

d) calculer la taille d'effet globale

e) identifier les paires de groupes qui diffèrent, en contrôlant l'erreur de type I

23. (QRU) Le poids de 30 patients est mesuré au début et à la fin d'un programme d'amaigrissement ; les différences suivent une loi normale. Le test approprié est :

a) le test t apparié

b) le test t pour échantillons indépendants

c) le test de Mann-Whitney

d) le test du khi-deux

e) l'ANOVA à un facteur

24. (QRM) Les conditions du test t classique (de Student) pour échantillons indépendants sont :

a) observations indépendantes

b) distribution approximativement normale dans chaque groupe (ou grands échantillons)

c) variances égales dans les deux groupes

d) effectifs égaux des deux groupes

e) variable de résultat ordinale

25. (QRU) Pour un tableau de contingence à 2 lignes et 3 colonnes, le nombre de degrés de liberté du test du khi-deux est :

a) 1

b) 2

c) 3

d) 5

e) 6

26. (QRU) Dans un test de Mann-Whitney, le groupe 1 compte 4 patients, le groupe 2 compte 5 patients, et la somme des rangs du groupe 1 est 12. La valeur $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$ est :

a) 2

b) 12

c) 18

d) 20

e) 10

27. (QRM) Dans mg2026, une perte de poids d'au moins 5% est survenue chez 6/53 patients dans le bras A, 16/53 dans le bras B et 16/54 dans le bras C ; χ^2 = 6,77, ddl = 2, p = 0,034, et la plus petite fréquence attendue est de 12,6. Les affirmations correctes sont :

a) La condition relative aux fréquences attendues est remplie

b) La proportion de patients ayant perdu au moins 5% de leur poids diffère significativement entre les bras

c) Le test montre que le bras C diffère significativement du bras B

d) Les degrés de liberté se calculent comme (3 - 1) x (2 - 1)

e) Le test exact de Fisher était obligatoire

28. (QRU) La glycémie a été mesurée chez les mêmes patients à 0, 3 et 6 mois, et la distribution est fortement asymétrique sur un petit échantillon. Le test approprié est :

a) l'ANOVA à un facteur

b) le test de Kruskal-Wallis

c) le test du khi-deux

d) le test de Friedman

e) le test de Mann-Whitney

29. (QRU) L'HbA1c (distribution normale dans les deux groupes) est comparée entre les patients avec et sans diabète confirmé ; le test de Levene donne p = 0,40. Le test approprié est :

a) le test t apparié

b) le test de Wilcoxon

c) le test t pour échantillons indépendants (version classique ou de Welch)

d) le test du khi-deux

e) le test de Friedman

30. (QRM) Le stade de la fibrose hépatique (F0-F4) est une variable ordinale. Les méthodes appropriées sont :

a) le test de Mann-Whitney pour comparer deux groupes

b) le test de Kruskal-Wallis pour comparer trois groupes

c) le coefficient de Spearman pour l'association avec une variable numérique

d) le test t pour échantillons indépendants appliqué aux codes 0-4, comme s'il s'agissait d'une variable continue normale

e) le test du khi-deux pour le tableau stade x groupe, en acceptant la perte de l'information sur l'ordre

31. (QRU) Dans une régression linéaire simple, $R^2 = 0{,}25$, et la pente de la droite est négative. Le coefficient de corrélation de Pearson est :

a) 0,25

b) -0,25

c) 0,5

d) -0,5

e) 0,0625

32. (QRM) À propos de la régression linéaire, les affirmations correctes sont :

a) Le résidu est la différence entre la valeur observée et la valeur prédite

b) La méthode des moindres carrés minimise la somme des carrés des résidus

c) L'erreur standard de l'estimation s'exprime dans les unités de la variable dépendante

d) La prédiction pour des valeurs de x très en dehors du domaine observé est tout aussi sûre

e) Un R^2 élevé prouve une relation causale

33. (QRU) Dans une régression logistique, le coefficient du tabagisme est b = 0,693. Le rapport de cotes correspondant est d'environ :

a) 0,693

b) 2,0

c) 1,693

d) 0,5

e) 6,93

34. (QRU) Sur un diagramme de dispersion, un seul patient a des valeurs très élevées sur les deux axes, loin des autres. Ce point :

a) ne peut pas influencer le coefficient de corrélation

b) diminue toujours le coefficient de corrélation

c) transforme automatiquement la corrélation en une corrélation non linéaire

d) doit être supprimé sans vérification

e) peut modifier fortement le coefficient de corrélation et doit être vérifié

35. (QRU) Dans un essai randomisé, l'événement survient chez 8 patients traités sur 200 et chez 16 patients sous placebo sur 200. Le risque relatif est de :

a) 0,5

b) 2,0

c) 0,04

d) 0,08

e) 0,96

36. (QRM) Pour les données de la question 35, les affirmations correctes sont :

a) La réduction absolue du risque est de 4 points de pourcentage

b) La réduction relative du risque est de 50%

c) Le NST est de 25

d) L'OR est exactement égal au RR

e) Ne rapporter que la réduction relative du risque peut exagérer l'impression de bénéfice

37. (QRU) Une exposition a un RR = 3, et 25% de la population est exposée. La fraction attribuable dans la population (formule de Levin) est d'environ :

a) 0,25

b) 0,333

c) 0,50

d) 0,667

e) 0,75

38. (QRU) Une maladie chronique a une incidence faible, mais les malades survivent de très nombreuses années. Dans cette situation :

a) la prévalence est toujours égale à l'incidence

b) la prévalence est inférieure à l'incidence annuelle

c) la prévalence est relativement élevée par rapport à l'incidence annuelle

d) l'incidence ne peut pas être calculée

e) la létalité est élevée

39. (QRM) À propos des types d'études épidémiologiques, les affirmations correctes sont :

a) L'étude de cohorte permet de calculer l'incidence

b) L'étude cas-témoins est efficace pour les maladies rares

c) L'étude cas-témoins permet de calculer directement le risque relatif

d) Une étude de cohorte prospective peut être coûteuse et de longue durée

e) L'étude transversale établit l'ordre chronologique de l'exposition et de la maladie

40. (QRU) Un test a une sensibilité de 90% et une spécificité de 95%. Dans une population où la prévalence est de 1%, la VPP est d'environ :

a) 15,4%

b) 90%

c) 95%

d) 1%

e) 50%

41. (QRM) À propos des rapports de vraisemblance, les affirmations correctes sont :

a) $\mathrm{RV}^{-} = \frac{1-\mathrm{Se}}{\mathrm{Sp}}$

b) Pour Se = 80% et Sp = 90%, le RV- est d'environ 0,22

c) Un RV égal à 1 ne modifie pas la probabilité de la maladie

d) Le nomogramme de Fagan relie la probabilité pré-test, le rapport de vraisemblance et la probabilité post-test

e) Le RV+ dépend de la prévalence de la maladie dans la population testée

42. (QRU) La cote (odds) d'un événement est de 3. La probabilité de l'événement est de :

a) 0,25

b) 0,33

c) 0,75

d) 3,00

e) 0,67

43. (QRU) Un test diagnostique donne : a = 80 vrais positifs, b = 20 faux négatifs, c = 10 faux positifs, d = 90 vrais négatifs. Le rapport de cotes diagnostique (DOR) est de :

a) 4

b) 8

c) 9

d) 36

e) 72

44. (QRM) Pour le test de la question 43, les affirmations correctes sont :

a) La sensibilité est de 80%

b) La spécificité est de 90%

c) Le taux d'erreur de classement est de (20 + 10)/200 = 15%

d) Le taux d'erreur de classement est de (20 + 90)/200

e) L'exactitude est de 85%

45. (QRU) Dans le scénario du jeu de données mg2026, le standard de référence pour le diagnostic du diabète, par rapport auquel on évalue le score FINDRISC, est :

a) la glycémie à jeun

b) l'HbA1c

c) le score FINDRISC lui-même

d) la déclaration du patient

e) l'hyperglycémie provoquée par voie orale

46. (QRM) Dans mg2026, l'aire sous la courbe ROC du score FINDRISC est de 0,852 (IC 95% 0,762-0,943), et celle de la glycémie à jeun de 0,844. Les affirmations correctes sont :

a) Le FINDRISC a une très bonne capacité de discrimination

b) L'IC de l'aire ne contient pas 0,5, donc le score discrimine mieux que le hasard

c) La différence entre 0,852 et 0,844 démontre la supériorité du FINDRISC

d) L'aire sous la courbe ne dépend pas du seuil choisi

e) L'aire est la probabilité qu'un malade choisi au hasard ait un score plus élevé qu'un non-malade choisi au hasard

47. (QRU) Dans une étude d'exactitude, la biopsie (le standard de référence) n'est réalisée que chez les patients dont le test est positif. L'erreur introduite s'appelle :

a) biais de mémorisation

b) biais de vérification

c) erreur de type II

d) facteur de confusion

e) biais de publication

48. (QRU) La phase d'un essai clinique dans laquelle le médicament est comparé de façon randomisée, sur de grands effectifs, au traitement standard ou à un placebo, avant l'autorisation de mise sur le marché, est :

a) la phase I

b) la phase II

c) la phase IV

d) la phase III

e) la phase préclinique

49. (QRM) À propos de la randomisation et de l'insu, les affirmations correctes sont :

a) La randomisation par blocs maintient les bras numériquement équilibrés tout au long du recrutement

b) La dissimulation de l'allocation empêche de connaître le bras avant l'inclusion du patient

c) Une étude ouverte (sans insu) ne peut pas être randomisée

d) L'utilisation d'un placebo permet l'insu des participants

e) L'analyse per protocole préserve toujours les avantages de la randomisation

50. (QRU) Un chercheur veut savoir si la proportion de patients ayant perdu au moins 5% de leur poids diffère entre les trois bras, puis quantifier l'effet du bras C par rapport au bras A. La succession correcte des analyses est :

a) ANOVA, puis test de Tukey

b) test t apparié, puis corrélation de Pearson

c) test du khi-deux sur le tableau 3 x 2, puis RR, différence de risques et NST avec IC 95% pour C par rapport à A

d) régression linéaire simple, puis R^2

e) test de Kruskal-Wallis, puis test de Friedman

Réponses et explications

Vérifiez d'abord vos réponses avec la grille, sans lire les explications, et notez votre score. Lisez ensuite les explications de toutes les questions auxquelles vous avez mal répondu et de celles pour lesquelles vous avez répondu au hasard. Tous les calculs ci-dessous ont été vérifiés avec un logiciel (Python, bibliothèque scipy).

Test 1 : grille des réponses

1. d
2. a, c, e
3. b
4. e
5. a, b, c
6. c
7. b, c, d
8. a
9. a, b, c
10. b
11. c
12. d
13. a, b, c
14. b
15. d
16. a

17. a, b, d
18. c
19. a, c, e
20. e
21. a, b, d, e
22. c
23. b, c, d
24. b
25. a
26. c
27. a, c, d
28. d
29. a, c, d
30. b
31. b
32. a, b, d
33. e
34. a, b, c, e
35. b
36. a, b, c, e
37. a
38. c
39. a, b, d
40. d
41. a, b, c, e
42. b
43. a, b, c
44. e
45. b
46. a, b, c
47. c
48. a
49. a, c, e
50. d

Test 1 : explications

Question 1. Toutes les autres propositions décrivent l'échantillon dont nous disposons (statistique descriptive). L'inférence consiste à généraliser à la population, et la proposition d fait exactement cela, au moyen d'un intervalle de confiance.

Question 2. Le piège est la proposition d : le biais de sélection est une erreur systématique et ne disparaît pas avec de grands échantillons ; un échantillon de convenance de 10 000 personnes reste tout aussi peu représentatif. La proposition b est fausse pour la même raison.

Question 3. Règle d'addition : $P(F \text{ ou HTA}) = 0{,}30 + 0{,}20 - 0{,}08 = 0{,}42$. Celui qui additionne simplement 0,30 + 0,20 = 0,50 compte deux fois les fumeurs hypertendus.

Question 4. Les classes NYHA ont un ordre naturel (I plus légère que IV), mais les distances entre les classes ne sont pas égales, la variable est donc ordinale.

Question 5. Moyenne = (4 + 7 + 7 + 9 + 13)/5 = 40/5 = 8. Au milieu des valeurs ordonnées (position 3) se trouve la valeur 7, la médiane est donc 7 ; le mode est également 7. L'étendue est 13 - 4 = 9, et non 13. Si 13 devient 43, la moyenne monte à 14, mais la médiane reste à 7 : la médiane est robuste aux valeurs extrêmes.

Question 6. $\text{CV} = \frac{100 \times 12}{80} = 15\%$. La proposition e (0,15%) correspond à l'erreur typique qui consiste à oublier de multiplier par 100, et 6,7% au rapport inversé (80/12 = 6,67).

Question 7. La ligne dans la boîte est la médiane, et non la moyenne (piège a). Le box-plot décrit des variables numériques ou ordinales ; pour les variables nominales, on utilise le diagramme en barres ou le diagramme circulaire.

Question 8. $z = \frac{260-200}{40} = 1{,}5$. D'après la table E.1, Φ(1,5) = 0,9332, donc l'aire à droite est 1 - 0,9332 = 0,0668, soit 6,7%. La proposition b (2,3%) correspond à z = 2.

Question 9. L'aire totale sous toute courbe de densité est égale à 1, quel que soit l'écart-type (d est fausse). Un p = 0,30 au test de Shapiro-Wilk ne "prouve" pas la normalité, il montre seulement qu'aucun écart à la normalité n'a été détecté (e est fausse).

Question 10. Dans une distribution asymétrique, la moyenne et l'écart-type sont tirés par la queue longue. La médiane et l'écart interquartile (151,5 mg/dL ; 107-203,25) décrivent correctement les données. Une autre option correcte, qui ne figure pas parmi les propositions, serait la moyenne géométrique après transformation logarithmique.

Question 11. $z = \frac{160-131{,}13}{13{,}88} = \frac{28{,}87}{13{,}88} = 2{,}08$. La proposition e (28,87) n'est que le numérateur.

Question 12. $\text{ES} = \frac{20}{\sqrt{100}} = 2$; $\text{IC}\,95\% = 150 \pm 1{,}96 \times 2 = 150 \pm 3{,}92$, c'est-à-dire 146,1-153,9. La proposition a (110,8-189,2) est l'intervalle des valeurs individuelles $(m \pm 1{,}96s)$, et non celui de la moyenne : c'est la confusion la plus fréquente à l'examen.

Question 13. $\text{ES} = \frac{\text{ÉT}}{\sqrt{n}}$; pour un n 4 fois plus grand, $\sqrt{n}$ double et l'ES diminue de moitié. L'ÉT ne diminue pas systématiquement avec n : il décrit la variabilité biologique, qui ne dépend pas du nombre de patients mesurés. L'intervalle $m \pm 1{,}96\text{ES}$ contient la moyenne de la population avec une confiance de 95%, et non 95% des patients.

Question 14. $p = \frac{40}{200} = 0{,}20$; $\text{ES} = \sqrt{\frac{0{,}2 \times 0{,}8}{200}} = 0{,}0283$; $\text{IC} = 0{,}20 \pm 1{,}96 \times 0{,}0283 = 0{,}20 \pm 0{,}055$, c'est-à-dire 14,5-25,5%.

Question 15. Lorsque la prévalence est inconnue, on utilise p = 0,5 (l'option la plus prudente) : $n = \frac{1{,}96^2 \times 0{,}5 \times 0{,}5}{0{,}05^2} = 384{,}16$, arrondi à l'entier supérieur, soit 385. La valeur 246 correspond à p = 0,20.

Question 16. La définition de la proposition a est la seule correcte. Les propositions b et c sont les interprétations erronées les plus répandues : la valeur p se calcule en supposant H0 vraie, elle ne peut donc pas être la probabilité que H0 soit vraie.

Question 17. La puissance est égale à 1 - β, et non à 1 - α (c fausse). Un α plus strict (0,01) rend H0 plus difficile à rejeter et, à n égal, diminue la puissance (e fausse).

Question 18. C'est l'exemple classique de résultat non significatif de mg2026. H0 n'est pas "acceptée" et l'égalité n'est pas "démontrée" : l'intervalle de -4,9 à 3,8 mmHg montre que les données sont compatibles aussi avec des différences de quelques mmHg. La valeur 0,81 est le p, et non la puissance.

Question 19. La valeur 1,645 est la valeur critique d'un test unilatéral à α = 0,05. Pour z = 2,30, p bilatéral = 2 x (1 - 0,9893) = 0,021 : inférieur à 0,05, mais supérieur à 0,01, donc d est fausse.

Question 20. ES des différences = 8/√16 = 8/4 = 2 ; $t = \frac{4}{2} = 2{,}0$. La proposition a (0,5) apparaît si l'on divise la moyenne par l'écart-type (4/8), c'est-à-dire si l'on calcule dz, et non t.

Question 21. ddl = n - 1 = 15 ; t critique bilatéral (α = 0,05) = 2,131 (tableau E.3). Comme 2,0 < 2,131, on ne rejette pas H0 (c fausse). La valeur critique pour α = 0,10 en bilatéral est 1,753 < 2,0, donc 0,05 < p < 0,10 (exactement, p = 0,064). Dans un test unilatéral prédéfini, le seuil est 1,753, et le p unilatéral = 0,032, le résultat serait donc significatif (e correcte). Cette proposition ne signifie pas que nous avons le droit de passer à un test unilatéral après avoir vu les données.

Question 22. Trois tests t sans correction augmentent le risque global d'erreur de type I. Pour trois groupes indépendants dont les conditions sont remplies, on utilise l'ANOVA à un facteur.

Question 23. ddl1 = k - 1 = 2, donc k = 3 ; ddl2 = N - k = 157, donc N = 160. $\eta^2 = 0{,}226$ signifie que 22,6% de la variabilité est expliquée par le bras. L'ANOVA seule dit uniquement qu'au moins une moyenne diffère ; les paires sont identifiées par le test post-hoc (Tukey). Un test de Levene à p = 0,163 n'impose pas l'alternative non paramétrique.

Question 25. ddl = (3 - 1) x (4 - 1) = 2 x 3 = 6. La proposition c (12) correspond au nombre de cases.

Question 26. $E = \frac{60 \times 50}{200} = \frac{3000}{200} = 15$.

Question 27. Le test s'applique aux fréquences absolues ; sur des pourcentages, la valeur de χ^2 dépendrait de la base arbitraire 100. Le χ^2 indique à quelle distance les données se trouvent de l'indépendance, mais ni le sens ni l'ampleur de l'association ; pour cela, on utilise le RR, l'OR ou la différence des proportions.

Question 28. La satisfaction sur l'échelle de Likert est ordinale, et les bras sont des groupes indépendants, donc test de Mann-Whitney (dans mg2026 : U = 2 265,5, p < 0,001).

Question 29. Les yeux d'un même patient et les résultats de deux tests appliqués aux mêmes patients sont des observations appariées (pour des variables dichotomiques appariées, on utilise le test de McNemar). La proposition e décrit trois groupes indépendants.

Question 30. Avec 25 patients par groupe, une forte asymétrie et un test de Shapiro-Wilk à p = 0,003, l'option sûre pour deux groupes indépendants est le test de Mann-Whitney (ou bien une transformation logarithmique suivie du test t).

Question 31. $R^2 = 0{,}6^2 = 0{,}36$.

Question 32. PAS prédite = 89,47 + 1,435 x 30 = 89,47 + 43,05 = 132,5 mmHg. $R^2 = 0{,}405^2 = 0{,}164$. L'ordonnée à l'origine correspond à un IMC de 0, valeur impossible et très éloignée des données ; ce n'est donc la PAS "réelle" d'aucun patient. La corrélation ne démontre pas la causalité.

Question 33. Le coefficient de Spearman travaille sur les rangs et mesure une relation monotone ; il convient aux variables ordinales ou aux relations monotones non linéaires. Pour les variables nominales, on utilise des tableaux de contingence, et pour une équation de prédiction, la régression.

Question 34. $R^2 = 0{,}15^2 = 0{,}0225$, c'est-à-dire 2,25%. Avec n = 210, même un r de 0,15 devient significatif (tableau E.6 : pour n = 100, le seuil à α = 0,05 est 0,197, et pour n = 210 le seuil est plus bas). Selon les règles empiriques présentées dans le cours, un |r| inférieur à 0,25 correspond à une association pratiquement négligeable. La signification statistique ne signifie ni pertinence ni causalité.

Question 35. Risque chez les exposés = 30/100 = 0,30 ; chez les non-exposés = 10/100 = 0,10 ; RR $= \frac{0{,}30}{0{,}10} = 3{,}0$.

Question 36. OR $= \frac{30\times90}{70\times10} = \frac{2700}{700} = 3{,}86$, supérieur au RR (d fausse : lorsque la maladie n'est pas rare, l'OR dépasse le RR pour un facteur de risque). Différence de risques = 0,30 - 0,10 = 0,20 ; NNH $= \frac{1}{0{,}20} = 5$. Fraction attribuable chez les exposés = (3 - 1)/3 = 0,667.

Question 37. Dans l'étude cas-témoins, la proportion de malades est fixée par le chercheur, les risques et les incidences ne peuvent donc pas être calculés ; l'OR est la seule mesure valide.

Question 38. RAR = 20% - 15% = 5% = 0,05 ; NNT $= \frac{1}{0{,}05} = 20$. La proposition 33 correspondrait à 1/0,03 (une erreur de soustraction), et 5 à 1/0,20.

Question 39. Un facteur de confusion est lié à la fois à l'exposition et au résultat, mais ce n'est pas un maillon intermédiaire (celui-ci s'appelle un médiateur). La randomisation le contrôle ; l'augmentation de la taille de l'échantillon, non, car il s'agit d'une erreur systématique.

Question 40. Pour 1 000 personnes : 20 malades, dont 19 positifs (95%) ; 980 non-malades, dont 98 faux positifs (10%). VPP $= \frac{19}{19+98} = \frac{19}{117} = 0{,}162$, soit environ 16%. C'est l'exemple typique : avec une maladie rare, même un bon test donne plus de faux positifs que de vrais positifs.

Question 41. Cote pré-test = 0,10/0,90 = 0,111 ; cote post-test = 0,111 x 9 = 1,0 ; probabilité post-test = 1/(1 + 1) = 0,50. Un test de Se 90% et de Sp 90% a $RV^+ = \frac{0{,}90}{0{,}10} = 9$.

Question 43. Tableau complet : a = 45, b = 5, c = 20, d = 130, N = 200. Se $= \frac{45}{50} = 90\%$; Sp $= \frac{130}{150} = 86{,}7\%$; VPP $= \frac{45}{65} = 69{,}2\%$; VPN $= \frac{130}{135} = 96{,}3\%$; exactitude = (45 + 130)/200 = 87,5%. Le taux d'erreur est $\frac{b+c}{N} = \frac{5+20}{200}$, et non (5 + 130)/200.

Question 44. Taux d'erreur de classement $= \frac{\text{FN+FP}}{N} = \frac{5+20}{200} = \frac{25}{200} = 12{,}5\% = 1 - 0{,}875$.

Question 45. Malades = 0,30 x 200 = 60 ; a = 0,80 x 60 = 48, b = 12. Non-malades = 140 ; d = 0,90 x 140 = 126, c = 14. VPP $= \frac{48}{48+14} = \frac{48}{62} = 77{,}4\%$. La proposition e (91,3%) correspond à VPN $= \frac{126}{138}$.

Question 46. L'AUC n'est pas le pourcentage de patients correctement classés : c'est la probabilité qu'un malade choisi au hasard ait un score plus élevé qu'un non-malade choisi au hasard. Le seuil optimal selon Youden maximise Se + Sp - 1, et non la seule sensibilité.

Question 47. J = 0,85 + 0,75 - 1 = 0,60.

Question 48. L'analyse en intention de traiter conserve les groupes tels que la randomisation les a constitués et c'est l'analyse principale des essais de supériorité.

Question 49. La dissimulation de l'allocation (on ne connaît pas à l'avance le bras suivant) et l'insu (on ne connaît pas le traitement reçu pendant la durée de l'étude) sont deux choses différentes. Le critère de jugement principal est fixé dans le protocole, avant la collecte des données.

Question 50. Sans IC, nous ne pouvons pas juger de la précision de l'estimation : un OR de 2,1 sur 80 patients peut avoir un intervalle très large, compatible aussi bien avec un effet faible qu'avec un effet important.

Test 2 : grille des réponses

1. a
2. a, b, d
3. b
4. c
5. a, b, d
6. d
7. e
8. b
9. a, b, c, e
10. c
11. a
12. b
13. d
14. a, c, d, e
15. e
16. a, b, d
17. c
18. a
19. a, b, d
20. c
21. a, b, c, e
22. d
23. a
24. e
25. a, b, c, e
26. b
27. a
28. a, b, d
29. d
30. a, b, c
31. b
32. a, b, c, d
33. a
34. c
35. e
36. a, b, c
37. a

38. d
39. a, b, c, e
40. b
41. a, b, d
42. c
43. a, b, c, d
44. c
45. a
46. d
47. a, b, c, e
48. c
49. a, b, d
50. e

Test 2 : explications

Question 1. Une valeur qui décrit la population et que nous estimons est un paramètre (μ) ; la moyenne calculée sur l'échantillon serait la statistique (m).

Question 2. L'anonymisation suppose que la personne ne puisse plus être réidentifiée ; le remplacement du nom par un code, avec conservation de la clé de correspondance, est une pseudonymisation (c fausse). Un jeu de données simulé comportant des patients fictifs ne contient pas de données de personnes réelles ; il ne s'agit donc pas de données à caractère personnel (e fausse).

Question 3. Pour des événements indépendants, $P(A \text{ et } B) = 0{,}3 \times 0{,}4 = 0{,}12$. La valeur 0,70 serait la somme, et 0,58 serait P(A ou B).

Question 4. n = 6 (pair), la médiane est donc la moyenne des valeurs des positions 3 et 4 : (8 + 10)/2 = 9. La moyenne est 58/6 = 9,67, le piège de la proposition b.

Question 5. Réponses valides = 200 - 20 = 180 ; pourcentage valide = 54/180 = 30% ; pourcentage du total $= \frac{54}{200} = 27\%$. La différence montre pourquoi le rapport doit préciser la base de calcul et le nombre de valeurs manquantes. Les valeurs manquantes ne se remplacent pas arbitrairement.

Question 6. EIQ = 30 - 20 = 10 ; limite supérieure = Q3 + 1,5 x EIQ = 30 + 15 = 45.

Question 7. L'histogramme montre la forme de la distribution d'une variable numérique. Le diagramme en barres et le diagramme circulaire sont destinés aux variables catégorielles, le diagramme de dispersion à deux variables numériques, et le forest plot aux estimations avec IC de plusieurs études.

Question 8. $z = \frac{11{,}6-14}{1{,}2} = -2{,}0$; Φ(-2) = 1 - Φ(2) = 1 - 0,9772 = 0,0228, soit 2,3%.

Question 9. m ± 1,96s = 120 ± 19,6 = 100,4-139,6 (95% des individus) ; $m \pm s = 110 - 130(68{,}3\%)$; $m \pm 3s = 90 - 150(99{,}7\%)$. Au-dessous de 100,4 se trouvent 2,5% des individus (la moitié des 5% restés en dehors de l'intervalle). L'intervalle 100,4-139,6 décrit les individus, et non la moyenne de la population : un IC de la moyenne utiliserait ES $= \frac{s}{\sqrt{n}}$ et serait beaucoup plus étroit.

Question 10. L'exponentielle de la moyenne des logarithmes est la moyenne géométrique ; dans mg2026, pour les triglycérides, elle est de 144,0 mg/dL, contre une moyenne arithmétique de 161,18 mg/dL.

Question 11. Le théorème porte sur la distribution des moyennes des échantillons, et non sur la distribution de la variable. La proposition d est fausse : l'écart-type ne diminue pas avec n ; c'est l'erreur standard qui diminue.

Question 12. ES $= \frac{8}{\sqrt{16}} = 2$; IC 95% $= 50 \pm 2{,}131 \times 2 = 50 \pm 4{,}26$, c'est-à-dire 45,7-54,3. La proposition a (46,1-53,9) utilise z = 1,96, ce qui est inadapté pour n = 16 avec σ inconnu.

Question 13. $n = \left(\frac{1{,}96 \times 10}{2}\right)^2 = 9{,}8^2 = 96{,}04$. La taille de l'échantillon s'arrondit toujours à l'entier supérieur, donc 97 ; avec 96 patients, la marge serait légèrement supérieure à 2 unités.

Question 14. L'intervalle de confiance se rapporte à la moyenne de la population, et non aux valeurs des patients (b fausse). Un niveau de confiance plus élevé exige un intervalle plus large ; pour un n 4 fois plus grand, l'ES diminue de moitié.

Question 15. $p = \frac{2}{25} = 0{,}08$; ES $= \sqrt{\frac{0{,}08 \times 0{,}92}{25}} = 0{,}054$; limite inférieure de Wald = 0,08 - 1,96 x 0,054 = -0,026, impossible pour une proportion. L'intervalle de Wilson résout le problème et reste compris entre 0 et 1.

Question 16. H0 n'est ni "acceptée" ni "démontrée" (c fausse). Pour une corrélation, H0 est ρ = 0 (absence de corrélation), et non ρ = 1.

Question 17. α' = 0,05/5 = 0,01.

Question 18. Avec 20 000 patients, même une diminution de 1,2 mmHg devient extrêmement significative sur le plan statistique. La borne supérieure de l'IC (1,6 mmHg) montre toutefois que l'effet est, tout au plus, très faible sur le plan clinique.

Question 19. La puissance augmente avec n, avec la taille de l'effet et avec α, et elle diminue avec la variabilité. Le schéma apparié élimine la variabilité interindividuelle ; y renoncer, pour le même nombre de mesures, diminue en règle générale la puissance.

Question 20. ddl = n1 + n2 - 2 = 25 + 30 - 2 = 53.

Question 21. ES $= \frac{11{,}57}{\sqrt{160}} = \frac{11{,}57}{12{,}649} = 0{,}915$; $t = \frac{10{,}19}{0{,}915} = 11{,}14$; ddl = 159 ; dz $= \frac{10{,}19}{11{,}57} = 0{,}88$ (effet fort selon les repères de Cohen). Le test t apparié n'exige pas l'égalité des variances aux deux temps de mesure : il travaille sur une seule série, celle des différences, dont la normalité doit être vérifiée (dans mg2026, Shapiro-Wilk p = 0,725).

Question 22. En cas de variances inégales, on utilise le test t de Welch, qui ajuste les degrés de liberté.

Question 23. $\eta^2 = \frac{\text{SSB}}{\text{SST}} = \frac{300}{1200} = 0{,}25$.

Question 24. $d = \frac{136-130}{12} = 0{,}5$, effet moyen selon les repères de Cohen. La proposition b donne le bon chiffre, mais une interprétation erronée.

Question 25. $\chi^2 = \frac{N(\text{ad}-\text{bc})^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)} = \frac{100 \times (800-300)^2}{50 \times 50 \times 30 \times 70} = 25\ 000\ 000/5\ 250\ 000 = 4{,}76$, avec ddl = 1. Comme 3,841 < 4,76 < 6,635, il en résulte 0,01 < p < 0,05 (exactement, p = 0,029). La fréquence attendue pour $a = \frac{(a+b)\,(a+c)}{N} = \frac{50 \times 30}{100} = 15$.

Question 26. Les résultats des deux tests proviennent des mêmes patients (données dichotomiques appariées), donc test de McNemar. Le test du khi-deux d'indépendance supposerait des groupes indépendants.

Question 27. Nous classons l'ensemble des 7 valeurs : 12 (A, rang 1), 14 (B, 2), 15 (A, 3), 18 (A, 4), 20 (B, 5), 22 (B, 6), 25 (B, 7). $R(A) = 1 + 3 + 4 = 8$; $U(A) = 8 - \frac{3\times4}{2} = 2$; $U(B) = 3 \times 4 - 2 = 10$. La plus petite valeur est U = 2 (p exact = 0,23, non significatif avec des échantillons aussi petits).

Question 28. Un test de Kruskal-Wallis significatif dit seulement qu'au moins un groupe diffère ; les paires sont identifiées par des tests post-hoc (Dunn). Le test de Friedman est destiné aux mesures répétées, et non aux groupes indépendants.

Question 29. Les deux variables sont dichotomiques, et les fréquences attendues sont supérieures à 5 : test du khi-deux sur le tableau 2 x 2 (suivi du calcul de l'OR ou du RR avec IC).

Question 30. Pour deux variables numériques normales liées par une relation linéaire, le coefficient de Pearson est le choix approprié ; le coefficient de Spearman n'est pas "obligatoire". Pour trois groupes indépendants à données normales, on utilise l'ANOVA à un facteur, et non le test de Friedman.

Question 31. rho $= 1 - \frac{6\times14}{6\times(36-1)} = 1 - \frac{84}{210} = 1 - 0{,}4 = 0{,}6$.

Question 32. ddl = 27 - 2 = 25 ; $t = \frac{0{,}5\times\sqrt{25}}{\sqrt{1-0{,}25}} = \frac{2{,}5}{0{,}866} = 2{,}89 > 2{,}060$, donc p < 0,05 (exactement, p = 0,008). $R^2 = 0{,}25$. Le tableau E.6 le confirme : pour n = 27, le r critique à 0,05 est inférieur à 0,5.

Question 33. 108,3 + 0,406 x 60 = 108,3 + 24,36 = 132,7 mmHg. La proposition b (24,4) n'est que le produit.

Question 34. Dans un modèle multiple, le coefficient d'un prédicteur est son effet ajusté sur les autres prédicteurs : +1,389 mmHg pour chaque kg/m², à âge, sexe et statut tabagique égaux.

Question 35. Cote d'exposition chez les cas = 40/60 ; chez les témoins = 20/80 ; OR $= \frac{40\times80}{60\times20} = \frac{3200}{1200} = 2{,}67$. La proposition a (2,0) est le rapport des proportions d'exposés (40%/20%), qui n'a pas le sens d'un risque dans une étude cas-témoins.

Question 36. L'IC contient 1, donc p > 0,05 et l'association n'est pas significative. Mais l'IC s'étend jusqu'à 3,6 : les données n'excluent pas un effet important. Le résultat non significatif ne démontre pas l'absence d'effet (d fausse).

Question 37. Incidence cumulée = 50/2 000 = 0,025 = 2,5% en 2 ans. La densité d'incidence exigerait des personnes-temps (environ 50/4 000 = 12,5 pour 1 000 personnes-années, si tous ont été suivis pendant 2 ans) ; les propositions d et e ne correspondent à aucun calcul correct.

Question 38. La sélection part de la maladie (cas et témoins), et l'exposition est recherchée rétrospectivement : étude cas-témoins.

Question 39. Prévalences : 18/54 = 33,3% et 11/106 = 10,4% ; RP = 3,21 ; OR $= \frac{18\times95}{36\times11} = \frac{1710}{396} = 4{,}32$. Une étude transversale mesure l'exposition et la maladie simultanément, elle ne peut donc pas établir leur ordre chronologique (d fausse). Le diabète n'est pas rare dans l'échantillon (18,1%), c'est pourquoi l'OR s'éloigne du RP.

Question 40. Cote pré-test = 0,2/0,8 = 0,25 ; cote post-test = 0,25 x 8 = 2,0 ; probabilité post-test = 2/3 = 66,7%. La proposition e (160%) apparaît lorsqu'on multiplie directement la probabilité par le RV, une erreur impardonnable : les probabilités ne peuvent pas dépasser 100%.

Question 41. L'opinion d'expert se trouve à la base de la hiérarchie des preuves, et une étude observationnelle, aussi grande soit-elle, n'élimine pas les facteurs de confusion comme le fait la randomisation.

Question 42. À Se et Sp constantes, la VPP diminue et la VPN augmente lorsque la prévalence diminue. Dans mg2026, pour FINDRISC ≥ 12, la VPP serait de 5,8% pour une prévalence de 2% et de 75,1% pour une prévalence de 50%.

Question 43. $a + c = \frac{a}{\text{VPP}} = \frac{40}{0,8} = 50$, donc c = 10. $b + d = \frac{b}{1-\text{VPN}} = \frac{10}{0,1} = 100$, donc d = 90. Se $= \frac{40}{40+10} = 80\%$; Sp $= \frac{90}{10+90} = 90\%$. N = 40 + 10 + 10 + 90 = 150, et non 200.

Question 44. κ = (0,85 - 0,50)/(1 - 0,50) = 0,35/0,50 = 0,70 (accord important). La proposition a (0,35) n'est que le numérateur.

Question 45. SnNOut : un test très sensible manque peu de malades, un résultat négatif exclut donc pratiquement la maladie.

Question 46. Un seuil plus élevé rend le test plus "sévère" : moins de positifs, donc une sensibilité plus faible et une spécificité plus élevée. L'aire sous la courbe ROC caractérise le test sur l'ensemble des seuils et ne change pas lorsqu'on déplace le seuil.

Question 47. L'utilisation de tests en série (positif seulement si les deux tests sont positifs) augmente la spécificité et diminue la sensibilité (d fausse) ; l'utilisation de tests en parallèle fait l'inverse.

Question 48. n ajusté $= \frac{64}{1-0,20} = \frac{64}{0,8} = 80$. La proposition b (77) provient du calcul erroné 64 x 1,2 = 76,8 : ajouter 20% à 64 ne compense pas une perte de 20% de l'effectif recruté.

Question 49. Les variables numériques se recueillent sous forme de nombres, avec l'unité fixée dans le formulaire et une plage de validation ; le texte libre avec unité ("87 mg/dl") est exactement l'erreur du fichier brut mg2026. Les essais cliniques sont enregistrés avant l'inclusion du premier patient.

Question 50. Une valeur p n'est jamais exactement égale à 0 ; le logiciel affiche "<.001" parce que la valeur est très petite. On rapporte p < 0,001.

Test 3 : grille des réponses

1. c
2. a, c, d
3. a, b, c, e
4. b
5. e
6. a, b, d
7. a
8. d
9. a, b, c, e
10. b
11. e
12. a
13. a, b, c, d
14. c
15. d

16. c
17. b, c, d
18. a
19. d
20. b
21. a, b, c, e
22. e
23. a
24. a, b, c
25. b
26. a
27. a, b, d
28. d
29. c
30. a, b, c, e
31. d
32. a, b, c
33. b
34. e
35. a
36. a, b, c, e
37. b
38. c
39. a, b, d
40. a
41. a, b, c, d
42. c
43. d
44. a, b, c, e
45. e
46. a, b, d, e
47. b
48. d
49. a, b, d
50. c

Test 3 : explications

Question 2. Le biais ne diminue pas avec la taille de l'échantillon (b fausse), et l'intervalle de confiance ne quantifie que l'erreur aléatoire, et non le biais (e fausse). Un IC étroit issu d'une étude biaisée est une estimation précise d'une valeur erronée.

Question 3. $P(A) \times P(B) = 0{,}3 \times 0{,}4 = 0{,}12 = P(A \text{ et } B)$, donc A et B sont indépendants ; $P(A \mid B) = \frac{0{,}12}{0{,}4} = 0{,}3 = P(A)$, même conclusion. $P(A \text{ ou } B) = 0{,}3 + 0{,}4 - 0{,}12 = 0{,}58$. Les événements ne sont pas incompatibles, car P(A et B) ≠ 0. $P(\text{non} - A) = 1 - 0{,}3 = 0{,}7$.

Question 4. Moyenne géométrique = (10 x 100 x 1 000)^(1/3) = (1 000 000)^(1/3) = 100. La moyenne arithmétique (370) est dominée par la valeur la plus élevée.

Question 5. $\text{CV}(A) = \frac{5}{50} = 10\%$; $\text{CV}(B) = \frac{30}{200} = 15\%$. B a une variabilité à la fois absolue (s) et relative plus grande.

Question 6. Le groupe sanguin est nominal (les catégories n'ont pas d'ordre). La température en degrés Celsius se situe sur une échelle d'intervalle : 0 °C ne signifie pas l'absence de température, et 20 °C n'est pas "deux fois plus chaud" que 10 °C.

Question 9. L'asymétrie positive (1,42) tire la moyenne (29,04) au-dessus de la médiane (28,25). Pour une distribution non normale, l'intervalle $m \pm 1{,}96s$ ne contient pas exactement 95% des patients.

Question 10. L'aire à droite de 1,96 est de 2,5% (Φ(1,96) = 0,9750) ; 1,645 laisse 5% à droite, et 2,576 laisse 0,5%.

Question 11. Le 90e percentile signifie que 90% des enfants de référence ont des valeurs inférieures ; ce n'est pas un diagnostic d'obésité, et cela n'exprime pas un pourcentage de la moyenne.

Question 12. $p = \frac{18}{60} = 0{,}30$; $\text{ES} = \sqrt{\frac{0{,}3 \times 0{,}7}{60}} = 0{,}0592$; $\text{IC} = 0{,}30 \pm 1{,}96 \times 0{,}0592 = 0{,}30 \pm 0{,}116$, c'est-à-dire 18,4-41,6%. La proposition e confond la proportion et le pourcentage.

Question 13. $n = \frac{z^2 \times p_i(1-p_i)}{E^2}$ augmente lorsque E diminue, lorsque z augmente (99% contre 95%), et pour les moyennes $n = \left(\frac{z\sigma}{E}\right)^2$ augmente avec σ. Le produit $p_i(1 - p_i)$ est maximal pour $p_i = 0{,}5$. La taille de la population ne compte que lorsque l'échantillon en représente une fraction importante ; passer de 1 à 2 millions ne change pratiquement rien.

Question 14. La largeur de l'IC est proportionnelle à $\frac{1}{\sqrt{n}}$: √(40/160) = √0,25 = 0,5, l'intervalle devient donc environ deux fois plus étroit.

Question 15. $n = \frac{1{,}96^2 \times 0{,}10 \times 0{,}90}{0{,}03^2} = \frac{3{,}8416 \times 0{,}09}{0{,}0009} = 384{,}16$, donc 385.

Question 16. $\text{ES} = \frac{16}{\sqrt{64}} = 2$; $t = \frac{128-120}{2} = 4$, avec ddl = 63 ($p < 0{,}001$).

Question 17. La valeur p n'est pas la probabilité que H0 soit vraie (a fausse). Les valeurs 0,049 et 0,051 apportent pratiquement la même preuve ; le seuil de 0,05 est une convention, et non une frontière naturelle (e fausse).

Question 18. Choisir un test unilatéral après avoir vu les données (par exemple pour "sauver" un p bilatéral de 0,08) est une pratique incorrecte.

Question 19. Probabilité qu'aucun test ne soit faussement significatif = 0,95^20 = 0,358 ; probabilité d'au moins un résultat faussement positif = 1 - 0,358 = 0,64. D'où la nécessité des corrections pour comparaisons multiples.

Question 20. $F = \frac{\text{MSB}}{\text{MSW}} = \frac{150}{30} = 5$.

Question 21. ddl intergroupes = k - 1 = 2 ; ddl résiduels = N - k = 60 - 3 = 57. Comme 5 > 3,16, $p < 0{,}05$ (exactement, p = 0,010). Une ANOVA significative montre seulement qu'au moins une moyenne diffère, et non que toutes diffèrent entre elles.

Question 24. Le test t n'exige pas des groupes d'effectifs égaux ; il exige des observations indépendantes, la normalité (ou de grands échantillons) et, dans sa version classique, des variances égales.

Question 25. ddl = (2 - 1) x (3 - 1) = 2.

Question 26. $U_1 = 12 - \frac{4\times5}{2} = 12 - 10 = 2$; U2 = n1 x n2 - U1 = 20 - 2 = 18.

Question 27. La fréquence attendue minimale (12,6) dépasse 5, le test du khi-deux est donc approprié et le test de Fisher n'est pas nécessaire. Le test global sur le tableau 3 x 2 montre seulement que les proportions ne sont pas toutes égales ; de plus, les bras B et C ont des proportions presque identiques (16/53 et 16/54), la proposition c est donc fausse.

Question 28. Mesures répétées (mêmes patients, trois temps de mesure) et données non normales sur un petit échantillon : test de Friedman. Le test de Kruskal-Wallis conviendrait pour trois groupes indépendants.

Question 29. Deux groupes indépendants, données normales, variances homogènes (Levene p = 0,40) : test t pour échantillons indépendants ; la version de Welch est elle aussi acceptable.

Question 30. Pour une variable ordinale, on utilise des tests sur les rangs ou, en perdant l'information sur l'ordre, le test du khi-deux. Traiter les codes 0-4 comme des valeurs continues normales suppose des distances égales entre les stades, ce qui n'est pas justifié.

Question 31. |r| = √0,25 = 0,5 ; le signe de r est celui de la pente, donc r = -0,5.

Question 32. L'extrapolation en dehors du domaine observé est risquée, et un R^2 élevé indique une association forte, et non une causalité.

Question 33. $\text{OR} = e^{0,693} = 2,0$. Le coefficient b est sur l'échelle du logarithme de la cote et doit être exponentié.

Question 34. Une valeur extrême sur les deux axes peut créer ou gonfler une corrélation. Le point est vérifié dans le document source et l'analyse est faite avec et sans lui, mais il n'est pas supprimé simplement parce qu'il "gâche" le résultat.

Question 35. Risque chez les traités = 8/200 = 0,04 ; risque sous placebo = 16/200 = 0,08 ; $\text{RR} = \frac{0,04}{0,08} = 0,5$.

Question 36. RAR = 0,08 - 0,04 = 0,04 (4 points de pourcentage) ; $\text{RRR} = \frac{0,04}{0,08} = 50\%$; $\text{NNT} = \frac{1}{0,04} = 25$. $\text{OR} = \frac{8\times184}{192\times16} = \frac{1472}{3072} = 0,48$, proche du RR (l'événement est rare), mais pas égal. Un "risque réduit de moitié" paraît spectaculaire, alors qu'il ne signifie que 4 événements évités pour 100 patients traités.

Question 37. $\text{FAP} = \frac{0,25\times(3-1)}{1+0,25\times(3-1)} = \frac{0,5}{1,5} = 0,333$.

Question 38. La prévalence est approximativement égale à l'incidence multipliée par la durée de la maladie ; une durée longue fait que la prévalence est élevée même avec une incidence faible.

Question 39. Dans l'étude cas-témoins, on ne peut pas calculer de risques, donc pas non plus de RR direct (c fausse). L'étude transversale n'établit pas l'ordre chronologique (e fausse).

Question 40. Pour 10 000 personnes : 100 malades, 90 positifs ; 9 900 non-malades, 495 faux positifs (5%). $\text{VPP} = \frac{90}{90+495} = \frac{90}{585} = 15,4\%$. Avec une prévalence de 20%, le même test aurait une VPP d'environ 82%.

Question 41. $\mathrm{RV}^- = \frac{1-0{,}8}{0{,}9} = \frac{0{,}2}{0{,}9} = 0{,}22$. Les rapports de vraisemblance se calculent uniquement à partir de la Se et de la Sp, ils ne dépendent donc pas de la prévalence (e fausse) ; c'est pourquoi ils sont plus faciles à transposer d'une étude à l'autre que les valeurs prédictives.

Question 42. p = cote/(1 + cote) = 3/4 = 0,75. La proposition b (0,33) est l'inverse de la cote.

Question 43. $\mathrm{DOR} = \frac{a \times d}{b \times c} = \frac{80 \times 90}{20 \times 10} = \frac{7200}{200} = 36$. Vérification : $\mathrm{RV}^+ = \frac{0{,}8}{0{,}1} = 8$, $\mathrm{RV}^- = \frac{0{,}2}{0{,}9} = 0{,}222$, et 8/0,222 = 36.

Question 44. $\mathrm{Se} = \frac{80}{100} = 80\%$; $\mathrm{Sp} = \frac{90}{100} = 90\%$; exactitude = (80 + 90)/200 = 85% ; taux d'erreur = $\frac{b+c}{N} = \frac{20+10}{200} = 15\% = 1 - 0{,}85$. La formule (b + d)/N, présente dans certains supports, est erronée.

Question 45. Dans le scénario mg2026, le diabète a été confirmé par l'hyperglycémie provoquée par voie orale (HGPO), standard de référence par rapport auquel on calcule la Se et la Sp du score FINDRISC.

Question 46. Une aire comprise entre 0,81 et 0,90 indique une très bonne discrimination. L'IC 0,762-0,943 exclut 0,5. La différence de 0,008 entre les deux aires est bien inférieure à l'incertitude de chaque estimation ; une supériorité ne pourrait être affirmée qu'après un test formel de comparaison des aires (c fausse).

Question 47. Lorsque le standard de référence est appliqué de manière sélective, en fonction du résultat du test étudié, apparaît le biais de vérification, qui surestime habituellement la sensibilité.

Question 48. La phase III compare de façon randomisée, sur de grands effectifs, le nouveau traitement au traitement standard ou à un placebo, avant l'autorisation de mise sur le marché. La phase IV a lieu après l'autorisation.

Question 49. De nombreux essais randomisés sont ouverts (par exemple ceux qui portent sur des interventions éducatives, comme les bras de mg2026) ; la randomisation ne dépend pas de l'insu. L'analyse per protocole exclut les patients qui n'ont pas respecté le protocole et peut détruire l'équilibre créé par la randomisation.

Question 50. Le critère de jugement est dichotomique et il y a trois groupes, donc test du khi-deux sur le tableau 3 x 2 (χ^2 = 6,77, p = 0,034 dans mg2026). Pour la comparaison clinique C par rapport à A, on rapporte un RR = 2,62 (IC 95% 1,11-6,17), une différence de risques de 18,3 points de pourcentage (IC 95% 3,4-33,2) et un NST = 5,5, soit 6.

Bibliographie

Les références sont numérotées dans l'ordre de leur première apparition dans le texte (style Vancouver). Chaque article a été vérifié dans PubMed ou CrossRef avant d'être cité ; les pages officielles ont été consultées à la date indiquée.

1. Muntean C. Informatică medicală: metode, sisteme și practică. Orizont 2026. Curs și lucrări practice pentru studenții Facultății de Medicină. Timișoara: Editura Victor Babeș; 2026. ISBN 978-606-786-641-4.
2. Lupușoru R, Sporea I, Rațiu I, Lungeanu D, Popescu A, Dănilă M, et al. Contrast-Enhanced Ultrasonography with Arrival Time Parametric Imaging as a Non-Invasive Diagnostic Tool for Liver Cirrhosis. Diagnostics (Basel). 2022;12(12):3013. doi:10.3390/diagnostics12123013 PMID: 36553020
3. Moga TV, Sporea I, Lupușoru R, Popescu A, Popa A, Bota S, et al. Performance of a Noninvasive Time-Harmonic Elastography Technique for Liver Fibrosis Evaluation Using Vibration Controlled Transient Elastography as Reference Method. Diagnostics (Basel). 2020;10(9):653. doi:10.3390/diagnostics10090653 PMID: 32878078
4. Moleriu RD, Piroș T, Moleriu LC, Muntean C, Lupușoru R, Dicu AM, et al. Convenience-Oriented Dietary Behavioral Patterns Across BMI Classes in University Students: Associations with Overweight and Obesity Risk During the Transition to University Life. Nutrients. 2026;18(14):2368. doi:10.3390/nu18142368 PMID: 42514438
5. Muntean C, Piroș T, Marin RC, Moleriu LC, Lupușoru R, Dicu AM, et al. Chrononutrition Behaviors and BMI Change During the Transition to University Life in Young Adults: A Prospective Study. Nutrients. 2026;18(12):1975. doi:10.3390/nu18121975 PMID: 42356360
6. Piroș T, Lupusoru R, Moleriu LC, Muntean C, Moleriu RD, Cîmpian DM, et al. Dietary Patterns, Cooking Methods, and Their Association with Prediabetes Risk Markers in Romanian University Students: A Cross-Sectional Analysis. Nutrients. 2026;18(6):977. doi:10.3390/nu18060977 PMID: 41901152
7. Frandes M, Gherbon A, Timar B, Muntean C. Digital Engagement in Diabetes Care: A Multi-Domain Analysis of Psychosocial and Clinical Determinants. Healthcare (Basel). 2026;14(6):800. doi:10.3390/healthcare14060800 PMID: 41897253
8. Cioca F, Timar R, Ignuta F, Vlad A, Bratosin F, Rosca O, et al. Comparative Analysis of COVID-19 Outcomes in Type 1 and Type 2 Diabetes: A Three-Year Retrospective Study. Medicina (Kaunas). 2024;60(2):210. doi:10.3390/medicina60020210 PMID: 38399498
9. Popa DI, Bălulescu LA, Mederle OA, Levai CM, Buleu T, Tudor A, et al. Sex-Related Differences in Clinical Characteristics and Outcomes Among Patients Clinically Managed for Suspected Bacterial Infection in the Emergency Department. Medicina (Kaunas). 2026;62(9):1776. doi:10.3390/medicina62091776 PMID: 42796382
10. Marin RC, Păunescu H, Muntean C, Balaban DV, Coman LI, Radu N, et al. Assessing Acceptance and Intention to Use of Home Telemonitoring Among Patients with Liver Cirrhosis: Development and Initial Psychometric Validation of a TAM-Based Questionnaire. Diagnostics (Basel). 2026;16(16):2635. doi:10.3390/diagnostics16162635 PMID: 42651036
11. Cirstea N, Tit DM, Toma MM, Purza AL, Radu A, Bungau GS, et al. Patient-Reported Experiences in Chronic Dermatological Conditions: Validation of the Romanian PSPSQ 2.0 Within Contemporary Dermatologic Care Pathways. Diagnostics (Basel). 2026;16(13):2112. doi:10.3390/diagnostics16132112 PMID: 42449893
12. Focșa M, Tudor A, Cătuș C, Muntean C, Mada L. Informatică medicală: îndrumător de lucrări practice. Timișoara: Editura Victor Babeș; 2010.

13. Vernic C, Mada L, Lungeanu D, Muntean C, Apostol AS, Cătuș CO. Aplicații practice de informatică medicală și biostatistică. Timișoara: Editura Victor Babeș; 2010.
14. Vernic C, Frandeș M, Tudor A, Apostol AS, Voicu A, Șerban C, et al. Prelucrarea datelor biomedicale: aplicații practice. Timișoara: Eurostampa; 2018.
15. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312(7023):71-2. doi:10.1136/bmj.312.7023.71 PMID: 8555924
16. Lungeanu D, Petrica A, Lupusoru R, Marza AM, Mederle OA, Timar B. Beyond the Digital Competencies of Medical Students: Concerns over Integrating Data Science Basics into the Medical Curriculum. Int J Environ Res Public Health. 2022;19(23):15958. doi:10.3390/ijerph192315958 PMID: 36498065
17. Lungeanu D, Dragan I, Muntean C, Timar B. Harnessing Information Technology while Teaching in Distress: the Pandemic Toll vs. the Opportunity of Long-Awaited Paradigm Change. In: 2020 International Conference on e-Health and Bioengineering (EHB); Iași, Romania. IEEE; 2020. p. 1-4. doi:10.1109/ehb50910.2020.9280122
18. Muntean C. Metode computaționale în cercetarea medicală, stomatologică și farmaceutică: ghid metodologic pentru studii doctorale și cercetare aplicată. Timișoara: Editura Victor Babeș; 2026. ISBN 978-606-786-644-5.
19. Mihalaș GI, Lungeanu D. Biostatistică. Timișoara: Editura Victor Babeș; 2011. ISBN 978-606-8054-35-3.
20. Mischie A, Jinga V, Nodiți-Cuc AR, Popovici RA, Marian D, Strete GE, et al. Self-Reported Disordered Eating and Weight-Control Behaviours Across BMI Categories in Romanian Young Adults: A Cross-Sectional Online Convenience Survey. Nutrients. 2026;18(16):2677. doi:10.3390/nu18162677 PMID: 42654257
21. Vesa CM, Tit DM, Babes EE, Bungau G, Radu AF, Moleriu RD. Use of Botanical Supplements Among Romanian Individuals with Diabetes: Results from an Online Study on Prevalence, Practices, and Glycemic Control. Nutrients. 2025;17(15):2440. doi:10.3390/nu17152440 PMID: 40806023
22. Popa MD, Sharma A, Kundnani NR, Gag OL, Rosca CI, Mocanu V, et al. Identification of Heavy Tobacco Smoking Predictors-Influence of Marijuana Consuming Peers and Truancy among College Students. Healthcare (Basel). 2021;9(12):1666. doi:10.3390/healthcare9121666 PMID: 34946393
23. Christodorescu RM, Muntean C, Andor M, Lighezan D, Pop Moldovan A, Blajevschi A, et al. Hemoglobin-to-Red Cell Distribution Width Ratio as a Prognostic Marker in Decompensated Heart Failure Patients: A Prospective Observational Study. Life (Basel). 2026;16(4):551. doi:10.3390/life16040551 PMID: 42073361
24. Schiavenato M, Chu F. PICO: What it is and what it is not. Nurse Educ Pract. 2021;56:103194. doi:10.1016/j.nepr.2021.103194
25. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370(9596):1453-7. doi:10.1016/S0140-6736(07)61602-X PMID: 18064739
26. Cordoș AA, Ștefănigă SA, Muntean C, Vernic CV, Bolboacă SD. ALSATION Study Protocol: Romanian Translation of Three Health Literacy Surveys. Stud Health Technol Inform. 2024;321:12-16. doi:10.3233/SHTI241052 PMID: 39575771
27. Jianu N, Buda VO, Căpățână D, Muntean C, Onea TN, Jivulescu MA, et al. Osteoporosis: a problem still faulty addressed by the Romanian healthcare system. Results of a questionnaire survey of people aged 40 years and over. Front Med (Lausanne). 2024;11:1485382. doi:10.3389/fmed.2024.1485382 PMID: 39507714

28. Shortliffe EH, Chiang MF. Biomedical Data: Their Acquisition, Storage, and Use. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. 5th ed. Cham: Springer; 2021. p. 45-75. doi:10.1007/978-3-030-58721-5_2
29. Melton GB, McDonald CJ, Tang PC, Hripcsak G. Electronic Health Records. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. 5th ed. Cham: Springer; 2021. p. 467-509. doi:10.1007/978-3-030-58721-5_14
30. World Health Organization. International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10), version 2019 [Internet]. Geneva: WHO; 2019 [cited 2026 Oct 8]. Available from: https://icd.who.int/browse10/2019/en
31. Regenstrief Institute. LOINC Users Guide: 2 Major Parts of a LOINC term [Internet]. Indianapolis (IN): Regenstrief Institute; 2026 [cited 2026 Oct 8]. Available from: https://loinc.org/kb/users-guide/major-parts-of-a-loinc-term/
32. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union. 2016 May 4;L 119:1-88. Available from: https://eur-lex.europa.eu/eli/reg/2016/679/oj
33. Legea nr. 190/2018 privind măsuri de punere în aplicare a Regulamentului (UE) 2016/679. Monitorul Oficial al României, Partea I. 2018 Jul 18;nr. 651. Available from: https://legislatie.just.ro/Public/DetaliiDocument/203151
34. Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847. Official Journal of the European Union. 2025 Mar 5;L 2025/327. Available from: https://eur-lex.europa.eu/eli/reg/2025/327/oj
35. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human participants. JAMA. 2025;333(1):71-4. doi:10.1001/jama.2024.21972 PMID: 39425955
36. International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals [Internet]. ICMJE; 2026 [cited 2026 Oct 8]. Available from: https://www.icmje.org/recommendations/
37. Goodman KW, Miller RA. Ethics in Biomedical and Health Informatics: Users, Standards, and Outcomes. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. 5th ed. Cham: Springer; 2021. p. 391-423. doi:10.1007/978-3-030-58721-5_12
38. Baneu P, Prelipcean A, Buda VO, Jianu N, Tudor A, Andor M, et al. Under-Prescription of Drugs in the Elderly Population of Western Romania: An Analysis Based on STOPP/START Version 2 Criteria. J Clin Med. 2024;13(19):5970. doi:10.3390/jcm13195970 PMID: 39408034
39. Marc L, Mihaescu A, Lupusoru R, Schiller O, Bob F, Chisavu L, et al. Hepatitis C and hepatitis B virus infection in hemodialysis patients after nationwide direct antiviral agents therapy-experience of 10 Romanian HD centers. Int Urol Nephrol. 2023;55(11):2951-8. doi:10.1007/s11255-023-03587-0 PMID: 37024632
40. Gal-Nădășan EG, Popescu IM, Bădițoiu LM, Gal-Nădășan N, Cioca F, Popovici ED, et al. Healthcare Workers' Vulnerability to SARS-CoV-2 in Western Romania: A Study on Incidence and Risk Factors for Non-Vaccination and Reinfection. Int J Gen Med. 2023;16:5621-32. doi:10.2147/IJGM.S442098 PMID: 38045906
41. Mihalaș GI, Lungeanu D. Informatică medicală și biostatistică. Timișoara: Editura Victor Babeș; 2009.
42. Kirkwood BR, Sterne JAC. Essential Medical Statistics. 2nd ed. Oxford: Blackwell Science; 2003.
43. Google. Google Sheets function list [Internet]. Mountain View (CA): Google; 2026 [cited 2026 Oct 8]. Available from: https://support.google.com/docs/table/25273?hl=en

44. Google. Create and use pivot tables. Google Docs Editors Help [Internet]. Mountain View (CA): Google; 2026 [cited 2026 Oct 8]. Available from: https://support.google.com/docs/answer/1272900
45. Google. How to set rules for your form. Google Docs Editors Help [Internet]. Mountain View (CA): Google; 2026 [cited 2026 Oct 8]. Available from: https://support.google.com/docs/answer/3378864
46. Google. Choose where to save form responses. Google Docs Editors Help [Internet]. Mountain View (CA): Google; 2026 [cited 2026 Oct 8]. Available from: https://support.google.com/docs/answer/2917686
47. JASP Team. JASP, version 0.98.1 [software]. Amsterdam: University of Amsterdam; 2026. Available from: https://jasp-stats.org/
48. Borelli M. Medical Statistics with JASP: lecture notes for the ICTP Master of Advanced Studies in Medical Physics. 1st ed. Trieste: ICTP; 2022.
49. Goss-Sampson MA. Statistical Analysis in JASP: A Guide for Students [Internet]. JASP; 2018 [cited 2026 Oct 8]. Available from: https://jasp-stats.org/jasp-materials/
50. MedCalc Software Ltd. MedCalc Statistical Software, version 24.0 [software]. Ostend: MedCalc Software Ltd; 2026. Available from: https://www.medcalc.org/
51. Centers for Disease Control and Prevention. Epi Info 7 for Windows, version 7.2.7 [software]. Atlanta (GA): CDC; 2025. Available from: https://www.cdc.gov/epiinfo/index.html
52. Universitatea de Medicină și Farmacie "Victor Babeș" Timișoara. UMFT eLearning platform [Internet]. Universitatea de Medicină și Farmacie "Victor Babeș" Timișoara; [cited 2026 Oct 9]. Available from: https://moodle.umft.ro/
53. Altman DG. Practical Statistics for Medical Research. London: Chapman and Hall; 1991.
54. Frandes M, Deiac AV, Timar B, Lungeanu D. Instrument for assessing mobile technology acceptability in diabetes self-management: a validation and reliability study. Patient Prefer Adherence. 2017;11:259-269. doi:10.2147/PPA.S127922 PMID: 28243069
55. Feier CVI, Santoro RR, Faur AM, Muntean C, Olariu S. Assessing Changes in Colon Cancer Care during the COVID-19 Pandemic: A Four-Year Analysis at a Romanian University Hospital. J Clin Med. 2023;12(20):6558. doi:10.3390/jcm12206558 PMID: 37892695
56. Hopewell S, Chan AW, Collins GS, Hróbjartsson A, Moher D, Schulz KF, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ. 2025;389:e081123. doi:10.1136/bmj-2024-081123 PMID: 40228833
57. Rațiu I, Lupușoru R, Lungeanu D, Popescu A, Sporea I, Goldiș A, et al. Diagnosis of malignant biliary obstruction: pondering over the ERCP, MRCP and histology. J Int Med Res. 2022;50(2):3000605221076924. doi:10.1177/03000605221076924 PMID: 35170356
58. Bende F, Sporea I, Șirli R, Bâldea V, Lazăr A, Lupușoru R, et al. Ultrasound-Guided Attenuation Parameter (UGAP) for the quantification of liver steatosis using the Controlled Attenuation Parameter (CAP) as the reference method. Med Ultrason. 2021;23(1):7-14. doi:10.11152/mu-2688 PMID: 33220028
59. Gherbon A, Frandes M, Dîrpeș D, Timar R, Timar B. Impact of SGLT-2 inhibitors on modifiable cardiovascular risk factors in Romanian patients with type 2 diabetes mellitus. Diabetol Metab Syndr. 2024;16(1):85. doi:10.1186/s13098-024-01326-8 PMID: 38627784
60. Dragomir AC, Buzatu AR, Marin RC, Moleriu RD, Muntean C, Piroș T, et al. Development and Initial Psychometric Validation of the Multidimensional Menopausal Burden Questionnaire (MMBQ-44) in a Sample of Romanian Women. Diagnostics (Basel). 2026;16(16):2574. doi:10.3390/diagnostics16162574 PMID: 42650976
61. Popa DI, Buleu F, Iancu A, Tudor A, Williams CG, Sutoi D, et al. Performance of GFAP and UCH-L1 for Early Acute Stroke Diagnosis in the Emergency Department. J Clin Med. 2025;14(13):4746. doi:10.3390/jcm14134746 PMID: 40649119

62. Popa DI, Buleu F, Williams C, Tudor A, Sutoi D, Trebuian CI, et al. Evaluating Thrombolysis Rates and Emergency Department Time Targets in Acute Ischemic Stroke: Need for Personalized Medicine. J Pers Med. 2024;14(9):955. doi:10.3390/jpm14090955 PMID: 39338208

63. Rosner B. Fundamentals of biostatistics. 8th ed. Boston (MA): Cengage Learning; 2016. ISBN 978-1-305-26892-0.

64. Băicuș C, Bălănescu P. Epidemiologie clinică și biostatistică. București: UMF Carol Davila; 2019.

65. Nica C, Sporea I, Popa A, Lupușoru R, Voicu Stan Erimescu M, Șirli R, et al. The value of Ultrasound Fat Fraction as compared to Controlled Attenuation Parameter for the severity of liver steatosis assessment - preliminary results. Med Ultrason. 2026;28(2):137-44. doi:10.11152/mu-4575 PMID: 41562217

66. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965;52(3-4):591-611. doi:10.1093/biomet/52.3-4.591

67. Frandes M, Gherbon A, Tudor A, Albai O, Vlad MM, Muntean C. TyG-Centered Endocrine-Metabolic Architecture in Patients with Thyroid Dysfunction: A Systems Phenotyping Study. Metabolites. 2026;16(7):490. doi:10.3390/metabo16070490 PMID: 42506441

68. Marin RC, Moleriu RD, Bungau GS, Tit DM, Muntean C. Discordant Immune-Virologic Responses During Antiretroviral Therapy: Immune Dysregulation Patterns, CD4/CD8 Ratio Inversion, and Clinical Predictors in a Romanian HIV Cohort. Viruses. 2026;18(5):512. doi:10.3390/v18050512 PMID: 42198715

69. Gherbon A, Roman D, Nicula-Neagu M, Frandes M. Demographic Vulnerability and Mental Health in Older Adults with Type 2 Diabetes: A Lifespan-Informed Analysis of Diabetes Distress. Diseases. 2026;14(9):332. doi:10.3390/diseases14090332 PMID: 42783225

70. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337-50. doi:10.1007/s10654-016-0149-3 PMID: 27209009

71. Moleriu LC, Lupusoru R, Muntean C, Piroș T, Popescu A, Sirli R, et al. Dietary Behaviors, Digestive Symptoms, and Neurovegetative Features in Disorders of Gut-Brain Interaction: A Cross-Sectional Clinical Study. Nutrients. 2026;18(7):1023. doi:10.3390/nu18071023 PMID: 41978074

72. Cirstea N, Tit DM, Purza AL, Toma MM, Radu A, Bungau GS, et al. Associations Between Patient Satisfaction with Pharmacist Services, Social Support, and Medication Adherence in Chronic Dermatological Conditions: A Cross-Sectional Study. Healthcare (Basel). 2026;14(15):2346. doi:10.3390/healthcare14152346 PMID: 42588315

73. Dicu AM, Marin RC, Muntean C, Moleriu RD, Tudor A, Piroș T, et al. Development and Validation of a Multidimensional Consumer Assessment Framework for Alternative-Ingredient Bakery and Confectionery Products: Implications for Nutrition Methodology and Sustainable Diet Research. Foods. 2026;15(17):2950. doi:10.3390/foods15172950 PMID: 42737195

74. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. Am Stat. 2016;70(2):129-33. doi:10.1080/00031305.2016.1154108

75. Albai O, Frandes M, Sima A, Timar B, Vlad A, Timar R. Practical Applicability of the ISARIC-4C Score on Severity and Mortality due to SARS-CoV-2 Infection in Patients with Type 2 Diabetes. Medicina (Kaunas). 2022;58(7):848. doi:10.3390/medicina58070848 PMID: 35888567

76. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale (NJ): Lawrence Erlbaum; 1988.

77. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007;39(2):175-91. doi:10.3758/bf03193146 PMID: 17695343

78. Popovici SE, Ritiu SA, Miutescu B, Moga TV, Sporea I, Sandesc D, et al. Entropy-Guided Sedation Is Associated with Improved Hemodynamic Stability and Recovery During ERCP: A

Prospective Real-World Cohort Study. J Clin Med. 2026;15(10):3665. doi:10.3390/jcm15103665 PMID: 42194633

79. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995;311(7003):485. doi:10.1136/bmj.311.7003.485 PMID: 7647644
80. American Diabetes Association Professional Practice Committee for Diabetes. 2. Diagnosis and classification of diabetes: Standards of Care in Diabetes-2026. Diabetes Care. 2026;49(Suppl 1):S27-49. doi:10.2337/dc26-S002 PMID: 41358893
81. Al Barri L, Yasar II, Mercea N, Tudor A, Stanca HT, Roşca C, et al. Interdevice Agreement of Keratometry, Astigmatism Vectors, and Ocular Biometry in Cataract Candidates: SS-OCT (Argos) vs. OLCI (Aladdin) vs. Scheimpflug-Placido (Sirius). Bioengineering (Basel). 2026;13(3):296. doi:10.3390/bioengineering13030296 PMID: 41899827
82. Padurean LL, Frandes M, Stefanescu HE, Muntean C, Horhat ID. Impact of Smart Hearing Aid Technology on Self-Esteem in Patients with Sensorineural Hearing Loss: A Cross-Sectional Multivariate Study. Healthcare (Basel). 2026;14(10):1336. doi:10.3390/healthcare14101336 PMID: 42194428
83. Student. The probable error of a mean. Biometrika. 1908;6(1):1-25. doi:10.2307/2331554
84. Ursoiu F, Moleriu L, Lungeanu D, Puschită M. The association between HIV clinical disease severity and psychiatric disorders as seen in Western Romania. AIDS Care. 2018;30(11):1368-1371. doi:10.1080/09540121.2018.1455959 PMID: 29592527
85. Gherbon A, Frandes M, Timar R, Nicula M. Beneficial effects of Aloe ferox on lipid profile, blood pressure, and glycemic control in obese persons: A CONSORT-clinical study. Medicine (Baltimore). 2021;100(50):e28336. doi:10.1097/MD.0000000000028336 PMID: 34918714
86. Buda V, Andor M, Cristescu C, Tomescu MC, Muntean DM, Bâibâță DE, et al. Thrombospondin-1 Serum Levels In Hypertensive Patients With Endothelial Dysfunction After One Year Of Treatment With Perindopril. Drug Des Devel Ther. 2019;13:3515-3526. doi:10.2147/DDDT.S218428 PMID: 31631975
87. Buleu FN, Luca CT, Tudor A, Badalica-Petrescu M, Caraba A, Pah A, et al. Correlations between Vascular Stiffness Indicators, OPG, and 25-OH Vitamin D3 Status in Heart Failure Patients. Medicina (Kaunas). 2019;55(6):309. doi:10.3390/medicina55060309 PMID: 31242663
88. Turi V, Iurciuc S, Crețu OM, Tit DM, Bungau S, Apostol A, et al. Arterial function in hypertensive pregnant women. Is arterial stiffness a marker for the outcomes in pregnancy?. Life Sci. 2021;264:118723. doi:10.1016/j.lfs.2020.118723 PMID: 33160988
89. Uivarosan D, Tit DM, Iovan C, Nistor-Cseppento DC, Endres L, Lazar L, et al. Effects of combining modern recovery techniques with neurotrophic medication and standard treatment in stroke patients. Sci Total Environ. 2019;679:80-87. doi:10.1016/j.scitotenv.2019.05.070 PMID: 31100563
90. Stoichescu-Hogea G, Buleu FN, Christodorescu R, Sosdean R, Tudor A, Ember A, et al. Contribution of Global and Regional Longitudinal Strain for Clinical Assessment of HFpEF in Coronary and Hypertensive Patients. Medicina (Kaunas). 2021;57(12):1372. doi:10.3390/medicina57121372 PMID: 34946317
91. Miulescu RG, Roşca I, Marin RC, Muntean C, Pavel AN, Stoleru S, et al. Clinical Outcomes Following Personalized Gut Microbiota-Targeted Therapy in Children with Atopic Dermatitis and Laboratory-Confirmed Dysbiosis: A Single-Arm Prospective Pilot Study. Nutrients. 2026;18(15):2572. doi:10.3390/nu18152572 PMID: 42588195
92. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat. 1947;18(1):50-60. doi:10.1214/aoms/1177730491
93. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bull. 1945;1(6):80-3. doi:10.2307/3001968

94. Munteanu LA, Frandes M, Timar B, Tudorache E, Fildan AP, Oancea C, et al. The efficacy of a mobile phone application to improve adherence to treatment and self-management in people with chronic respiratory disease in Romanian population - a pilot study. BMC Health Serv Res. 2020;20(1):475. doi:10.1186/s12913-020-05340-0 PMID: 32460752

95. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc. 1952;47(260):583-621. doi:10.1080/01621459.1952.10483441

96. Hasan EM, Calma CL, Tudor A, Oancea C, Tudorache V, Petrache IA, et al. Coping, Anxiety, and Pain Intensity in Patients Requiring Thoracic Surgery. J Pers Med. 2021;11(11):1221. doi:10.3390/jpm11111221 PMID: 34834573

97. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc. 1937;32(200):675-701. doi:10.1080/01621459.1937.10503522

98. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Lond Edinb Dublin Philos Mag J Sci. 1900;50(302):157-75. doi:10.1080/14786440009463897

99. Marin RC, Streinu-Cercel A, Moleriu LC, Bungau SG. Analysis of virological response to therapy and resistance profile in treatment-experienced and naive HIV-1 infected Romanian patients receiving regimens containing darunavir boosted with ritonavir or cobicistat. Biomed Pharmacother. 2022;150:113077. doi:10.1016/j.biopha.2022.113077 PMID: 35658217

100. Popa DI, Buleu F, Iancu A, Tudor A, Williams CG, Militaru M, et al. Long COVID and Acute Stroke in the Emergency Department: An Analysis of Presentation, Reperfusion Treatment, and Early Outcomes. J Clin Med. 2025;14(18):6514. doi:10.3390/jcm14186514 PMID: 41010719

101. Iovanescu D, Frandes M, Lungeanu D, Burlea A, Miutescu BP, Miutescu E. Diagnosis reliability of combined flexible sigmoidoscopy and fecal-immunochemical test in colorectal neoplasia screening. Onco Targets Ther. 2016;9:6819-6828. doi:10.2147/OTT.S122425 PMID: 27853376

102. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307-10. PMID: 2868172

103. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74. PMID: 843571

104. Masood M, Masood Y, Newton JT. The clustering effects of surfaces within the tooth and teeth within individuals. J Dent Res. 2015;94(2):281-8. doi:10.1177/0022034514559408 PMID: 25421840

105. Danila AI, Cioca F, Gadde ST, Daruvuri SP, Timar R, Hogea E. Prognostic Utility of dNLR, ALRI, APRI, and SII in COVID-19 Patients with Diabetes: A Cross-Sectional Study. Diagnostics (Basel). 2024;14(15):1685. doi:10.3390/diagnostics14151685 PMID: 39125561

106. Ion R, Kimidi JSSK, Kalapala C, Fnu O, Chandrababu VR, Desireddygari OR, et al. Evaluating the Utility of Clinical Scores APACHE, CURB, SOFA, and NEWS2 at Admission and 5-Days after Symptom Onset in Predicting Severe COVID-19 in Patients with Diabetes. J Pers Med. 2024;14(8):868. doi:10.3390/jpm14080868 PMID: 39202059

107. Bâldea V, Sporea I, Tudor A, Popescu A, Bende F, Șirli R. Virtual Touch Quantification using Acoustic Radiation Force Impulse Imaging Technology versus Transient Elastography for the Noninvasive Assessment of Liver Fibrosis in Patients with Chronic Hepatitis B or C using Liver Biopsy as the Gold Standard. J Gastrointestin Liver Dis. 2020;29(2):181-190. doi:10.15403/jgld-2256 PMID: 32530985

108. Hill AB. The environment and disease: association or causation?. Proc R Soc Med. 1965;58(5):295-300. doi:10.1177/003591576505800503 PMID: 14283879

109. Mateoc-Sîrb T, Suciu IM, Gaiță D, Minodora A, Popescu R, Vlad T, et al. Guideline-Directed Medical Therapy Intensity, Ventricular Remodeling, and Clinical Outcomes After Acute

Myocardial Infarction: A Single-Center Real-World Retrospective Cohort Study. Biomedicines. 2026;14(5):1067. doi:10.3390/biomedicines14051067 PMID: 42193392

110. Fofiu R, Bende F, Popescu A, Şirli R, Lupuşoru R, Ghiuchici AM, et al. Spleen and Liver Stiffness for Predicting High-Risk Varices in Patients with Compensated Liver Cirrhosis. Ultrasound Med Biol. 2021;47(1):76-83. doi:10.1016/j.ultrasmedbio.2020.09.004 PMID: 33067019

111. Wassertheil-Smoller S, Smoller J. Biostatistics and Epidemiology: A Primer for Health and Biomedical Professionals. 5th ed. Cham: Springer; 2024. ISBN 978-3-031-53042-5.

112. Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. Lancet. 2002;359(9303):341-5. doi:10.1016/S0140-6736(02)07500-1 PMID: 11830217

113. Schulz KF, Grimes DA. Case-control studies: research in reverse. Lancet. 2002;359(9304):431-4. doi:10.1016/S0140-6736(02)07605-5 PMID: 11844534

114. Pirtea L, Grigoraş D, Matusz P, Pirtea M, Moleriu L, Tudor A, et al. Age and HPV type as risk factors for HPV persistence after loop excision in patients with high grade cervical lesions: an observational study. BMC Surg. 2016;16(1):70. doi:10.1186/s12893-016-0185-7 PMID: 27716233

115. Pantea-Roşan LR, Pantea VA, Bungau S, Tit DM, Behl T, Vesa CM, et al. No-Reflow after PPCI-A Predictor of Short-Term Outcomes in STEMI Patients. J Clin Med. 2020;9(9):2956. doi:10.3390/jcm9092956 PMID: 32932736

116. Feier CVI, Ardelean MV, Gaborean V, Muntean C, Faur AM, Rus VI, et al. Diagnostic and Prognostic Value of AISI, SII, and SIRI in Predicting Gangrenous Evolution of Acute Lithiasic Cholecystitis. Diagnostics (Basel). 2026;16(3):441. doi:10.3390/diagnostics16030441 PMID: 41681759

117. Petrusan AM, Feier CVI, Muntean C, Gaborean V, Petrusan AS, Nicoara D, et al. Association of BRCA Mutation Status with Clinical Outcomes in High-Grade Serous Ovarian Cancer. Healthcare (Basel). 2026;14(9):1193. doi:10.3390/healthcare14091193 PMID: 42121636

118. Gherbon A, Frandes M, Toderescu CD, Dirpes D, Timar R, Nicula MN, et al. Maternal and Fetal-Neonatal Complications of Romanian Women with Gestational Diabetes: A Retrospective Comparative Study. Medicina (Kaunas). 2025;61(7):1190. doi:10.3390/medicina61071190 PMID: 40731821

119. Alexandru GC, Muntean C, Chioran D, Riviş M, Gligor LN, Pricop MO, et al. Epidemiological Assessment of Oral and Oropharyngeal Squamous Cell Carcinomas Before, During, and After the COVID-19 Pandemic, 2018-2024. J Clin Med. 2026;15(18):6963. doi:10.3390/jcm15186963 PMID: 42795735

120. Owens DK, Goldhaber-Fiebert JD, Sox HC. Biomedical Decision Making: Probabilistic Clinical Reasoning. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. 5th ed. Cham: Springer; 2021. p. 77-120. doi:10.1007/978-3-030-58721-5_3

121. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ. 2004;329(7458):168-9. doi:10.1136/bmj.329.7458.168 PMID: 15258077

122. Fagan TJ. Nomogram for Bayes's theorem. N Engl J Med. 1975;293(5):257. doi:10.1056/NEJM197507312930513 PMID: 1143310

123. Barbulescu A, Ratiu I, Sporea I, Lungeanu D, Lupusoru R, Miutescu B, et al. Extending and validating the Baveno VI criteria for the exclusion of high-risk varices. Med Ultrason. 2021;23(3):265-70. doi:10.11152/mu-2929 PMID: 33657192

124. Musen MA, Middleton B, Greenes RA. Clinical Decision-Support Systems. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. 5th ed. Cham: Springer; 2021. p. 795-840. doi:10.1007/978-3-030-58721-5_24

125. Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care. 2003;26(3):725-31. doi:10.2337/diacare.26.3.725 PMID: 12610029

126. Sipos SI, Vlad DC, Enatescu VR, Moleriu RD, Petre I, Balasan CG, et al. The Role of IL-6 in the Diagnosis of Neonatal Sepsis and Its Influence on Maternal Mental Health. Cureus. 2025;17(3):e80693. doi:10.7759/cureus.80693 PMID: 40242671

127. Pavel AL, Kundnani NR, Morariu SI, Tudor A, Man DE, Duda-Seiman DM, et al. Importance of H-FABP in Early Diagnosis of Acute Myocardial Infarction. Int J Gen Med. 2024;17:4335-4346. doi:10.2147/IJGM.S476736 PMID: 39346631

128. Guyatt G, Agoritsas T, Brignardello-Petersen R, Mustafa RA, Rylance J, Foroutan F, et al. Core GRADE 1: overview of the Core GRADE approach. BMJ. 2025;389:e081903. doi:10.1136/bmj-2024-081903

129. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71 PMID: 33782057

130. Suciu BS, Turi VR, Crisan S, Luca CT, Lazar DC, Faur-Grigori AA, et al. Heart Failure with Preserved Ejection Fraction (HFpEF), Pulse Wave Velocity, and Heart Rate Recovery Interconnections-A Brief Literature Review. J Clin Med. 2025;14(24):8781. doi:10.3390/jcm14248781 PMID: 41464682

131. Moleriu LC, Sas I, RaȚiu AC, Margan MM, Secoşan CA, Ilina RȘ, et al. HPV genotyping and p16÷Ki-67 dual staining in the detection of high-grade cervical lesion in patients with LSIL on Pap smear. Rom J Morphol Embryol. 2017;58(2):433-437. PMID: 28730227

132. Golu I, Sporea I, Moleriu L, Tudor A, Cornianu M, Vlad A, et al. 2D-Shear Wave Elastography in the Evaluation of Parathyroid Lesions in Patients with Hyperparathyroidism. Int J Endocrinol. 2017;2017:9092120. doi:10.1155/2017/9092120 PMID: 28845158

133. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ. 1994;309(6947):102. doi:10.1136/bmj.309.6947.102 PMID: 8038641

134. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32-5. doi:10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3 PMID: 15405679

135. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ. 1994;308(6943):1552. doi:10.1136/bmj.308.6943.1552 PMID: 8019315

136. Monaghan TF, Rahman SN, Agudelo CW, Wein AJ, Lazar JM, Everaert K, et al. Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. Medicina (Kaunas). 2021;57(5):503. doi:10.3390/medicina57050503

137. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36. doi:10.1148/radiology.143.1.7063747 PMID: 7063747

138. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. BMJ. 1994;309(6948):188. doi:10.1136/bmj.309.6948.188 PMID: 8044101

139. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378

140. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527. doi:10.1136/bmj.h5527 PMID: 26511519

141. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. 2025;31(10):3283-9. doi:10.1038/s41591-025-03953-8

142. Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. Lancet. 2002;359(9307):696-700. doi:10.1016/S0140-6736(02)07816-9 PMID: 11879884

143. Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC (Text with EEA relevance). OJ L 158, 27.5.2014, p. 1-76. Available from: https://eur-lex.europa.eu/eli/reg/2014/536/oj

144. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH harmonised guideline: guideline for good clinical practice E6(R3). Final version, adopted on 6 January 2025 [Internet]. ICH; 2025 [cited 2026 Oct 9]. Available from: https://database.ich.org/sites/default/files/ICH_E6%28R3%29_Step4_FinalGuideline_2025_0106.pdf

145. Putnoky S, Banu AM, Moleriu LC, Putnoky S, Șerban DM, Niculescu MD, et al. Reliability and validity of a General Nutrition Knowledge Questionnaire for adults in a Romanian population. Eur J Clin Nutr. 2020;74(11):1576-1584. doi:10.1038/s41430-020-0616-5 PMID: 32235889

146. Serban DM, Serban CL, Ursoniu S, Putnoky S, Moleriu RD, Putnoky S. Mindful Eating Questionnaire: Validation and Reliability in Romanian Adults. Int J Environ Res Public Health. 2022;19(17):10517. doi:10.3390/ijerph191710517 PMID: 36078231

147. Google. View and manage form responses. Google Docs Editors Help [Internet]. Mountain View (CA): Google; 2026 [cited 2026 Oct 8]. Available from: https://support.google.com/docs/answer/139706

148. Google. Show questions based on answers. Google Docs Editors Help [Internet]. Mountain View (CA): Google; 2026 [cited 2026 Oct 8]. Available from: https://support.google.com/docs/answer/141062

149. Bolboacă SD, Drugan T. Aplicații practice de biostatistică. Cluj-Napoca: Editura Medicală Universitară "Iuliu Haţieganu"; 2023. ISBN 978-606-075-206-6 (PDF). Coordinator role ('coord.') could not be confirmed from a primary source; both sources list them as authors

150. Iancu M, Bolboacă SD. Informatică medicală și biostatistică: noțiuni fundamentale pentru profesioniștii în științele sănătății. Cluj-Napoca: Editura Medicală Universitară "Iuliu Haţieganu"; 2025. ISBN 978-606-075-342-1 (PDF).

151. Țarcă V. Noțiuni fundamentale de biostatistică. Iași: Junimea; 2025. ISBN 978-973-37-2931-0.