„VICTOR BABEŞ" UNIVERSITY
OF MEDICINE AND PHARMACY
FROM TIMIŞOARA

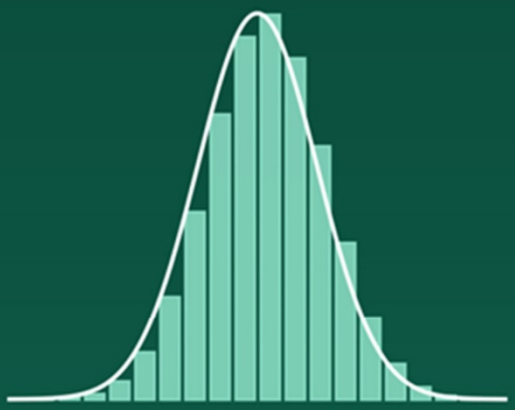
CĂLIN MUNTEAN

APPLIED BIOSTATISTICS

IN GENERAL MEDICINE

Data, Methods, Analysis and Decision. Horizon 2026

Lectures and practical sessions for medical students





ROC

95% CI	128.96-133.30
RR	2.62
AUC	0.85

MANUALE

„Victor Babeş" Publishing House
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 978-606-786-653-7

The patients in the examples are fictitious. The teaching dataset "mg2026" was generated by simulation and does not come from patient records. Figures taken from published articles are reproduced as they appear in the abstracts indexed in PubMed. Information on software, versions and regulations reflects the situation as of October 9, 2026. This book does not replace clinical guidelines or the regulations in force, which must be consulted in their current form.

Contents

Preface 11

Acknowledgments 12

How to Read This Book 12

List of Abbreviations 14

Part I. Lectures 16

Chapter 1. Introduction to Biostatistics: Fundamental Concepts, Role in Medical Research and Elements of Probability 17

1.1 Why a Physician Needs Biostatistics 17

1.2 Basic Concepts: Statistical Unit, Population, Sample 19

1.3 The mg2026 Teaching Dataset 20

1.4 Sampling: Representativeness and Methods 21

1.5 Systematic Errors (Bias) 23

1.6 Descriptive Statistics and Inferential Statistics 25

1.7 The Stages of a Statistical Study and the Research Question 27

1.8 Data in the Era of the Electronic Health Record: The Link with Medical Informatics 28

1.9 Elements of Probability Theory 31

1.10 Software Used in This Course 36

1.11 Worked Problems 37

1.12 Proposed Exercises 41

Chapter 2. Describing and Organizing Data: Types of Variables, Descriptive Measures and Graphical Displays 45

2.1 From the Patient to the Data Table 45

2.2 Frequency Tables 49

2.3 Measures of Central Tendency 52

2.4 Measures of Dispersion 54

2.5 Quantiles, Percentiles, Quartiles and the Five-Number Summary 57

2.6 The Shape of the Distribution: A First Look at Skewness and Kurtosis 58

2.7 Graphical Displays 58

2.8 The Descriptive Table of an Article: "Table 1" 62

2.9 Complete Descriptive Analysis of the mg2026 Dataset 63

2.10 Worked Problems 66

2.11 Proposed Exercises 70

Chapter 3. Statistical Distributions: The Normal Distribution, Checking Normality, Skewness, Kurtosis, the z-Score and Data Transformation 74

3.1 Random Variables and Probability Distributions 74

3.2 Two Discrete Distributions Useful in Medicine 75

3.3 The Normal Distribution 77
3.4 The Standard Normal Distribution and the z-Score 80
3.5 Reference Intervals in Laboratory Medicine 82
3.6 Sampling Distributions: t, Chi-Square and F 83
3.7 The Central Limit Theorem 85
3.8 Checking Normality 86
3.9 Data Transformation 91
3.10 Extreme Values 93
3.11 Overlapping Distributions: From the z Score to Clinical Decisions 95
3.12 Worked Problems 96
3.13 Proposed Exercises 99
Chapter 4. Parameter Estimation and Confidence Intervals; Principles of Sample Size Calculation 103
4.1 Point Estimation and Interval Estimation 103
4.2 The Sampling Distribution of the Mean and the Standard Error 104
4.3 The Interval for Individual Values and the Interval for the Mean 105
4.4 The Confidence Interval for the Mean Using the t Distribution 106
4.5 Correct Interpretation of a Confidence Interval 108
4.6 Width of the Confidence Interval and Sample Size 109
4.7 The Confidence Interval for a Proportion 110
4.8 The Confidence Interval for the Difference between Two Means and between Two Proportions 112
4.9 Confidence Intervals for Skewed Data 113
4.10 Confidence Intervals for Ratios (OR, RR) and for the ROC Area 114
4.11 The Bootstrap, Explained Simply 115
4.12 The Confidence Interval and the p Value 116
4.13 Principles of Sample Size Calculation 117
4.14 Worked Problems 120
4.15 Proposed Exercises 122
Chapter 5. Foundations of Statistical Testing: Hypotheses, Significance Level, Type I and II Errors, Decision and Interpretation 127
5.1 From the Clinical Question to a Testable Hypothesis 127
5.2 The Logic of the Test: The Test Statistic and Its Distribution under H0 129
5.3 The p Value 130
5.4 The Significance Level α and the Decision Rule 131
5.5 Type I and II Errors, Statistical Power 132
5.6 Effect Size and Clinical Significance 136

5.7 Three Routes to the Same Decision: The Critical Value, the p Value and the Confidence Interval ... 138
5.8 The Algorithm of a Statistical Test ... 139
5.9 The z and t Tests for a Single Mean ... 139
5.10 Multiple Comparisons and Inflation of the Type I Error ... 141
5.11 p-Hacking, HARKing and Preregistration ... 144
5.12 Equivalence and Non-Inferiority, in Brief ... 145
5.13 How to Report a Statistical Test ... 145
5.14 Worked Problems ... 146
5.15 Proposed Exercises ... 149
Chapter 6. Common Parametric Statistical Tests: The t-Test for Paired and Independent Samples, Analysis of Variance (ANOVA) ... 153
6.1 The Family of Parametric Tests ... 153
6.2 Comparing a Mean with a Given Value: The z-Test and the t-Test ... 154
6.3 The Paired-Samples t-Test ... 154
6.4 The Independent-Samples t-Test ... 157
6.5 One-Way Analysis of Variance (ANOVA) ... 160
6.6 Post Hoc Tests ... 164
6.7 Repeated-Measures ANOVA ... 166
6.8 Two-Way ANOVA and ANCOVA, Briefly ... 167
6.9 Reporting t-Tests and ANOVA ... 167
6.10 Worked Problems ... 168
6.11 Proposed Exercises ... 171
Chapter 7. Common Nonparametric Statistical Tests: Mann-Whitney U, Wilcoxon, Kruskal-Wallis, Friedman and the Chi-Square Test ... 175
7.1 When We Can No Longer Use Parametric Tests ... 175
7.2 Ranks: The Raw Material of Nonparametric Tests ... 177
7.3 The Mann-Whitney U Test ... 178
7.4 The Wilcoxon Test for Paired Data and the Sign Test ... 181
7.5 The Kruskal-Wallis Test ... 183
7.6 The Friedman Test ... 186
7.7 The Chi-Square Test for Contingency Tables ... 188
7.8 Variants of the Chi-Square Test ... 193
7.9 Beyond the p Value: Phi, Cramér's V and Other Measures of Association ... 195
7.10 Reporting the Results of Nonparametric Tests and the Chi-Square Test ... 196
7.11 Worked Problems ... 196
7.12 Proposed Exercises ... 201

Chapter 8. Selecting and Applying Statistical Tests: Selection Criteria, Problem Solving and Case Studies 205

8.1 Choosing the Test Is a Design Decision, Not a Calculation Decision 205

8.2 The Algorithm for Choosing the Test 206

8.3 Summary Table of the Tests 208

8.4 The Unit of Analysis and Independence of Observations 210

8.5 The Test Is Specified Before Seeing the Data 211

8.6 Common Mistakes in Choosing and Applying the Test 211

8.7 Reporting a Result in the Standardized Format (a-g) 213

8.8 Worked Problems 214

8.9 Recognizing the Wrong Test: Ten "Methods" Descriptions 222

8.10 Critical Reading of the Methods Section: Articles by Colleagues in the Discipline 224

8.11 Proposed Exercises 229

Chapter 9. Analyzing Relationships between Variables: Correlation and Linear Regression for Numeric and Ordinal Variables 233

9.1 Bivariate Analysis: Two Variables Measured on the Same Patient 233

9.2 Covariance: The First Step Toward a Measure of Association 235

9.3 The Pearson Correlation Coefficient 235

9.4 Correlation for Ordinal or Skewed Data: The Spearman and Kendall Coefficients 240

9.5 Correlation Does Not Mean Causation 242

9.6 Correlation and Agreement: The Bland-Altman Method 244

9.7 Simple Linear Regression 244

9.8 Multiple Linear Regression: An Introduction 249

9.9 Logistic Regression: A Bridge to Chapter 10 252

9.10 Worked Problems 253

9.11 Proposed Exercises 256

Chapter 10. Statistical Applications in Epidemiology: Types of Observational Studies and Risk Analysis 260

10.1 The Risk Factor 260

10.2 Measuring Disease Frequency 261

10.3 Types of Epidemiological Studies 262

10.4 The 2 x 2 Contingency Table 266

10.5 The Relative Risk 267

10.6 The Odds Ratio 267

10.7 Confidence Intervals and Statistical Testing 269

10.8 Measures of Impact: Attributable Risk, Attributable Fraction, NNT and NNH 270

10.9 Confounding, Stratification and Effect Modification 272

10.10 From Association to Causation 274

10.11 Worked Problems 276
10.12 Proposed Exercises 280
Chapter 11. Statistical Methods in Medical Decision-Making: Bayesian Analysis and Evidence-Based Medicine 284
11.1 Uncertainty in Clinical Decision-Making 284
11.2 Conditional Probability, Revisited 285
11.3 Bayes' Theorem 287
11.4 Odds, Likelihood Ratios and the Odds Form of Bayes' Theorem 289
11.5 Sequential Testing and Combining Tests 293
11.6 Differential Diagnosis with Several Diseases 295
11.7 Decision Thresholds: To Test, to Treat 296
11.8 Screening in Family Medicine: The FINDRISC Example 297
11.9 Bayesian Inference Versus Frequentist Inference 299
11.10 Evidence-Based Medicine 301
11.11 Cognitive Errors in Clinical Reasoning 303
11.12 Communicating Risk to the Patient 303
11.13 Worked Problems 304
11.14 Proposed Exercises 308
Chapter 12. Evaluating the Performance of Diagnostic Tests: Sensitivity, Specificity, Predictive Values and the ROC Curve 312
12.1 The Index Test and the Reference Standard 312
12.2 The 2 x 2 Table and the Performance Indicators 313
12.3 Reconstructing a 2 x 2 Table 316
12.4 The Dependence of Predictive Values on Prevalence 317
12.5 Bias in Diagnostic Accuracy Studies 318
12.6 Tests with a Continuous Result and the Choice of Cutoff 319
12.7 The ROC Curve 320
12.8 Interrater Agreement 325
12.9 Evaluating Artificial Intelligence Classifiers 326
12.10 Reporting Diagnostic Accuracy Studies: STARD 2015 and STARD-AI 326
12.11 Worked Problems 327
12.12 Proposed Exercises 332
Chapter 13. Clinical Studies: Classification, Design, Protocol, Data Collection Instruments and Integrated Analysis of Biomedical Data 336
13.1 What a Clinical Study Is and What It Requires of Statistics 336
13.2 Classification of Clinical Studies 337
13.3 The Methodological Principles of a Randomized Trial 341
13.4 The Protocol, Registration and the Ethical Framework 345

13.5 Data Collection Instruments 346
13.6 Data Management and Data Cleaning 351
13.7 Reporting: CONSORT 2025 352
13.8 Integrated Analysis of the mg2026 Study: A Complete Example 353
13.9 Worked Problems 359
13.10 Proposed Exercises 363
Chapter 14. Review Examples: Integrated Interpretation of Statistical Results 367
14.1 The Map of the Discipline 367
14.2 How to Read the Statistics of a Medical Article 369
14.3 Integrated Case Studies 371
14.4 How to Write the Results Paragraph 382
14.5 Forty Statements: True or False 383
14.6 Summary Table of Formulas 385
14.7 Worked Problems 387
14.8 Proposed Exercises 389
Part II. Practical Sessions 396
Practical Session 1. Introduction to Biostatistics and Digital Tools: Google Sheets, JASP and MedCalc 397
Practical Session 2. Describing and Organizing Data: The Medical Questionnaire, Frequency Tables and Graphs 405
Practical Session 3. Distributions and Normality Testing 413
Practical Session 4. Estimation and Confidence Intervals: The Effect of Sample Size 420
Practical Session 5. Statistical Hypothesis Testing: Formulating Hypotheses, Errors and Decisions 426
Practical Session 6. Parametric Tests: The t-Test and ANOVA 434
Practical Session 7. Nonparametric Tests: Mann-Whitney, Wilcoxon, Kruskal-Wallis, Friedman and Chi-Square 441
Practical Session 8. Choosing the Statistical Test and Standardized Reporting of Results 448
Practical Session 9. Correlation and Regression Analysis: Pearson, Spearman and Simple Linear Regression 454
Practical Session 10. Epidemiological Analyses: Relative Risk, Odds Ratio and the Interpretation of Observational Studies 460
Practical Session 11. Bayesian Analysis: Probabilistic Reasoning in Diagnosis and Screening 467
Practical Session 12. Performance of Diagnostic Tests: Sensitivity, Specificity, Predictive Values and the ROC Curve 473
Practical Session 13. Designing a Mini Clinical Study and the Integrated Analysis of a Complete Dataset 479
Practical Session 14. The Practical Exam: Sample Paper, Grading Scheme and Preparation Plan 486

Appendix A. Answers to the Self-Assessment Questions and Proposed Exercises 496
Chapter 1 496
Chapter 2 497
Chapter 3 498
Chapter 4 498
Chapter 5 499
Chapter 6 500
Chapter 7 501
Chapter 8 502
Chapter 9 505
Chapter 10 505
Chapter 11 506
Chapter 12 507
Chapter 13 508
Chapter 14 509
Appendix B. Glossary 511
Appendix C. The mg2026 Teaching Dataset: Scenario, Data Dictionary and the Errors in the Raw File 532
C.1 The Scenario 532
C.2 The Files 532
C.3 The Data Dictionary 533
C.4 The Errors in the Raw File and How to Correct Them 535
C.5 How the Dataset Was Generated 537
C.6 How to Cite the Dataset in Reports and Assignments 537
Appendix D. Formula Sheet 539
D.1 Chapter 1. Probability 539
D.2 Chapter 2. Describing Data 540
D.3 Chapter 3. The Normal Distribution, the z-Score, Transformations 541
D.4 Chapter 4. Estimation, Confidence Intervals, Sample Size 542
D.5 Chapter 5. Fundamentals of Hypothesis Testing 543
D.6 Chapter 6. Parametric Tests: t and ANOVA 543
D.7 Chapter 7. Nonparametric Tests and the Chi-Square Test 545
D.8 Chapter 8. Choosing the Test (Numerical Benchmarks) 546
D.9 Chapter 9. Correlation and Regression 546
D.10 Chapter 10. Epidemiology and Risk Analysis 547
D.11 Chapter 11. Bayes and Evidence-Based Medicine 548
D.12 Chapter 12. Diagnostic Tests and the ROC Curve 548

D.13 Chapter 13. Clinical Trials (Planning Formulas) 549
Appendix E. Statistical Tables 551
E.1 The Standard Normal Distribution: The Cumulative Distribution Function Φ(z) 551
E.2 Critical Values of the Standard Normal Distribution 552
E.3 Critical Values of the Student (t) Distribution 552
E.4 Critical Values of the Chi-Square (χ^2) Distribution 553
E.5 Critical Values of the Fisher (F) Distribution for $\alpha = 0.05$ 554
E.6 Critical Values of the Pearson Correlation Coefficient (r) 555
E.7 Approximate Critical Values of the Spearman Correlation Coefficient (rho) 556
E.8 Table of 1,000 Random Digits 557
Appendix F. Practice Multiple-Choice Tests for the Exam 559
How the Questions Are Built and How They Are Scored 559
Test 1 559
Test 2 567
Test 3 576
Answers and Explanations 583
References 597

Preface

A family doctor receives the results of forty laboratory tests in the morning, reads a new hypertension guideline at lunch and, in the evening, an article about an antidiabetic drug. At each of these moments the doctor is doing statistics without realizing it. The doctor wonders whether a blood glucose of 118 mg/dL is a chance deviation or a signal, whether the drop in blood pressure reported in the study would matter for their own patients, and how far a risk score filled in within five minutes can be trusted. Biostatistics is not a discipline of calculations. It is the way medicine measures its uncertainty.

This book is the Biostatistics course for first-year students of the Faculty of Medicine of the "Victor Babeș" University of Medicine and Pharmacy of Timișoara. It continues the medical informatics textbook of the first semester [1], keeping its format and working rules, and follows the course syllabus for the 2026-2027 academic year: 14 lecture topics and 14 practical sessions, one pair per week, with a written multiple-choice exam in the format of the Romanian residency examination and a practical exam in the laboratory. I chose the title "Applied Biostatistics in General Medicine" because all the examples come from the medicine that every graduate will practice, whatever specialty they choose later: family practice, internal medicine, diabetes and metabolic diseases, hypertension, respiratory diseases, hepatology, cardiology, pediatrics, obstetrics, emergency medicine and infectious diseases.

The first part of the course, Chapters 1-4, lays the foundation: what data are and where they come from, how they are described, what a distribution is, and how a population parameter is estimated from a sample. Chapter 1 also links back to the medical informatics of the first semester, because a statistical analysis can be no better than the data it processes, and health data require ethical and protection rules that every physician must know. Chapters 5-8 cover hypothesis testing: the logic of a test, the usual parametric and nonparametric tests and, above all, how to choose the right test, the most frequent source of errors both in the exam and in published articles. Chapters 9-12 deal with association and decision: correlation and regression, the measures of risk used in epidemiology, Bayesian reasoning and the evaluation of diagnostic tests. Chapter 13 shows how to design a clinical study and how to build a data collection form, and Chapter 14 brings everything together in examples of integrated interpretation.

Biostatistics is hard to learn from definitions and much easier to learn from examples the reader recognizes. For this reason the book draws on two sources of examples. The first is the teaching dataset "mg2026", a simulated randomized trial in family practices, with 160 patients at cardiometabolic risk followed for six months. The same dataset appears in every chapter and in every practical session, so that students can recalculate the same mean, the same confidence interval and the same test with the free software available in the laboratory and compare their result with the one in the book.

The second source is research published by colleagues from our discipline, most often together with the university clinics. Studies on liver elastography and contrast-enhanced ultrasound [2,3], on eating behaviors and body mass index in students [4-6], on the digital engagement of patients with diabetes [7], on the course of COVID-19 in patients with type 1 and type 2 diabetes [8], on differences between women and men in cardiovascular emergencies [9] or on the validation of questionnaires in Romanian [10,11] are read in this book as case studies: what question they asked, what design they used, what test they applied, and what their figures do and do not say. Students who learn to read these articles pencil in hand will be able to read any article in the specialty they eventually choose in the same way.

The tools are those listed in the course syllabus, all free or available under an academic license: Google Forms and Google Sheets for collecting and organizing data, JASP and MedCalc for analysis, and Epi Info for epidemiological tables and for offline work. Each practical session gives the exact menus and the expected results, so that students can check their own work.

I have kept the working rules of the medical informatics volume. Every reference was actually located in PubMed, in CrossRef or on the website of the issuing institution before being cited. Figures from published studies are reproduced as they appear in the indexed abstract. Graphs built on the teaching dataset are labeled as such. The patients in the examples are fictitious. Where a piece of information could not be verified, I preferred not to write it.

Acknowledgments

This book owes a great deal to my colleagues in the department, with whom I have taught the same discipline for many years, with whom I wrote the discipline's practical session guides [12-14], and alongside whom I have worked on the statistical analysis of many clinical studies. For the teaching and scientific collaboration from which this course grew, I thank Associate Professor Radu Moleriu, PhD, Associate Professor Anca Tudor, PhD, Lecturer Mirela Frandeș, PhD, Lecturer Lavinia Moleriu, PhD, Teaching Assistant Flavius Cioca, PhD, and Teaching Assistant Raluca Lupușoru, PhD. Many of the examples in this book come from articles to which they contributed, and many of the laboratory problems were polished in our discussions at the end of classes. I also thank the students of recent years, whose questions, sometimes uncomfortable ones, decided what deserved a fuller explanation. Any remaining errors are my own.

Timișoara, October 2026

Călin Muntean

How to Read This Book

The book can be read in order, week by week, but this is not compulsory. The chapters refer to one another, and the map of the discipline in Chapter 14 shows how they connect. Readers who quickly need a particular topic, for example how to calculate an odds ratio or how to choose between the t-test and the Mann-Whitney test, can go straight to the relevant chapter and its paired practical session.

Every lecture chapter has the same structure. It opens with operational objectives, phrased as actions the student will be able to perform at the end. The content moves from concepts to methods, with examples worked first by hand and then on the teaching dataset. Case studies drawn from the discipline's research are set in separate boxes. Next come step-by-step worked problems and proposed exercises, a box of common pitfalls, that is, the mistakes seen again and again in the laboratory and in the exam, and then a box of key points. Each chapter ends with ten multiple-choice questions in the exam format: single choice (SC), with one correct answer, and multiple choice (MC), with two to four correct answers. The answers and the solutions to the exercises are given in Appendix A.

Each practical session has an objective, the tools with the version used, the working steps with the expected results, additional exercises, what to submit with an indicative grading scheme, and a variant for rooms without network access. All of them use the "mg2026" dataset, supplied in two versions: the raw file, with data entry errors deliberately left in, and the clean file. The data dictionary is in Appendix C, the formula sheet in Appendix D, the statistical tables in Appendix E, and three complete practice multiple-choice tests in Appendix F.

For regular study I recommend three steps. First, read the objectives. Then work through the text with its figures, tables and worked problems, redoing the calculations on paper or in the

spreadsheet. Finally, answer the questions and solve the exercises without looking at the answers in Appendix A. The common pitfalls are worth reading twice, once before the practical session and once before the exam.

Citations are numbered in Vancouver style, in order of first appearance, and the full list is at the end of the book. Abbreviations are collected in a separate list, and the main terms are briefly defined in the glossary in Appendix B, with a reference to the chapter in which they are explained. The text uses the decimal point, as is standard in English (0.05, 131.13 mmHg), which is also the convention most analysis software displays.

List of Abbreviations

List of the abbreviations used in this book, in alphabetical order. For acronyms of established English names, the full name is given in parentheses.

Abbreviation	Full term
ANOVA	analysis of variance
ARR	absolute risk reduction
AUC	area under the ROC curve
BMI	body mass index
CI	confidence interval
CONSORT	reporting guideline for randomized trials (Consolidated Standards of Reporting Trials)
COPD	chronic obstructive pulmonary disease
CRF	case report form
CV	coefficient of variation
DBP	diastolic blood pressure
df	degrees of freedom
DOR	diagnostic odds ratio
EBM	evidence-based medicine
FINDRISC	Finnish Diabetes Risk Score
FN	false negative
FP	false positive
GDPR	General Data Protection Regulation, Regulation (EU) 2016/679
GRADE	system for rating the certainty of evidence (Grading of Recommendations Assessment, Development and Evaluation)
H0, H1	null hypothesis, alternative hypothesis
HbA1c	glycated hemoglobin
HR	hazard ratio
HTN	arterial hypertension
ICC	intraclass correlation coefficient
IQR	interquartile range
ITT	intention-to-treat analysis
LR+, LR-	positive and negative likelihood ratio
MC	multiple choice (exam question type)
NNH	number needed to harm (number of patients who must be exposed for one adverse event to occur)
NNT	number needed to treat (number of patients who must be treated to prevent one event)
NPV	negative predictive value
OGTT	oral glucose tolerance test
OR	odds ratio
PICO	population, intervention, comparison, outcome
PPV	positive predictive value
PR	prevalence ratio
PRISMA	reporting guideline for systematic reviews
RCT	randomized controlled trial
ROC	receiver operating characteristic (ROC) curve
RR	relative risk
RRR	relative risk reduction

Abbreviation	Full term
SBP	systolic blood pressure
SC	single choice (exam question type)
SD	standard deviation
SE	standard error
Se	sensitivity
SEE	standard error of the estimate in regression
Sp	specificity
STARD	reporting guideline for diagnostic accuracy studies
STROBE	reporting guideline for observational studies
TN	true negative
TP	true positive

Part I. Lectures

Chapter 1. Introduction to Biostatistics: Fundamental Concepts, Role in Medical Research and Elements of Probability

Objectives

By the end of this chapter, the student will be able to:

- explain why biological variability makes statistics necessary in medicine and give examples from family practice;
- define the statistical unit, the population, the sample and their sizes, and identify them in a published article;
- compare the census, screening and sampling, and choose a sampling method suited to a given question;
- recognize the main types of systematic error (selection, information, confounding) in a clinical scenario;
- distinguish descriptive from inferential statistics and use the notation for parameters (μ, σ, π) and for statistics (m, s, p_i);
- describe the stages of a statistical study and formulate a research question in PICO format;
- list the sources of biomedical data and the ethical and legal principles governing their processing (GDPR, Law 190/2018, anonymization, pseudonymization, consent, ethics committee approval);
- calculate simple, conditional and total probabilities from a contingency table and check whether two events are independent;
- calculate simple binomial probabilities and recognize the situations in which the Poisson distribution is used.

1.1 Why a Physician Needs Biostatistics

Many students meet biostatistics for the first time with an honest question: "What do I need this for, if I want to be a surgeon?" The answer does not come from mathematics but from an ordinary day in a doctor's life. In the morning, the family doctor has on the desk the laboratory results of the patients seen last week. Is a fasting plasma glucose of 118 mg/dL in a patient who had 104 mg/dL a year ago a real deterioration or an unimportant fluctuation? At lunch the doctor reads the abstract of a trial showing that a new antidiabetic drug lowers the risk of hospitalization for heart failure "by 30%". In the evening the doctor fills in a risk score for a patient and has to explain what a "high" result means. Each of these situations calls for a judgment about uncertainty, and biostatistics is the language in which that judgment is expressed.

Evidence-based medicine has been defined as the conscientious, explicit and judicious use of the current best evidence in making decisions about the care of individual patients, integrated with the clinical expertise of the physician and the values of the patient [15]. The word "evidence" almost always hides a statistical calculation: a confidence interval, a p value, a relative risk, a sensitivity. The clinical guidelines you will use throughout your career grade their recommendations according to the quality of the studies they are based on, and the quality of a study is judged largely by its design and statistical analysis. A physician who does not understand what "95% CI 0.62-0.91" means cannot check unaided whether a recommendation rests on solid data or on a fragile result.

There is also a very concrete and more immediate reason: the residency examination. Its reading list includes chapters on epidemiology, on the evaluation of diagnostic tests and on the interpretation of clinical trials. Questions about sensitivity and predictive value, relative risk or study types come up year after year. The exam for this course is built in the same style, with single-choice and multiple-choice questions, precisely so that you become used to this kind of reasoning early on.

What do medical students think about these subjects? An anonymous survey of 121 students at our university asked them to rate several courses on a scale from 1 to 10 [16]. The emergency medicine course received a median of 9 (IQR 7-10), biomedical informatics 7 (5-9), and biostatistics 8 (5-9). The most interesting result, however, lies elsewhere: the correlation between knowledge measured by a test and knowledge as self-assessed by the students was practically zero ($R = 0.107$, $p = 0.246$). In other words, students cannot judge their own level of knowledge in this field correctly; some believe they know more than they do, others underestimate themselves. Note also how the results were reported: median and interquartile range, because a mark on a scale from 1 to 10 is an ordinal variable, not a true measurement (Chapter 2). Colleagues from our discipline have published a separate analysis of the cost and the opportunities of teaching with information technology during the pandemic years [17]; many of the digital tools used in the practical sessions of this book became established at that time.

Last but not least, many of you will do research: a graduation thesis, a project in a student research group, a doctoral thesis. Without some grounding in biostatistics, collaboration with a statistician turns into a monologue. The statistician cannot choose the right test alone without understanding the clinical question, and the physician cannot formulate the question correctly without knowing what an analysis can and cannot offer. This book tries to prepare you for that dialogue, as do the reference textbooks of the discipline [14,18,19].

1.1.1 Variability, the Raw Material of Biology

Two patients of the same age, the same sex and with the same diagnosis never have the same blood pressure. Nor does the same patient have the same blood pressure at two consecutive measurements taken five minutes apart. This variability is the basic feature of biological phenomena, and it has several sources. There is variability between individuals (genetics, age, lifestyle, comorbidities), variability over time within the same individual (circadian rhythm, exertion, stress, diet) and measurement variability (the device, the technique, the observer, laboratory conditions).

If all values were identical, statistics would not be needed: a single measurement would tell us everything about the entire population. Precisely because values differ, we need methods that describe how much they differ (dispersion), around which value they cluster (central tendency) and how certain the conclusions drawn from a limited number of observations are. In the teaching dataset mg2026, presented in Section 1.3, systolic blood pressure at enrollment ranges from 102 to 169 mmHg, even though all 160 patients were recruited according to the same criteria. The mean is 131.13 mmHg and the standard deviation 13.88 mmHg; in Chapter 2 we will learn what these figures mean.

In the laboratory, students often ask whether variability is "an error". The answer is no. Real biological variability is information: it tells us how different patients are from one another. Measurement error is added on top of it, and we can reduce it with calibrated devices, standardized protocols and repeated measurements. The former cannot be eliminated, nor should it be. Biostatistics accepts it and learns to work with it.

1.1.2 Statistics and Biostatistics

Statistics is the science concerned with collecting, organizing, describing, analyzing and interpreting data on mass phenomena, that is, phenomena that occur in a large number of individuals and cannot be understood from a single case. Biostatistics is the application of these methods to data from biology, medicine and public health. The difference does not lie in the formulas, which are the same, but in the type of problems: clinical trials, diagnostic tests, risk factors, survival, severity scores, quality-of-life questionnaires.

Biostatistics has a few particular features. The data come from people, with their consent, and are protected by strict ethical and legal rules. Measurements often have clinical units (mmHg, mg/dL, kg/m²) and thresholds with medical meaning (140 mmHg for hypertension, 126 mg/dL for diabetes). Many variables are skewed (triglycerides, length of hospital stay, antibody titers), and many outcomes are binary (cured or not, dead or alive). Above all, the conclusions turn into decisions about patients, which means that mistakes have a real cost.

1.2 Basic Concepts: Statistical Unit, Population, Sample

Every study starts with a simple question: "Who are we talking about?" The precise answer to this question defines a few notions you will use throughout the rest of the course.

1.2.1 The Statistical Unit and the Population

The statistical unit (statistical individual) is the element on which observations are made, located in space and time. In most medical studies the statistical unit is the patient. This is not always the case, however. In a prescribing study the unit may be the prescription; in an ophthalmology study, the eye; in a dental study, the tooth or the tooth surface; in a health services study, the practice or the hospital. The choice of unit has consequences: the two eyes of the same patient are not independent observations, and the analysis must take this into account.

The population is the set of all statistical units that share at least one common property and that are, in turn, located in space and time. "Adults with prediabetes in Timiș County in 2026" is a well-defined population; "patients with prediabetes" is not, because it does not say where or when. The number of units in the population is called the population size and is usually denoted by N.

In research, a distinction is made between the target population, to which we want to generalize our conclusions, and the accessible (or source) population, from which we can actually recruit. For a study in family practices in the west of the country, the target population may be "adults at cardiometabolic risk in Romania", but the accessible population is "adults registered on the lists of the 12 participating practices". The gap between the two is one of the first questions the reader of an article should ask.

1.2.2 The Sample

The sample is a subset of the population chosen to be studied in its place. The number of units in the sample is the sample size and is denoted by n. The sample of 160 patients in the mg2026 dataset comes from an accessible population of several thousand adults registered with the participating practices; the analyses are carried out on the 160, but the question of interest concerns the whole population.

The sample size is not chosen to be "as large as possible", nor "as large as we can manage". It is calculated before the study begins, according to the desired precision or the difference we want to detect. We will see in Chapter 4 that, to estimate mean systolic blood pressure with a margin of ±3 mmHg when the standard deviation is about 14 mmHg, 84 patients are enough, but to estimate

a proportion of 20% with a margin of ±5%, 246 are needed. The difference comes from the nature of the variable and the precision required, not from the researcher's wishes.

1.2.3 Census, Screening, Sampling

There are three main ways of collecting data about a population. A census observes all the units of the population at a given moment. It is exact (there is no sampling error) and well located in time and space, but very costly, slow and rarely feasible in medicine. Romania's population census is carried out once every ten years; a complete national registry of a rare disease comes close to a census of the cases.

Screening systematically examines a population (or a large segment of it) in order to identify people with a particular condition, often still undiagnosed. Classic examples are screening for cervical cancer, colorectal cancer, hearing loss in newborns, or type 2 diabetes with the FINDRISC score in family practice. Screening does not require strict localization in time (it may take place over several years), has a relatively high cost and aims at an individual benefit, not merely a statistical estimate.

Sampling (selection) studies only part of the population and extends the conclusions to the whole. The cost is lower, the data can be collected more carefully, and the result is available sooner. The price paid is uncertainty: the result obtained in the sample inevitably differs from the true value in the population. Inferential statistics exists precisely to quantify this uncertainty. Almost all the studies you will read are studies on samples.

1.3 The mg2026 Teaching Dataset

Throughout the course we will work on the same dataset, so that you can follow how the same variables are described, estimated, compared and correlated. The dataset is called mg2026 and is available on Moodle in two versions: mg2026_curat.csv (160 rows, ready for analysis) and mg2026_brut.csv (161 rows, with errors deliberately left in, which you will clean in Practical Session 1).

The scenario is a simulated pragmatic randomized trial conducted in 12 family practices in the west of the country. It enrolled 160 adults at cardiometabolic risk: people who were overweight, had prediabetes or had previously undiagnosed diabetes. Patients were randomly allocated to three arms: A, standard lifestyle counseling (53 patients); B, counseling plus a self-monitoring mobile app (53); C, counseling, the app and a monthly teleconsultation with the family doctor (54). Measurements were taken at enrollment, at 3 months and at 6 months. Diabetes was confirmed at enrollment by the oral glucose tolerance test (OGTT), which is the reference standard against which the FINDRISC score is evaluated. The patients are fictitious, and the values were generated by simulation so as to resemble real data from family practice without coming from any actual person.

Table 1.1 summarizes the variables in the dataset. The full dictionary is given in Appendix C and will be used again in Chapter 2, when we discuss the types of variables.

Table 1.1. Variables of the mg2026 teaching dataset.

Variable	Meaning	Type (see Chapter 2)
id	patient code (P001-P160)	identifier
brat	treatment arm (A, B, C)	nominal
sex	F, M	nominal dichotomous
varsta	age at enrollment, years	numerical
mediu	urban, rural	nominal dichotomous
educatie	1 lower secondary school, 2 high school or vocational school, 3	ordinal

Variable	Meaning	Type (see Chapter 2)
	university, 4 postgraduate	
fumat	nefumator, fost fumator, fumator (never smoker, former smoker, current smoker)	nominal
activitate	self-rated physical activity, 1-5	ordinal
inaltime, greutate0, imc0, talie	height (cm), weight (kg), BMI (kg/m²), waist circumference (cm)	numerical
tas0, tad0, fc0	systolic and diastolic blood pressure (mmHg), heart rate (beats/min)	numerical
colesterol, hdl, trigliceride	lipid profile (mg/dL)	numerical
glicemie0, glicemie3, glicemie6	fasting plasma glucose at 0, 3 and 6 months (mg/dL)	numerical
hba1c0	glycated hemoglobin (%)	numerical
findrisc	diabetes risk score (0-26)	discrete numerical (score)
diabet_ogtt	diabetes confirmed by OGTT (0/1)	nominal dichotomous
greutate6, tas6	weight and SBP at 6 months	numerical
scadere5	weight loss ≥ 5% at 6 months (0/1)	nominal dichotomous
satisfactie	satisfaction with the program, 1-5 (Likert)	ordinal

A few figures will help you get your bearings from the outset. Of the 160 patients, 86 are women and 74 men; 95 live in urban and 65 in rural areas; 74 are never smokers, 33 former smokers and 53 current smokers. The mean age is 56.29 years, with a range of 35 to 79 years. Diabetes was confirmed by OGTT in 29 patients, that is, 18.1%. These figures will reappear, in different forms, in almost every chapter.

One design detail is worth remembering from now on: at enrollment, the mg2026 data can be analyzed as a cross-sectional study (for example, the association between obesity and diabetes), whereas the follow-up at 3 and 6 months in randomized arms turns them into a controlled clinical trial. The same dataset can therefore answer questions of different kinds, provided we know what type of conclusion each part of it allows.

1.4 Sampling: Representativeness and Methods

A sample is useful only if it resembles the population it comes from. We say that a sample is representative when it reproduces, in similar proportions, the structure of the population for the characteristics that matter in the study: age, sex, area of residence, disease severity and so on. Representativeness cannot be guaranteed by intuition. It is achieved through the way individuals are selected.

1.4.1 Conditions for a Representative Sample

Two conditions underlie correct selection. The first is equal probability: every individual in the population must have the same probability of being included in the sample (or, in more complex designs, a known, nonzero probability). The second is independence: the inclusion of one individual must not influence the inclusion of another. If a doctor invites not only the patient but also all members of the patient's family into the study, independence is lost, because members of the same family resemble each other in diet, habits and genes.

Any condition that gives certain individuals a greater chance of being selected distorts the sample. The result is a systematic selection error (selection bias), which we discuss in Section 1.5. A practical rule: if you cannot describe exactly how each patient came to be in the sample, you cannot claim that the sample is representative.

1.4.2 Simple Random Sampling

Simple random sampling is the theoretical model on which almost all statistical formulas are based. One starts from a complete list of the population (the sampling frame), each individual receives a number, and the numbers to be included are chosen with a random number generator. In Google Sheets, for example, you can add a column with =RAND() next to the list of patients, sort by this column and keep the first n rows.

Suppose a practice has 1,850 adults aged 40 to 75 on its list and wants to invite 120 of them for a FINDRISC assessment. With simple random sampling, each patient has the same probability of being invited, 120/1,850 = 0.065 (6.5%). The method is correct, but it has two weak points in practice: it requires a complete and up-to-date list, and it can produce, by chance, an unbalanced sample (for example, with too few men), especially when n is small.

1.4.3 Systematic Sampling

In systematic sampling, a sampling interval $k = \frac{N}{n}$ is calculated, a random starting point between 1 and k is chosen, and then every k-th individual on the list is taken. For 1,200 patients and a sample of 80, the interval is k = 15. If the random starting number drawn is 7, patients numbered 7, 22, 37, 52 and so on are included, up to 1,192.

The method is simple and fits the workflow of a practice very well: "every fifth patient who comes in for a consultation in March". The danger arises when the list has a periodicity that coincides with the interval. If appointments are organized so that every fifth appointment of the day is reserved for urgent checkups, an interval of 5 may select only such patients. Before using the method, check how the list is ordered.

1.4.4 Stratified Sampling

The population is divided into homogeneous strata according to an important characteristic (sex, age group, urban or rural residence), and then a random sample is drawn from each stratum. In the proportional version, each stratum contributes a number of individuals proportional to its size in the population. If, of 1,200 patients, 720 live in urban areas (60%) and 480 in rural areas (40%), a stratified sample of 80 will contain 48 urban and 32 rural patients.

Stratification increases the precision of the estimates when the stratification characteristic influences the variable under study, and it guarantees that small groups are not missing from the sample. Disproportionate stratification is sometimes used on purpose, oversampling a rare group (for example, patients over 75), followed by weighting of the results at the analysis stage.

1.4.5 Cluster Sampling

When there is no list of all individuals but there is a list of natural groups (practices, schools, villages, hospital wards), a few groups are chosen at random and either all individuals in the chosen groups or a sample from each are studied (two-stage sampling). For a national study on blood pressure control, it is much easier to choose 40 family practices at random and include 25 patients from each than to choose 1,000 patients scattered across the whole country.

The logistic advantage comes at a statistical cost. Patients from the same practice resemble one another (same doctor, same area, same prescribing habits), so 1,000 patients from 40 practices carry less information than 1,000 individually chosen patients. The analysis must take this clustering into account. The mg2026 dataset has a similar structure, with 12 participating practices, even though in this course we will analyze it, for simplicity, as a sample of individuals.

1.4.6 Convenience Sampling

A convenience sample includes the individuals who are easiest to recruit: patients who come for a consultation during a given period, volunteers who respond to an advertisement, people who fill in an online questionnaire shared on social media. It is by far the most widely used method in clinical research, because it is fast and cheap. It is also the most exposed to selection bias, because nothing guarantees that those recruited resemble the target population.

A convenience sample is not automatically useless. Randomized clinical trials almost always use consecutively recruited patients, that is, a convenience sample, and the validity of the comparison between arms comes from randomization, not from representativeness. The problem arises when a result obtained in a convenience sample (for example, a prevalence) is presented as valid for the whole population. A physician reading such a study should ask who responded, who did not respond, and why.

Table 1.2. Sampling methods at a glance, with examples from family medicine.

Method	Principle	Example	Advantage	Limitation
Simple random	every individual has the same chance	120 patients drawn with RAND() from the practice list	theoretical basis of the formulas	requires a complete list; chance imbalances
Systematic	every k-th individual, random start	every 15th patient on the list, start 7	simple, easy to apply at triage	periodicity of the list
Stratified	random sample within each stratum	48 urban + 32 rural, proportional	greater precision, small groups represented	requires information about the strata
Cluster	groups are chosen at random	40 practices, 25 patients each	easy logistics	correlated observations within clusters
Convenience	whoever is available	patients seen in March	fast, cheap	high risk of selection bias

Case study: the same cohort of students, two different questions

Two studies published in 2026 by colleagues from our discipline used the same group of 921 first-year students aged 18 to 24. In the cross-sectional study on convenience-oriented eating behaviors, 24.1% of participants were classified as overweight or obese, of whom 19.4% were overweight and 4.7% obese [4]. In the prospective study on chrononutrition, the same students were followed over their first academic year, and the mean change in BMI was -0.45 ± 1.07 kg/m² (95% CI -0.52 to -0.38; $p < 0.001$); 48.2% had a decrease in BMI, 38.7% remained stable and 13.1% increased, and clinically relevant weight gain occurred in 7.2% of the cohort [5]. A third study, of 693 students of the same age, reported a prevalence of prediabetes of 21.1% (diabetes 1.4%) [6]. The question the reader should ask is not whether the figures are correct but to which population they can be extended. The students of a single university of medicine and pharmacy do not represent all 18- to 24-year-olds in Romania: they have a higher level of education, better access to medical information and a specific lifestyle. The conclusions are valid for the source population and, with caution, for students at similar universities. What should additionally be reported in order to judge representativeness: how many students were invited, how many agreed (the participation rate), and how those included compare with those who declined, at least by sex and faculty.

1.5 Systematic Errors (Bias)

Any result obtained in a sample differs from the true value in the population. The difference has two components. Random (sampling) error arises because we studied only part of the population; it shrinks as n increases and can be quantified with confidence intervals. Systematic error (bias)

always shifts the result in the same direction and does not disappear when we enlarge the sample. A scale that reads 2 kg too high will give a mean weight that is 2 kg off whether we weigh 10 or 10,000 patients.

This distinction is fundamental for critical reading. A large study has a small random error, that is, narrow confidence intervals, but it may have a large systematic error. Precision does not mean accuracy. Systematic errors are usually classified into three broad families: selection bias, information (measurement) bias and confounding.

1.5.1 Selection Bias

Selection bias occurs when the way participants enter the study (or remain in it) depends on both the exposure and the outcome, so that the sample no longer reflects the population. Some common forms:

The "healthy volunteer" effect: people who volunteer for a prevention program are, on average, more health-conscious, smoke less and exercise more than the general population. A screening program evaluated only on volunteers will appear more effective than it really is.

Nonresponse: if only 40% of the people invited fill in a questionnaire on alcohol consumption, heavy drinkers are likely to be underrepresented among the respondents. The estimated prevalence will be too low.

Self-selection in online surveys: a questionnaire shared on a social network is filled in by those who use that network, are interested in the topic and have the time. The age, sex and educational level of the respondents may differ greatly from those of the population.

Two recent studies illustrate the problem well. An online convenience survey on eating behaviors included 753 young people aged 18 to 30, 93.5% of them women; 11.6% were underweight and 29.3% overweight or obese [20]. The authors correctly described the design in the title itself and reported the prevalences with Wilson confidence intervals (a method we will learn in Chapter 4), but a proportion of 93.5% women shows that the results actually concern the young women who chose to respond. Another online survey, of people with diabetes, obtained 329 validated responses, and 44.07% of respondents reported using herbal supplements [21]. It is quite possible that patients interested in supplements were more willing to fill in a questionnaire on this topic, which would inflate the prevalence. By contrast, a cross-sectional study of 2,976 students in Timiș County, 62.5% female and 37.5% male, aged 18 to 25, had a large sample, but even a large size does not guarantee the absence of selection bias: the question remains who was invited and who responded [22].

Loss to follow-up (in prospective studies): if patients who do not respond to treatment drop out of the study more often than the others, those remaining at the end show a better effect than the real one.

Spectrum bias (in diagnostic test studies): a test evaluated in patients with advanced disease, compared with perfectly healthy volunteers, looks much better than it will be in family practice, where it has to distinguish mild forms of the disease from other, similar conditions. We will return to this bias in Chapter 12.

1.5.2 Information (Measurement) Bias

Information bias occurs when variables are measured or classified incorrectly and the error differs between groups. Recall bias is the classic example: mothers of children born with a malformation remember the medicines taken during pregnancy in much more detail than mothers of healthy children, not because they took more, but because they have thought more about possible causes. In a case-control study, this difference can create an association that does not exist.

Other forms are observer bias (the doctor who knows which treatment the patient received tends, involuntarily, to rate the patient's progress more favorably), instrument bias (an uncalibrated blood pressure monitor), reporting bias (patients underreport smoking or alcohol consumption to the doctor) and "terminal digit preference for zero", very common in blood pressure measurement: an unusually large proportion of values ending in 0 (130, 140, 150) shows that the measurements were rounded. Blinding (masking) of assessors and standardized measurement protocols are the main means of prevention.

1.5.3 Confounding

Confounding occurs when the association between an exposure and an outcome is distorted by a third variable that is related to both the exposure and the outcome without being an intermediate step between them. The textbook example: in some old studies, coffee drinkers had lung cancer more often. Coffee does not cause lung cancer; smokers drink more coffee, and smoking causes lung cancer. Smoking is the confounder.

In family medicine, age is the most common confounder. Patients who receive more medicines have more hospital admissions, but they are also older and have more diseases. A crude association of "more medicines, more admissions" says nothing about the effect of the medicines until age and comorbidities are taken into account. Confounding is prevented by design (randomization, restriction, matching) and controlled at the analysis stage by stratification or by multivariable models, which we will meet in Chapters 9 and 10.

A point that often comes up in the exam: randomization is the only method that balances both known and unknown confounders between groups. This is why randomized trials sit at the top of the evidence hierarchy for questions about the effects of treatments.

> **Case study: when an association disappears after adjustment**
>
> In a prospective study of 278 patients hospitalized with decompensated heart failure and followed for 12 months, the hemoglobin-to-red cell distribution width ratio (HRR) was analyzed as a prognostic marker [23]. The median HRR was 0.89, and the primary event occurred in 167 patients (60.1%). Survival curves showed a difference between patients with low and high HRR (log-rank p = 0.027), but after adjustment for age, sex, NT-proBNP, ejection fraction and glomerular filtration rate, the association was no longer statistically significant (p = 0.240). The correct interpretation is that a large part of the crude association was explained by the adjustment variables: patients with low HRR were probably older, with worse renal and cardiac function. HRR remains an easily obtained marker, but its prognostic information largely overlaps with that of the classic markers. The authors acted correctly in reporting both analyses; the reader should remember that a crude association is not proof of an independent effect. What would help further: the hazard ratio (HR) with its confidence interval for both models, to show how much the effect shrank, not merely whether it crossed the 0.05 threshold.

1.6 Descriptive Statistics and Inferential Statistics

Biostatistics has two main branches, which correspond to two different questions. Descriptive statistics answers the question "What did we observe?": it organizes data into frequency tables, summarizes them with indicators (mean, median, standard deviation, proportions) and presents them graphically. It concerns only the sample studied and makes no statements about the population. Chapter 2 is devoted entirely to it.

Inferential statistics answers the question "What can we say about the population on the basis of what we observed?" It uses probability theory to estimate population parameters (with

confidence intervals) and to test hypotheses (with statistical tests). Figure 1.1 shows this path schematically: a sample is drawn from the population, statistics are calculated on the sample, and from the statistics conclusions are drawn, with a certain probability of error, about the population parameters.

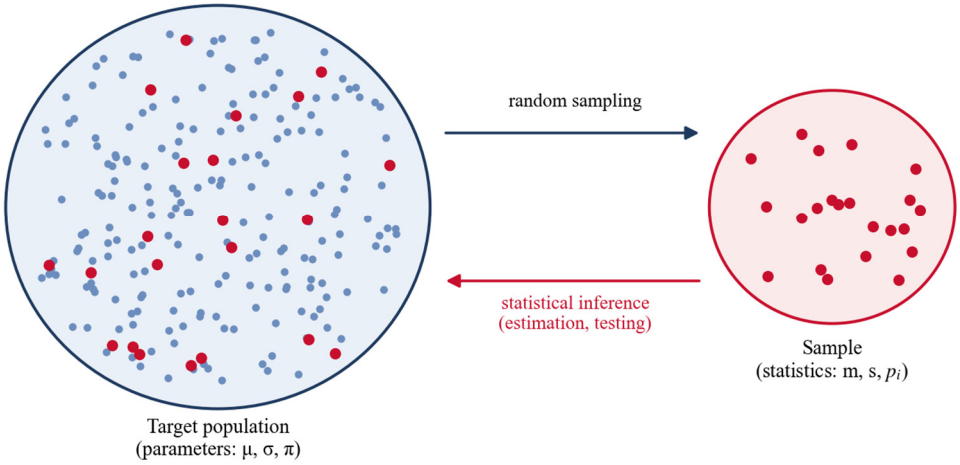


Figure 1.1. Population, sample and statistical inference. A sample is drawn from the target population (with unknown parameters μ, σ, π), and statistics (m, s, p_i) are calculated on it; inference goes in the opposite direction, from the sample to the population, and is always accompanied by a measure of uncertainty.

1.6.1 Parameters and Statistics: Notation

A constant source of confusion in the exam is mixing up the notation. The characteristics of a population are called parameters and are denoted, by convention, with Greek letters: the population mean μ (mu), the population standard deviation σ (sigma), the population proportion π (pi, which here has nothing to do with 3.14). Parameters are fixed but usually unknown.

Characteristics calculated on the sample are called statistics (or estimators) and are denoted with Latin letters: the sample mean m ($\bar{x}$ in some books), the sample standard deviation s, the sample proportion p_i (p with subscript i), that is, the proportion of subjects in class i of the variable out of all subjects studied: $p_i = \frac{n_i}{n}$, where n_i is the number of subjects in class i. Statistics can be calculated exactly, but they vary from one sample to another. Table 1.3 summarizes the convention, with values from the mg2026 dataset.

Table 1.3. Parameters and statistics: notation and examples from the mg2026 dataset.

Characteristic	Parameter (population)	Statistic (sample)	Value in mg2026 (n = 160)
mean SBP at enrollment	μ (unknown)	m	131.13 mmHg
standard deviation of SBP	σ (unknown)	s	13.88 mmHg
proportion of patients with confirmed diabetes	π (unknown)	p_i	$\frac{29}{160} = 0.181 (18.1\%)$
size	N (thousands of patients)	n	160

The value $p_i = 0.181$, where class i is "confirmed diabetes", is a point estimate of π. Another sample of 160 patients from the same practices would almost certainly have given a slightly different proportion, perhaps 15% or 21%. In Chapter 4 we will see that the 95% confidence interval for π, calculated with the Wilson method, is 12.9-24.8%. This interval says something about the population; the figure of 18.1% says something about the sample.

Beware of a vocabulary trap: this is why the sample proportion is written with a subscript, p_i, and not just with the letter p. Without a subscript, in the chapters on statistical tests the letter p

denotes the "p value", that is, a probability calculated by the test, and the two notions must not be confused.

1.7 The Stages of a Statistical Study and the Research Question

A study does not begin with the data, nor with the statistical software. It begins with a question. Figure 1.2 shows the sequence of stages, from the clinical question to the decision; each stage depends on the quality of the one before it.

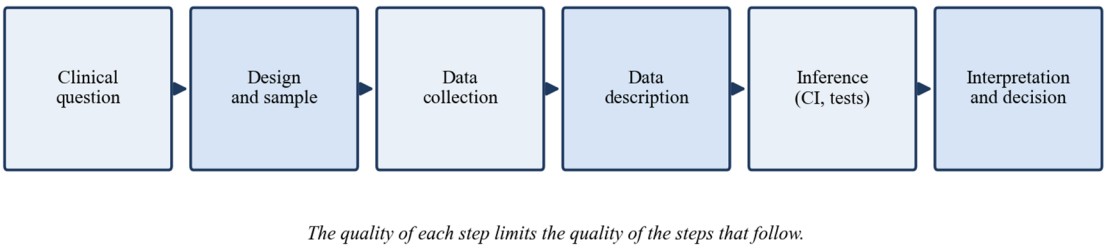


Figure 1.2. The stages of a statistical study: the clinical question, the design and the sample, data collection, data description, inference (confidence intervals and tests), and interpretation with the clinical decision. Errors made in the early stages can no longer be corrected by analysis.

1.7.1 The Research Question and the PICO Format

A good research question is precise, clinically relevant and can be checked against data. The PICO format helps to structure it into four elements: P (population or patients), I (intervention or exposure), C (comparison) and O (outcome) [24]. Sometimes T (time) and S (study type) are added.

For the mg2026 scenario, a vaguely worded question would be "Do mobile apps help patients with prediabetes?" In PICO format it becomes: "In adults at cardiometabolic risk in family practices (P), does adding a self-monitoring app and monthly teleconsultation to standard counseling (I), compared with standard counseling alone (C), increase the proportion of patients who lose at least 5% of their body weight within 6 months (O)?" The second wording states exactly what data must be collected, which groups are compared and which variable will be analyzed. The answer, which we will calculate in Chapter 10, is that the proportion was 29.6% in arm C versus 11.3% in arm A.

The PICO format is not a formality. It forces the researcher to choose a single primary outcome before seeing the data, which prevents one of the most common problems in the literature: testing dozens of variables and reporting only the "significant" ones.

1.7.2 From the Question to the Design and Data Collection

Once the question has been formulated, the design is chosen (cross-sectional, case-control, cohort, randomized clinical trial; see Chapters 10 and 13), the population is defined, the inclusion and exclusion criteria are set, the sample size is calculated, and the collection instrument is built: the observation sheet, the electronic form or the questionnaire. Each variable receives an operational definition, a unit of measurement and a coding scheme. Only then does recruitment begin.

The statistical analysis is also planned before data collection, in a statistical analysis plan. This specifies the primary outcome, the tests to be used, the significance level and how missing data will be handled. Analysis "after seeing the data" is not forbidden, but it must be presented as exploration, not as the testing of a hypothesis. Reporting guidelines, such as STROBE for observational studies, require these elements to be described explicitly in the article [25].

> **Case study: the stages of a questionnaire study, from translation to report**
>
> A questionnaire used in a language other than the one in which it was created cannot simply be translated by a single bilingual speaker. The ALSATION protocol, published by a team that includes the author of this course, describes the plan for translating three health literacy questionnaires into Romanian [26]. The fact that the translation of an instrument warranted a separately published protocol says a great deal: the "instrument building" stage requires the same rigor as the analysis. In general, adapting a questionnaire involves independent translations, their reconciliation, back-translation, expert review and testing on a small group of respondents, and each of these steps can change the meaning of a question. On the other hand, a questionnaire study conducted in 10 urban and rural pharmacies included 189 people aged 40 and over, 78.8% women and 69.3% urban residents; 75.1% said they knew what osteoporosis was, but 77.3% had never had a DXA bone densitometry scan, and 42.9% of those who reported a diagnosis of osteoporosis were not receiving treatment [27]. Both studies show that the result of a questionnaire depends on stages that do not appear in the final table: how the questions were worded, where they were administered, who agreed to answer. In the pharmacy study, the predominance of women (78.8%) and of urban respondents (69.3%) partly reflects who walks into a pharmacy and agrees to fill in a questionnaire, so the prevalences should be read as valid for this type of respondent, not for all Romanians over 40.

1.8 Data in the Era of the Electronic Health Record: The Link with Medical Informatics

In the first semester, in medical informatics, you learned how health data are represented, stored and transmitted: the electronic health record, interoperability standards, classifications and nomenclatures, information security [1]. Biostatistics starts exactly where that course left off. A statistical analysis can be no better than the data it processes, and most of the problems you will meet in practice have nothing to do with the choice of test but with the quality, structure and legality of the data.

1.8.1 Where Biomedical Data Come From

Data sources can be grouped into a few broad categories, each with its own advantages and limitations [28].

The electronic health record (EHR) and the information systems of practices and hospitals contain diagnoses, prescriptions, laboratory results and consultations recorded during routine clinical practice [29]. The advantage is the huge volume and low cost of data that already exist. The limitation is that they were collected for patient management and reimbursement, not for research: diagnoses may be coded to justify a prescription, some values are missing, and definitions differ from one unit to another. Diagnoses are usually coded in the International Classification of Diseases [30], and laboratory tests can be identified by standard codes such as LOINC [31].

Disease registries (cancer, diabetes, dialysis and rare disease registries) and epidemiological surveillance data reported to public health directorates systematically track all cases in an area. They are valuable for estimating incidence and trends over time, but they depend on the completeness of reporting.

Research questionnaires and forms (on paper or electronic, for example in Google Forms) collect exactly the variables the study needs, with controlled definitions. The cost is higher, and quality depends on the wording of the questions, the validation of the instrument and the response rate.

Wearable devices and mobile apps (smartwatches, continuous glucose monitors, connected blood pressure monitors, self-monitoring apps such as the one in arm B of the mg2026 study) produce long series of measurements, at a frequency impossible to achieve in the practice. Their problems are calibration, missing data when the patient does not wear the device, and the fact that users of these technologies are usually younger and more motivated than the average patient.

Laboratory and imaging data come from instruments and information systems that follow their own standards; they are generally precise, but reference values and methods may differ between laboratories, which matters when data from several centers are combined.

1.8.2 Data Quality and the Data Dictionary

Data quality has several dimensions: completeness (how many values are missing), correctness (the values reflect reality), consistency (the same information is coded the same way everywhere), plausibility (the values are biologically possible) and timeliness. In the file mg2026_brut.csv you will find typical errors: sex written once as "F", another time as "f" or "Feminin"; a systolic blood pressure of 1540 mmHg (one digit too many); a height of 1.72 entered in meters in a column in centimeters; a weight of 7.26 kg (decimal separator in the wrong place); a blood glucose entered as "87 mg/dl", that is, text instead of a number; a duplicate row. Each of them would spoil a mean, a frequency or a test if it were not found in time.

The basic tool for data quality is the data dictionary. For each variable it specifies the name (short, without spaces or diacritics, for example tas0), the description, the type (numeric, text, calendar date), the unit of measurement, the permitted values or plausible range (for example, SBP between 70 and 260 mmHg), the codes for categorical variables (1 = lower secondary school, 2 = high school...) and the code used for missing values. The dictionary is written before data collection, not after; it then becomes the basis for the validation rules in the electronic form and in the spreadsheet. Practical Session 1 asks you to build such a dictionary.

The golden rule of data cleaning is simple: a suspicious value is not corrected by intuition but checked against the source document. A blood pressure of 1540 mmHg looks obviously like 154 mmHg with an extra digit, but it may just as well be a mistyped 150 or 140, or a completely different value; even the most likely hypothesis must be confirmed. If the source document is not available, the value is marked as missing and the decision is recorded in a change log. The original raw data are never deleted; one works on a copy.

1.8.3 Ethics and Data Protection in the Processing of Biomedical Data

For a passing grade of 5, the course syllabus requires familiarity with the principles of ethics in the processing of biomedical data. This is not a formality. Health data are among the most sensitive information about a person, and a data leak can affect the patient's employment, insurance or social relationships.

The General Data Protection Regulation (GDPR) has applied throughout the European Union since 2018 and lists data concerning health, genetic data and biometric data among the special categories of data, whose processing is prohibited in principle, with the exceptions expressly provided for by the regulation, including the explicit consent of the person, the provision of health care and scientific research accompanied by appropriate safeguards [32]. In Romania, Law no. 190/2018 lays down the measures for implementing the regulation, including specific rules for processing genetic, biometric and health data and for processing the personal numeric code (the

Romanian national identification number) [33]. The Regulation on the European Health Data Space, adopted in 2025, adds a framework for the secondary use of health data in research through authorized access bodies [34].

Several GDPR principles translate directly into working rules for a student doing research. Purpose limitation: data collected for care are not freely reused for another purpose. Data minimization: only the variables needed for the research question are collected; the full date of birth is not requested if age in years is enough, and the address is not recorded if urban or rural residence is enough. Integrity and confidentiality: files are kept on protected media, not on a USB stick left in the laboratory, nor in personal cloud storage accounts.

The difference between anonymization and pseudonymization is a frequent exam question. Pseudonymization replaces direct identifiers (name, personal numeric code, telephone number) with a code, for example P001, but keeps a separate correspondence list through which the patient can be re-identified. Pseudonymized data remain personal data and still fall under the GDPR. Anonymization irreversibly removes any reasonable possibility of re-identification, including by combining several variables; truly anonymous data are no longer personal data. Note: deleting the name is not anonymization. A file containing the date of birth, sex, place of residence and the diagnosis of a rare disease can identify a patient even without a name. The mg2026 dataset has codes P001-P160, but because it is simulated it does not correspond to any real person.

In addition to data legislation, research on human beings is governed by ethical principles. The Declaration of Helsinki of the World Medical Association, revised in 2024, requires that any research protocol involving human subjects be approved in advance by an independent ethics committee and that participants give freely expressed informed consent after receiving clear information about the purpose, risks, benefits and their right to withdraw at any time [35]. For retrospective studies on existing data, the ethics committee may approve the use of the data without individual consent, under the conditions laid down by law, if the risk is minimal and the data are pseudonymized or anonymized. Medical journals require ethics committee approval and the way consent was obtained to be stated in the methods section [36].

The ethics of data processing does not stop at protecting identity. It also covers the honesty of the analysis: "inconvenient" patients are not removed without a rule set in advance, dozens of hypotheses are not tested with only the favorable results reported, missing data are not hidden, and an association is not presented as a causal relationship. These rules, discussed at length in the literature on the ethics of biomedical informatics [37], will come up again in the chapters on statistical tests and clinical trials.

> **The rule that prevents the error**
>
> Before uploading a file with patient data to an online program, a form, an artificial intelligence tool or a storage platform, ask yourself three questions: (1) Does the file contain direct identifiers or combinations of variables that could identify a person? (2) Do I have approval (from the ethics committee, from the institution) to process these data in this place? (3) Have I kept the code correspondence list separate from the data, in a protected location? If the answer to any of these questions is "I don't know", do not upload the file. In the practical sessions you will use only the simulated mg2026 dataset and the teaching files on Moodle, which contain no real identifying data.

1.8.4 Routine Data in Research: Examples from Our Region

Many studies published by colleagues from our discipline use existing data originally collected for other purposes. A cross-sectional study of electronic prescriptions from urban areas of Timiș and Arad counties analyzed 1,498 prescriptions for older adults using the STOPP/START criteria [38]. Of the prescriptions, 57% were issued for women, the mean age of the patients was 74.1 ±

6.95 years, and the mean number of medicines per prescription was 4.7 ± 1.51; statins appeared on 47.4% of prescriptions. The prevalence of potentially inappropriate medications was 18.58%, and that of potential prescribing omissions 49.2%. Note that the statistical unit is the prescription, not the patient: if a patient received several prescriptions during the study period, that patient contributes several times, and this must be kept in mind when we read "57% women".

Another type of routine data comes from surveillance systems and from the registries of treatment centers. A hemodialysis study compared two cross-sectional cohorts from 10 Romanian centers: 1,401 patients in 2015 and 1,698 in 2019, before and after the national treatment program with direct-acting antivirals [39]. The prevalence of hepatitis C virus infection fell from 16.9% to 10.5% ($p < 0.0001$), whereas the prevalence of hepatitis B virus infection remained practically unchanged (4.7% versus 4.8%, $p = 0.604$), and one-year mortality did not differ significantly (9.9% versus 10.7%, $p = 0.46$).

Case study: surveillance data and their limitations

A retrospective study used surveillance data from the Timiș County Public Health Directorate for all confirmed COVID-19 cases among healthcare workers between January 1, 2021 and March 31, 2022 [40]. Incidence among auxiliary staff and nurses tracked the course in the community very closely, with Spearman correlation coefficients of 0.852 and 0.821 ($p < 0.001$), and vaccination coverage exceeded 70%. It is a fine example of making use of existing data at very low cost. What the result says: the waves of infection among healthcare workers coincided in time with the waves in the population. What it does not say: the correlation between two aggregated time series does not show where the workers became infected (in the hospital or in the community), and it is not proof of causality. It is an ecological, group-level association, which may be explained by the fact that both series depend on the same factor, the circulation of the virus. The hemodialysis study discussed above has a related limitation: the two cohorts, from 2015 and 2019, do not consist of the same patients, so the drop in HCV prevalence describes the dialysis population at the two time points, not the cure of individually followed patients. The cohorts also differed in age (59.7 ± 12.92 versus 61 ± 12.93 years, $p = 0.005$); a difference of 1.3 years is statistically significant in a sample of more than 3,000 patients but clinically negligible, a distinction we will discuss at length in Chapter 5. In both cases, routine data make it possible to address questions that would otherwise be impossible to tackle, provided the interpretation respects the way the data were collected.

1.9 Elements of Probability Theory

Inferential statistics rests on probability theory. You do not need rigorous proofs, but you must master a few simple rules, because they underlie p values, confidence intervals, relative risk and the predictive values of diagnostic tests [41,42]. All the examples in this section use the mg2026 dataset.

1.9.1 Random Experiment, Outcomes, Events

A random experiment is a process whose outcome cannot be predicted with certainty, although the set of possible outcomes is known. Rolling a die is the classic example. In medicine, a random experiment may be choosing a patient at random from a list, measuring a patient's blood glucose, the result of a rapid test, or a patient's course after treatment.

The set of all possible outcomes is called the sample space (or outcome space) and is denoted by Ω. If we choose a patient at random from mg2026 and record smoking status, Ω = {never smoker, former smoker, current smoker}. An event is a subset of the sample space, for example "the patient has ever smoked" = {former smoker, current smoker}. The certain event is Ω itself (it always occurs), and the impossible event is the empty set.

1.9.2 The Classical and the Frequentist Definitions

The classical definition of probability applies when all outcomes are equally likely:

P(A) = number of favorable cases / number of possible cases

If we choose a patient completely at random from the 160 in the mg2026 dataset, each patient has the same chance of being chosen, and the probability that the patient is a woman is $P(F) = \frac{86}{160} = 0.5375$. The probability that the patient has diabetes confirmed by OGTT is $P(D) = \frac{29}{160} = 0.18125$, that is, about 18.1%.

The frequentist (statistical) definition starts from observation: the probability of an event is the limit approached by the relative frequency of the event when the experiment is repeated a very large number of times under the same conditions. If 5,140 of 10,000 newborns in a maternity hospital are boys, we estimate the probability of the birth of a boy at 0.514. The law of large numbers guarantees that, as the number of repetitions increases, the relative frequency approaches the true probability. Over 10 births, the proportion of boys may be 0.3 or 0.7; over 10,000 it will be very close to the true value.

In medicine we almost always use the frequentist definition. When we say that the risk of diabetes in a population is 18%, we mean a relative frequency observed in a sample and used as an estimate of a probability. The sample proportion (p) estimates the population probability (π).

1.9.3 The Axioms of Probability and the Complementary Event

Every probability obeys three basic rules (Kolmogorov's axioms): (1) the probability of any event is a number between 0 and 1, $0 \leq P(A) \leq 1$; (2) the probability of the certain event is 1, $P(\Omega) = 1$; (3) if two events cannot occur at the same time (they are mutually exclusive), the probability that one of them occurs is the sum of their probabilities. All the other rules follow from these three.

The complementary event of A, denoted "non-A", occurs when A does not occur. Its probability is:

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

The probability that a patient chosen at random from mg2026 does not have diabetes is 1 - 0.18125 = 0.81875. The rule seems trivial, but it is extremely useful for "at least one" calculations: it is much simpler to calculate the probability that no event occurs and subtract it from 1.

1.9.4 The Addition Rule

For two mutually exclusive events (which cannot occur together), the probability that one or the other occurs is:

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

A patient cannot be both a current smoker and a former smoker. The probability that a patient chosen at random has ever smoked is therefore $P(\text{ smoker }) + P(\text{former smoker }) = \frac{53}{160} + \frac{33}{160} = \frac{86}{160} = 0.5375$.

If the events can occur together, the sum above would count twice the individuals who have both characteristics, and the formula becomes:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Of the 160 patients, 53 are smokers, 54 are obese (BMI ≥ 30 kg/m²), and 22 are both smokers and obese. The probability that a patient chosen at random is a smoker or obese (or both) is:

$$P(\text{ smoker or obese}) = \frac{53}{160} + \frac{54}{160} - \frac{22}{160} = \frac{85}{160} = 0.531$$

Had we simply added them, we would have obtained 107/160 = 0.669, that is, we would have counted the 22 obese smokers twice. In the exam, this is the most frequent mistake in problems involving the addition rule.

1.9.5 Conditional Probability

The conditional probability of event A, given that event B has occurred, is denoted P(A|B) and read "the probability of A given B". It is calculated by restricting the sample space to the individuals who have B:

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

This concept is probably the most important in the whole chapter, because relative risk, sensitivity, specificity and predictive values all rest on it. Let us apply it to Table 1.4, which cross-tabulates obesity (BMI ≥ 30 kg/m²) and diabetes confirmed by OGTT at enrollment in the mg2026 study.

Table 1.4. Obesity and diabetes confirmed by OGTT at enrollment, mg2026 dataset (n = 160).

	Diabetes (D)	No diabetes	Total
Obese (O), BMI ≥ 30	18	36	54
Non-obese, BMI < 30	11	95	106
Total	29	131	160

Several types of probability can be calculated from the table, and the difference between them must be understood very well.

Marginal probabilities concern a single variable and are calculated from the totals in the margins: $P(O) = \frac{54}{160} = 0.3375$; $P(D) = \frac{29}{160} = 0.181$.

The joint probability (of the intersection) concerns both characteristics at once and is calculated from the corresponding cell divided by the grand total: $P(O \text{ and } D) = \frac{18}{160} = 0.1125$.

Conditional probabilities are calculated by dividing the cell by the total of the row or column that expresses the condition. The probability of having diabetes if the patient is obese is $P(D \mid O) = \frac{18}{54} = 0.333 (33.3\%)$. The probability of having diabetes if the patient is not obese is $P(D \mid \text{non} - O) = \frac{11}{106} = 0.104 (10.4\%)$. This can also be checked with the formula: $P(D \mid O) = \frac{P(O \text{ and } D)}{P(O)} = \frac{0.1125}{0.3375} = 0.333$.

Note that P(D|O) is not the same as P(O|D). The probability of being obese if the patient has diabetes is $P(O \mid D) = \frac{18}{29} = 0.621 (62.1\%)$. The first probability answers the question "How many of the obese patients have diabetes?", the second the question "How many of the patients with diabetes are obese?" Confusing them, sometimes called the fallacy of the transposed conditional, is the source of much faulty reasoning, in medicine and elsewhere. In Chapter 11 you will see that the difference between the sensitivity of a test, P(positive test|disease), and its positive predictive value, P(disease|positive test), is exactly the difference between P(O|D) and P(D|O).

The multiplication rule, valid for any two events, follows from the formula for conditional probability:

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

For Table 1.4: $P(O \text{ and } D) = P(O) \times P(D \mid O) = 0.3375 \times 0.333 = 0.1125$, the same result as 18/160.

1.9.6 Independent Events

Two events are independent if the occurrence of one does not change the probability of the other, that is, P(A|B) = P(A). For independent events, the multiplication rule becomes simpler:

$P(A \text{ and } B) = P(A) \times P(B)$ (only for independent events)

This relationship also provides a way of checking independence. If obesity and diabetes were independent in mg2026, the joint probability should be $P(O) \times P(D) = 0.3375 \times 0.18125 = 0.0612$. The observed probability, however, is 0.1125, almost double. Equivalently, P(D|O) = 0.333 is very different from P(D) = 0.181 and from P(D|non-O) = 0.104. We conclude that, in this sample, the two events are not independent: obesity and diabetes occur together more often than they would by chance. Whether the observed difference can be explained by chance (by sampling variability) is a question for the chi-square test in Chapter 7; the answer, anticipated here, is no ($\chi^2 = 12.70$, $p < 0.001$).

An example of approximate independence from the same dataset: the probability of diabetes is 15/86 = 0.174 in women and 14/74 = 0.189 in men, both very close to the overall probability of 0.181. Sex and diabetes appear to be independent in mg2026.

Do not confuse independent events with mutually exclusive events. Two mutually exclusive events with nonzero probabilities cannot be independent: if we know that one has occurred, we know for certain that the other has not, so its probability has dropped to zero. "Current smoker" and "former smoker" are mutually exclusive and, precisely for that reason, strongly dependent.

The multiplication rule for independent events is the one we use when patients are chosen independently of one another. If the probability of diabetes in a patient is 0.18, the probability that two independently chosen patients both have diabetes is 0.18 x 0.18 = 0.0324, and the probability that none of ten patients has diabetes is 0.82 to the tenth power, that is, 0.137. Hence the probability that at least one of the ten has diabetes is 1 - 0.137 = 0.863.

1.9.7 The Law of Total Probability and a First Look at Bayes' Theorem

Sometimes we know the probability of an event in each of several subgroups that partition the population, and we want the overall probability. If the events B1, B2,..., Bk are mutually exclusive and together cover the whole population, then:

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

It is a weighted average of the conditional probabilities, with weights equal to the size of each subgroup. For Table 1.4:

$$P(D) = P(D \mid O) \times P(O) + P(D \mid \text{non} - O) \times P(\text{non} - O)$$
$$= 0.333 \times 0.3375 + 0.104 \times 0.6625 = 0.1125 + 0.0688 = 0.181$$

We obtain, as expected, 29/160. The formula becomes truly useful when we do not have the full table but only the probabilities by subgroup. Suppose a family doctor works with two lists: the practice in town has 60% of the patients, with a diabetes prevalence of 15%; the satellite practice in the rural commune has 40% of the patients, with a prevalence of 25%. The probability that any given patient of this doctor has diabetes is 0.15 x 0.60 + 0.25 x 0.40 = 0.09 + 0.10 = 0.19.

Bayes' theorem answers the reverse question: if we know that the patient has diabetes, what is the probability that the patient comes from the rural commune? Using the multiplication rule and the law of total probability:

$$P(\text{ pooled } \mid D) = \frac{P(D \mid \text{ pooled }) \times P(\text{ pooled })}{P(D)} = \frac{0.10}{0.19} = 0.526$$

Although only 40% of the patients are from the commune, more than half of the patients with diabetes (52.6%) come from there, because prevalence is higher in that subgroup. For the mg2026 table, the same theorem gives $P(O \mid D) = \frac{P(D|O) \times P(O)}{P(D)} = \frac{0.333 \times 0.3375}{0.181} = 0.621$, that is, the value calculated directly above. Bayesian logic, which updates an initial (prior) probability on the basis of new information and yields a final (posterior) probability, is the foundation of the interpretation of diagnostic tests and will be covered at length in Chapter 11.

1.9.8 Probability Distributions: A First Look

A random variable is a variable whose value depends on a random experiment, for example the number of patients with diabetes among ten patients chosen at random. Its probability distribution tells us what values it can take and with what probability. Distributions are divided into discrete ones (for counts: 0, 1, 2,...) and continuous ones (for measurements such as blood pressure or blood glucose). The most important continuous distribution, the normal distribution, is the subject of Chapter 3. Here we present two discrete distributions that you will meet often.

The binomial distribution describes the number of "successes" in n independent trials, each with the same probability of success p. The word "success" is conventional and simply denotes the event being counted, even if that event is a disease. The conditions are: a fixed number of trials, two possible outcomes in each trial, a constant probability, and independence between trials. The probability of obtaining exactly k successes is:

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

where $C(n,\ k) = \frac{n!}{k! \times (n-k)!}$ is the number of ways in which the k successes can be chosen among the n trials (combinations of n taken k at a time), and p^k means p to the power k.

Let us take an example straight from our scenario. The proportion of patients with confirmed diabetes in mg2026 is about 18%. If a doctor assesses 10 independently chosen patients with a similar profile by OGTT, what is the probability that exactly 2 of them have diabetes? We have n = 10, p = 0.18, k = 2:

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

$$P(X = 2) = 45 \times 0.18^2 \times 0.82^8 = 45 \times 0.0324 \times 0.2044 = 0.298$$

Table 1.5 gives the whole distribution, and Figure 1.3A shows it graphically. The mean of the binomial distribution is n x p = 10 x 0.18 = 1.8 patients, and the standard deviation is $\sqrt{[n \times p \times (1-p)]} = \sqrt{10 \times 0.18 \times 0.82} = \sqrt{1.476} = 1.21$.

Table 1.5. Binomial distribution of the number of patients with diabetes out of 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

We can read the answers to practical questions directly from the table. The probability that no patient has diabetes is 0.137; the probability that at least one has diabetes is 1 - 0.137 = 0.863; the probability that at least three have diabetes is $1 - P(X \leq 2) = 1 - 0.737 = 0.263$. If the doctor finds 6 patients with diabetes out of 10, that result would be very unlikely (0.003) under the hypothesis that p = 0.18, which suggests that this doctor's patients are at higher risk than those in

mg2026. This line of reasoning, "how probable would the observed result be if our hypothesis were true", is exactly the logic of the statistical tests in Chapter 5.

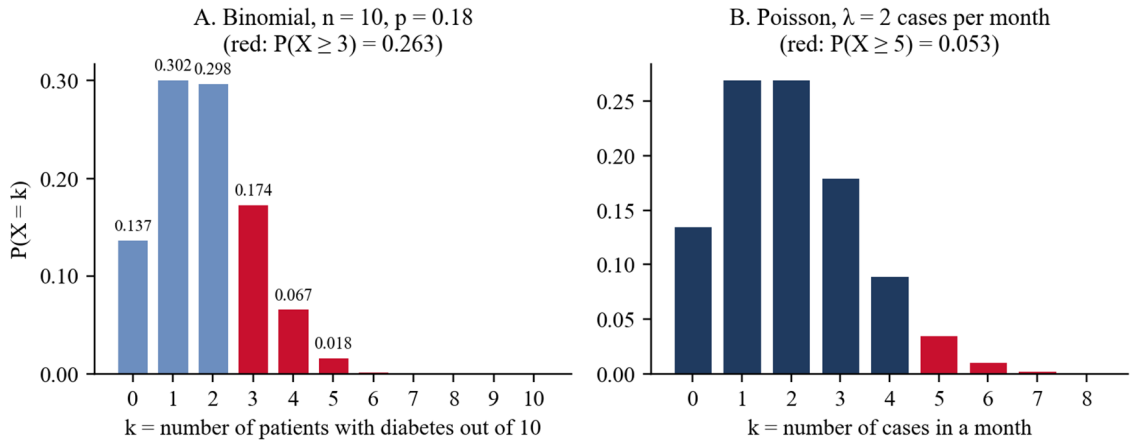


Figure 1.3. Two discrete distributions. A: binomial distribution of the number of patients with diabetes among 10 patients assessed, when the individual probability is 0.18; the red bars correspond to the event "at least 3 patients with diabetes" (total probability 0.263). B: Poisson distribution for a rare event with a mean of 2 cases per month; the red bars correspond to the event "at least 5 cases" (probability 0.053). Theoretical calculation.

The Poisson distribution describes the number of rare events that occur in a given interval of time or region of space, when the events occur independently of one another at a constant mean rate λ (lambda). Typical examples are the number of new cases of a rare disease reported monthly in a county, the number of serious adverse reactions to a vaccine per 100,000 doses, or the number of emergency admissions during one night at a small hospital. The probability of observing exactly k events is:

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

where e ≈ 2.718 is the base of natural logarithms. The particular feature of the Poisson distribution is that its mean and variance are equal, both being λ. If a public health directorate receives on average 2 reports a month of a rare disease, the probability of a month with no cases at all is $e^{-2} = 0.135$, and the probability of at least 5 cases in a month is 0.053 (Figure 1.3B). A month with 5 cases does not prove an outbreak, but it is rare enough (about one month in 19) to justify a check. This is, in essence, how epidemiological alert systems work.

The Poisson distribution is also a good approximation of the binomial distribution when n is large and p very small, with λ = n x p. For an adverse reaction with a probability of 1/10,000 in 20,000 treated patients, the number of reactions approximately follows a Poisson distribution with λ = 2.

1.10 Software Used in This Course

Hand calculations are indispensable for understanding what a method does, but nobody calculates a standard deviation of 160 values by hand. The course syllabus provides for a few programs, chosen because they are free or have a free version, run on the laboratory computers and cover the methods in the course. Each has its own role, and you will use all of them in the practical sessions.

Google Sheets is the spreadsheet in the Google suite, accessible from the browser with the university account. We use it for entering and organizing data, for validation rules, for data cleaning, for descriptive calculations with functions such as AVERAGE, MEDIAN, STDEV and COUNTIF, for pivot tables and for simple charts [43,44]. It is the tool closest to the spreadsheets you will encounter in any workplace. With the Romanian locale settings, decimals are written

with a comma and function arguments are separated by semicolons (with an English locale, as in the examples in this book, the decimal point and the comma are used).

Google Forms makes it possible to build a questionnaire or a data collection form quickly, with question types that correspond to the types of variables and with response validation rules; the responses are saved automatically in a Google Sheets spreadsheet [45,46]. You will use it in Practical Session 2 and in the session on designing clinical studies.

JASP is a free, open-source statistical program developed at the University of Amsterdam [47]. It has a menu-driven interface, results appear immediately in tables formatted in academic style, and analyses update automatically when options change. It covers descriptive statistics, parametric and nonparametric tests, correlation, regression and much more, including Bayesian versions of the tests. Dedicated guides are available for students, for example for medical statistics [48,49].

MedCalc is a statistical program oriented toward medical research, widely used for ROC curves, diagnostic test analysis, method comparison studies and meta-analysis [50]. The full version is commercial, but the manufacturer's website offers free online calculators for many of the problems in the course: confidence intervals for proportions, relative risk, odds ratio, sensitivity and specificity, number needed to treat. You will use them mainly in Chapters 10-12.

Epi Info is developed by the US Centers for Disease Control and Prevention and is intended for field epidemiology: building forms, entering data, frequency tables, 2x2 tables with relative risk and odds ratio, maps [51]. Its advantage is that it also works without an internet connection, which makes it useful in epidemiological investigations conducted in remote locations. The latest version is 7.2.7; the CDC stopped development and support of the program in September 2025, so we use it for classic epidemiology exercises, knowing that it will no longer receive updates.

Table 1.6. The programs in the course syllabus and their main use.

Program	Type	Main use in the course
Google Sheets	online spreadsheet	data entry, validation and cleaning; descriptive statistics; pivot tables; charts
Google Forms	online forms	questionnaires and data collection forms
JASP	free statistical program	descriptive statistics, normality, t-tests, ANOVA, nonparametric tests, correlation, regression
MedCalc	medical statistical program, online calculators	ROC, diagnostic tests, relative risk, odds ratio, confidence intervals
Epi Info	epidemiology program, works offline	forms, frequencies, 2x2 tables, risk analysis

Course materials, data files and assignments are available on the university's Moodle platform (Emediqual), where submitted work is also uploaded [52]. A final practical tip: whatever the program, always check the number of observations (n) reported in each results table. If you expect 160 and the program reports 158, two values are missing or have been read as text, and this must be cleared up before any interpretation.

1.11 Worked Problems

Worked Problem 1.1. Smoking and Obesity in the mg2026 Dataset

Statement. Table 1.7 cross-tabulates smoking status and obesity (BMI $\geq$ 30 kg/m²) at enrollment for the 160 patients of the mg2026 dataset. A patient is chosen at random. Calculate: (a) the probability that the patient is a smoker; (b) the probability that the patient is a smoker and obese;

(c) the probability that the patient is a smoker or obese; (d) the probability of obesity among current smokers, former smokers and never smokers; (e) check whether the events "smoker" and "obese" are independent.

Table 1.7. Smoking status and obesity at enrollment, mg2026 dataset.

	Obese	Non-obese	Total
Current smokers	22	31	53
Former smokers	9	24	33
Never smokers	23	51	74
Total	54	106	160

Solution. All patients have the same probability of being chosen, so we use the classical definition, with 160 possible cases.

(a) $P(\text{ smoker }) = \frac{53}{160} = 0.331$.

(b) The joint probability is read from the cell at the intersection: $P(\text{ smoker and obese}) = \frac{22}{160} = 0.1375$.

(c) The events are not mutually exclusive (there are 22 patients who have both characteristics), so we use the general addition rule: $P(\text{ smoker or obese}) = \frac{53}{160} + \frac{54}{160} - \frac{22}{160} = \frac{85}{160} = 0.531$.

(d) The conditional probabilities are calculated by rows: $P(\text{obese} \mid \text{ smoker }) = \frac{22}{53} = 0.415$; $P(\text{obese} \mid \text{former smoker}) = \frac{9}{33} = 0.273$; $P(\text{obese} \mid \text{ non-smoker}) = \frac{23}{74} = 0.311$.

(e) If the two events were independent, we would have $P(\text{ smoker and obese}) = P(\text{ smoker }) \times P(\text{obese}) = 0.331 \times 0.3375 = 0.112$. The observed value is 0.1375, somewhat higher. Equivalently, P(obese|smoker) = 0.415 is higher than P(obese) = 0.3375. The events are not perfectly independent in the sample.

Interpretation. In the mg2026 dataset, obesity is somewhat more frequent among current smokers (41.5%) than among never smokers (31.1%) and former smokers (27.3%). The differences, however, are much smaller than in the case of obesity and diabetes, and in a sample of 160 patients they may also be produced by chance. The fact that P(A and B) differs from P(A) x P(B) in a sample does not prove dependence in the population; for that we need a statistical test (Chapter 7). From a clinical point of view, the association between smoking and obesity is not a direct causal one; both may be linked to other lifestyle factors.

Worked Problem 1.2. Binomial Probability: Hypertensive Patients in the Waiting Room

Statement. In the mg2026 dataset, 41 of the 160 patients (25.6%) have a systolic blood pressure ≥ 140 mmHg at enrollment. Assume that the probability that a patient with a similar profile has SBP ≥ 140 mmHg is p = 0.256 and that patients are independent. There are 5 such patients in the waiting room of the practice. What is the probability that: (a) none has SBP ≥ 140; (b) at least one has SBP ≥ 140; (c) exactly two have SBP ≥ 140; (d) at least two have SBP ≥ 140?

Solution. We check the conditions of the binomial distribution: a fixed number of patients (n = 5), two possible outcomes for each (SBP ≥ 140 or not), a constant probability (0.256), independence (assumed). They are met.

(a) $P(X = 0) = (1 - 0.256)^5 = 0.744^5$. We calculate step by step: 0.744² = 0.5535; 0.744^4 = 0.5535² = 0.3064; $0.744^5 = 0.3064 \times 0.744 = 0.228$.

(b) We use the complementary event: $P(X \geq 1) = 1 - P(X = 0) = 1 - 0.228 = 0.772$.

(c) $C(5,\ 2) = \frac{5\times4}{2} = 10$. $P(X = 2) = 10 \times 0.256^2 \times 0.744^3 = 10 \times 0.0655 \times 0.4118 = 0.270$.

(d) We also need $P(X = 1) = 5 \times 0.256 \times 0.744^4 = 5 \times 0.256 \times 0.3064 = 0.392$. Then $P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - 0.228 - 0.392 = 0.380$.

Interpretation. With a probability of about 77%, at least one of the five patients will have a systolic blood pressure of at least 140 mmHg when measured, and with a probability of 38%, at least two will. In practice, on an ordinary day, the doctor should expect raised values in almost every group of five patients at cardiometabolic risk. Caution: a single measurement ≥ 140 mmHg does not mean a diagnosis of hypertension, which requires repeated measurements according to a standardized protocol.

Worked Problem 1.3. Designing a Sample in a Family Practice

Statement. A family practice has 1,600 adults aged 40 to 75 on its list: 1,000 from the town and 600 from the surrounding villages it serves. The doctor wants to assess diabetes risk with the FINDRISC score in a sample of 80 patients. (a) Describe a systematic sampling scheme. (b) Describe a stratified sampling scheme proportional to area of residence. (c) What would happen if the doctor included the first 80 patients who come to the practice in January?

Solution. (a) The sampling interval is k = N/n = 1,600/80 = 20. A number between 1 and 20 is drawn at random, for example 13 (in Google Sheets, =RANDBETWEEN(1, 20)). The patients with serial numbers 13, 33, 53, 73 and so on are included, that is, 13 + 20 x (i - 1) for i = 1,..., 80. The last patient included is number 13 + 20 x 79 = 1,593. Before applying the scheme, the doctor must check that the list is not ordered by some periodic criterion.

(b) The proportions in the population are 1,000/1,600 = 62.5% for the town and 600/1,600 = 37.5% for the villages. The sample of 80 will include 80 x 0.625 = 50 patients from the town and 80 x 0.375 = 30 from the villages, chosen by simple random sampling within each stratum. The sampling fraction is the same in both strata: 50/1,000 = 30/600 = 5%.

(c) The first 80 patients who come to the practice form a convenience sample. Patients who see the doctor often are generally sicker or older (more chronic diseases, monthly prescriptions), and in January respiratory infections predominate. Healthy patients, who have not come to the practice for months, would be underrepresented. The prevalence of high risk would most likely be overestimated. This is selection bias.

Interpretation. Stratified sampling guarantees that village residents, who have more difficult access to the practice and perhaps a different risk profile, are present in exactly their proportion in the population. The systematic version is simpler, but the proportion of village residents in the sample could vary by chance around 37.5%.

Worked Problem 1.4. Recognizing Systematic Errors

Statement. For each of the situations below, identify the type of systematic error and propose a preventive measure. (1) In a case-control study on the use of nonsteroidal anti-inflammatory drugs and gastrointestinal bleeding, the cases are interviewed in hospital a few days after the episode, while the controls are interviewed by telephone. (2) A diabetes education program is evaluated by comparing program participants with patients who declined to take part. (3) A study shows that people who carry a lighter have coronary heart disease more often. (4) The nurse who measures blood pressure knows which patients are in the teleconsultation arm and rounds the values to the nearest value ending in 0.

Solution. (1) Information bias, more precisely recall bias: the cases, preoccupied with the cause of their illness, remember the medicines they took better than the controls; in addition, the

different interview mode (face to face versus telephone) can produce differences in reporting. Prevention: the same interview mode for cases and controls, a structured questionnaire, verification of drug use in the medical record or in the prescription database.

(2) Selection bias (self-selection, the healthy volunteer effect): patients who accept an educational program are usually more motivated and more adherent to treatment. The difference in glycemic control could be explained by motivation, not by the program. Prevention: randomization of eligible patients to a program group and a no-program group.

(3) Confounding: carrying a lighter is associated with smoking, and smoking causes coronary heart disease. The lighter has no effect of its own. Prevention: stratification by smoking status or adjustment for smoking in a multivariable model.

(4) Information bias: observer bias (knowing the arm may involuntarily influence the measurement) and terminal digit preference for zero. Prevention: assessors who do not know the patient's arm (blinding), automated blood pressure monitors, a standardized measurement protocol with recording of the exact value displayed.

Interpretation. Selection bias concerns who enters the study, information bias concerns how the variables are measured, and confounding concerns the interpretation of an association in the presence of a third variable. All three shift the result systematically and do not disappear when the sample is enlarged.

Worked Problem 1.5. Total Probability and Bayes' Theorem with Three Practices

Statement. A network of three family practices takes part in a diabetes screening program. Practice 1 contributes 50% of the patients assessed, practice 2 contributes 30% and practice 3 contributes 20%. The proportion of patients with previously undiagnosed diabetes is 12% in practice 1, 20% in practice 2 and 30% in practice 3. (a) What is the probability that any given patient in the program has diabetes? (b) A patient has confirmed diabetes; what is the probability that this patient comes from practice 3?

Solution. Let C1, C2, C3 denote the events "the patient comes from practice 1, 2, 3"; they are mutually exclusive and cover the whole program population.

(a) Law of total probability:

$$P(D) = P(D \mid C_1) \times P(C_1) + P(D \mid C_2) \times P(C_2) + P(D \mid C_3) \times P(C_3)\, P(D)$$
$$= 0.12 \times 0.50 + 0.20 \times 0.30 + 0.30 \times 0.20 = 0.06 + 0.06 + 0.06 = 0.18$$

(b) Bayes' theorem:

$$P(C_3 \mid D) = \frac{P(D \mid C_3) \times P(C_3)}{P(D)} = \frac{0.06}{0.18} = 0.333$$

Interpretation. The overall prevalence of diabetes in the program is 18%, a weighted average of the three prevalences, not their simple average (which $\frac{\text{ar fi} \quad (12+20+30)}{3} = 20.7\%$). Although practice 3 contributes only 20% of the patients, it yields one third of the diabetes cases, because its prevalence is the highest. Note the symmetry: each practice contributes 6 percentage points to the overall prevalence, so each yields one third of the cases. Planning the resources for confirming the diagnosis (OGTT, consultations) should take this distribution into account, not the number of patients assessed in each practice.

Worked Problem 1.6. The Poisson Distribution in an Alert System

Statement. The public health directorate of a county receives, on average, 3 reports a month of viral hepatitis A. Assume that the reports are independent and occur at a constant rate. (a) What is

the probability that no case is reported in a month? (b) What is the probability of exactly 3 cases? (c) In one month 7 cases are reported; how unusual is this number?

Solution. The monthly number of cases approximately follows a Poisson distribution with λ = 3.

(a) $P(X = 0) = \frac{e^{-3} \times 3^0}{0!} = e^{-3} = 0.0498$. About one month in 20 will have no cases.

(b) $P(X = 3) = \frac{e^{-3} \times 3^3}{3!} = \frac{0.0498 \times 27}{6} = 0.224$.

(c) We calculate the probability of at least 7 cases using the complementary event. The probabilities for k = 0, 1,..., 6 are 0.0498, 0.1494, 0.2240, 0.2240, 0.1680, 0.1008 and 0.0504, which sum to 0.966. Therefore $P(X \geq 7) = 1 - 0.9665 = 0.034$.

Interpretation. A month with 7 or more cases occurs, under ordinary conditions, with a probability of about 3.4%, that is, less often than once every two years. The result does not prove the existence of an outbreak, but it is unusual enough to justify an epidemiological investigation: do the cases have a common source, do they come from the same locality, from the same group of children? This is, in principle, how alert thresholds are set in the surveillance of communicable diseases.

1.12 Proposed Exercises

1. In the mg2026 dataset, 19 of the 95 urban patients and 10 of the 65 rural patients have diabetes confirmed by OGTT. Construct the 2x2 table and calculate: P(D|urban), P(D|rural), P(urban and D), P(urban) x P(D). Are the events "urban" and "diabetes" independent in the sample?
2. In mg2026 there are 86 women and 53 smokers, and 30 of the women are smokers. Calculate P(smoker|woman), P(woman|smoker) and P(woman or smoker). Explain why the first two probabilities are different.
3. The prevalence of arterial hypertension in a population of adults over 40 is 25%. A total of 8 people are chosen at random. Calculate the probability that: (a) none is hypertensive; (b) exactly one is hypertensive; (c) at least two are hypertensive.
4. A hospital has 2,400 patients with heart failure from the last three years in its discharge register. A systematic sample of 120 records is wanted. What is the sampling interval? If the random starting number drawn is 9, write down the first five record numbers included and the last one.
5. A practice has 1,500 adults on its list: 600 aged 18 to 39, 600 aged 40 to 64 and 300 aged 65 and over. How many patients from each age group must be included in a proportional stratified sample of 90 patients? What is the sampling fraction?
6. Classify the systematic error (selection, information or confounding) in each situation: (a) a questionnaire on alcohol consumption filled in in the presence of the family doctor; (b) a study on the effect of vitamin D on fractures, in which the people who take supplements are also those who exercise more; (c) a study on the efficacy of a COPD treatment in which 30% of the patients in the treated group drop out because of adverse effects and are no longer assessed; (d) a new test for prostate cancer evaluated in patients with metastases and in healthy student volunteers.
7. Read the description of the study on prescribing in older adults in Section 1.8.4 [38]. State: the statistical unit, the sample size, the likely target population, the type of study, one numerical variable and one qualitative variable. What problem may arise if the same patient has several prescriptions in the database?

8. A pediatric emergency department receives, on average, 1.5 cases a week of accidental medication poisoning. Assuming a Poisson distribution, calculate the probability that there are no cases in a week and the probability of at least 4 cases.
9. In a practice, 30% of adults are smokers. The prevalence of COPD is 15% in smokers and 3% in nonsmokers. Calculate the prevalence of COPD in the practice population and the probability that a patient with COPD is a smoker.
10.

    (a) Formulate in PICO format the question "Does teleconsultation help lower blood pressure?", using the mg2026 scenario and the variables tas0 and tas6. (b) A file contains the columns: patient code (P001...), age in years, sex, BMI, blood glucose; the list linking the codes to the patients' names is kept by the investigating physician in a separate, protected folder. Are the data anonymized or pseudonymized? What legal regime applies to them?

Common Pitfalls

- Confusing P(A|B) with P(B|A): the probability of diabetes among obese patients (33.3%) is not the probability of obesity among patients with diabetes (62.1%).
- Applying the simple addition rule, P(A) + P(B), to events that can occur together; the result may even exceed 1.
- Confusing independent events with mutually exclusive events.
- Believing that a large sample is automatically representative; a large size reduces random error, not systematic error.
- Using Greek letters (μ, σ) for values calculated on a sample; in a sample we write m and s.
- Correcting an outlying value "by intuition" instead of checking it against the source document.
- Considering a file anonymous just because the patients' names have been deleted.

Key Points

- Biological variability is the reason medicine needs statistics; it is described, not eliminated.
- The population is defined in space and time; the sample is a part of it, chosen so as to be representative (equal probability and independence).
- Simple random, systematic, stratified and cluster sampling are probability methods; the convenience sample is common but exposed to selection bias.
- Systematic errors (selection, information, confounding) do not decrease as the sample grows; they are prevented by design.
- Descriptive statistics summarizes the sample; inferential statistics generalizes to the population, with quantified uncertainty.
- Population parameters (μ, σ, π) are unknown; sample statistics (m, s, p_i) estimate them.
- Conditional probability $P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$ underlies relative risk and diagnostic tests.
- Independent events satisfy $P(A \text{ and } B) = P(A) \times P(B)$; checking this in a sample does not replace a statistical test.
- Health data are special-category data (GDPR, Law 190/2018); pseudonymized data remain personal data; research on human beings requires ethics committee approval and, as a rule, informed consent.

Self-Assessment Questions

1. (SC) In a study analyzing 1,498 electronic prescriptions issued for older adults, the statistical unit is:

    a) the older patient

   b) the prescription
   c) the pharmacy
   d) the prescribing physician
   e) the county

2. (SC) The correct notation for the mean of a variable in the population is:
   a) m
   b) s
   c) μ
   d) p
   e) σ

3. (SC) In Table 1.4 (obesity and diabetes, mg2026), the probability that an obese patient has diabetes is:
   a) 0,113
   b) 0,181
   c) 0,333
   d) 0,621
   e) 0,104

4. (SC) For a systematic sample of 60 patients from a list of 900, the sampling interval is:
   a) 60
   b) 15
   c) 9
   d) 6
   e) 90

5. (SC) If the probability that a patient has diabetes is 0.18 and patients are independent, the probability that none of 10 patients has diabetes is approximately:
   a) 0,18
   b) 0,82
   c) 0,137
   d) 0,863
   e) 0,298

6. (MC) Which of the following statements about data pseudonymization are correct?
   a) it replaces direct identifiers with codes
   b) pseudonymized data remain personal data within the meaning of the GDPR
   c) it is irreversible, and the patient can no longer be re-identified in any way
   d) the correspondence list between codes and identities is kept separately from the data
   e) it exempts the study from ethics committee approval

7. (MC) Which of the following statements about systematic errors are correct?
   a) they decrease as the sample size increases
   b) recall bias is a form of information bias
   c) randomization balances known and unknown confounders between groups
   d) the healthy volunteer effect is a form of selection bias
   e) a narrow confidence interval guarantees the absence of systematic errors

8. (MC) Let A and B be two mutually exclusive events, with P(A) > 0 and P(B) > 0. Which of the following statements are correct?
    a) $P(A \text{ and } B) = 0$
    b) $P(A \text{ or } B) = P(A) + P(B)$
    c) A and B are independent
    d) $P(A \mid B) = 0$
    e) P(A) + P(B) may be greater than 1
9. (MC) Inferential statistics:
    a) estimates population parameters with confidence intervals
    b) tests hypotheses about the population
    c) refers exclusively to the individuals in the sample
    d) rests on probability theory
    e) consists of building frequency tables
10. (MC) Which of the following statements about the convenience sample are correct?

a) it consists of the individuals who are easiest to recruit
b) it guarantees representativeness if its size exceeds 500
c) it can be used in a randomized trial, where the validity of the comparison between arms comes from randomization
d) an online survey shared on social media is an example
e) it eliminates selection bias

Chapter 2. Describing and Organizing Data: Types of Variables, Descriptive Measures and Graphical Displays

Objectives

By the end of this chapter, the student will be able to:

- classify a variable by its type (nominal, ordinal, discrete or continuous numerical) and by its measurement scale (nominal, ordinal, interval, ratio);
- organize the data of a study in a table with one row for each patient and one column for each variable, accompanied by a data dictionary, and recognize common data entry errors;
- construct a table of absolute, relative and cumulative frequencies, including for data grouped into classes, and reconstruct the missing values in such a table;
- calculate, by hand and with software, the mean, median, mode, weighted mean and geometric mean, and choose the measure appropriate to the type of data and the shape of the distribution;
- calculate the range, mean absolute deviation, variance, standard deviation, coefficient of variation and interquartile range;
- explain the difference between the standard deviation and the standard error of the mean;
- determine quartiles, percentiles and the five-number summary, and identify outliers with the 1.5 x IQR rule;
- choose the appropriate graph for each type of variable and recognize misleading graphs;
- read and construct a "Table 1"-type descriptive table from a medical article.

2.1 From the Patient to the Data Table

A patient who walks into the practice is not a string of numbers. To be analyzed statistically, however, the patient must be described by a set of measured or observed characteristics, the same for all patients in the study. Each characteristic is called a variable, because it takes different values from one patient to another. Age, sex, blood pressure, smoking status, educational level and satisfaction with treatment are variables. The value of a variable in a particular patient is called an observation.

The type of a variable determines almost everything that follows: which descriptive measures can be calculated, which graph is suitable, which statistical test can be applied. This is why the first step of any analysis, in the exam as in research, is the correct classification of the variables. Students who skip this step end up calculating the mean of sex or the standard deviation of the treatment arm, results that the software displays without protest but that make no sense at all.

2.1.1 Qualitative and Quantitative Variables

Figure 2.1 summarizes the basic classification. Qualitative (categorical) variables place each individual in a category. Quantitative (numerical) variables express a magnitude as a number, usually accompanied by a unit of measurement.

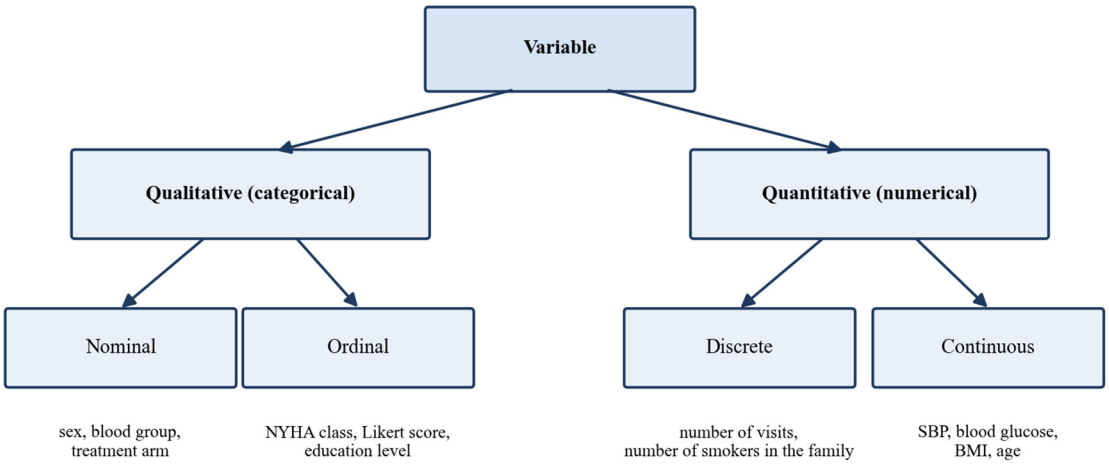


Figure 2.1. Classification of statistical variables. Qualitative variables may be nominal (categories without order) or ordinal (ordered categories); quantitative variables may be discrete (counts) or continuous (measurements). The examples come from general medicine and from the mg2026 dataset.

Nominal variables have categories with no natural order between them: sex (F, M), blood group (0, A, B, AB), treatment arm (A, B, C), smoking status (never smoker, former smoker, current smoker), area of residence (urban, rural). Even if we code them with numbers, for example 1 for women and 2 for men, the numbers are only labels. The mean of the codes makes no sense. A nominal variable with only two categories is called dichotomous (or binary): confirmed diabetes yes/no, dead yes/no, weight loss of at least 5% yes/no.

Ordinal variables have ordered categories, but the distances between them are not equal or are unknown. NYHA classes of heart failure, liver fibrosis stages F0-F4, steatosis grades S0-S3, educational level, Likert scales ("strongly disagree"... "strongly agree") and the 0 to 10 pain score are ordinal. We know that NYHA class III is more severe than class II, but we cannot say that the difference between I and II equals the difference between III and IV. In the mg2026 dataset, education (1-4), self-rated physical activity (1-5) and satisfaction (1-5) are ordinal.

Discrete quantitative variables take isolated values, usually whole numbers resulting from counting: the number of consultations per year, the number of medicines per prescription, the number of pregnancies, the number of decayed teeth. Continuous quantitative variables can take, at least in theory, any value within an interval, limited only by the precision of the instrument: blood pressure, blood glucose, weight, height, age, the concentration of a drug. The fact that a continuous variable is rounded to a whole number (an SBP of 131 mmHg, an age of 56 years) does not make it discrete.

2.1.2 Measurement Scales

A finer classification, widely used in methodology textbooks, distinguishes four measurement scales, in increasing order of the information they contain [19,53].

The nominal scale only allows us to check equality: two patients either have or do not have the same blood group. The ordinal scale also allows ordering: one patient has a more advanced stage than another. The interval scale also allows differences to be compared: the difference between 36 °C and 38 °C equals the difference between 38 °C and 40 °C. Its zero, however, is conventional, so ratios make no sense; a temperature of 40 °C is not "twice as high" as 20 °C. The ratio scale has an absolute zero, which means the absence of the measured quantity, and allows all operations, including ratios: a blood glucose of 200 mg/dL is double one of 100 mg/dL, and a patient weighing 90 kg weighs 1.5 times as much as one weighing 60 kg.

Most quantitative variables in medicine are on a ratio scale: weight, height, blood glucose, cholesterol, blood pressure, heart rate, length of hospital stay. The interval-ratio distinction rarely matters in practice for the choice of test, but it has a consequence you will meet in Section 2.4: the coefficient of variation makes sense only for variables on a ratio scale.

Table 2.1. Measurement scales and permitted operations.

Scale	Meaningful operations	Examples	Appropriate measure of central tendency
Nominal	=, ≠	sex, blood group, treatment arm, smoking	mode, proportions
Ordinal	=, ≠, <, >	NYHA class, education, Likert scale, pain score	median, mode
Interval	in addition, differences	temperature in °C, year of birth	mean, median
Ratio	in addition, ratios	blood glucose, weight, SBP, age, length of hospital stay	mean, median, geometric mean

2.1.3 Derived Variables and Scores

Many variables in an analysis are not measured directly but calculated from others. Body mass index is the typical example:

BMI = weight (kg) / [height (m)]²

For patient P003 in mg2026, with a weight of 127.4 kg and a height of 183 cm (1.83 m), $\text{BMI} = \frac{127.4}{1.83^2} = \frac{127.4}{3.3489} = \frac{38.0\ \text{kg}}{m^2}$. Other common derived variables are age (calculated from the date of birth and the date of enrollment), estimated glomerular filtration rate (from creatinine, age and sex), the change in a variable over time (glicemie0 - glicemie6) or the percentage weight loss.

Derived variables are calculated in the spreadsheet with a formula, not entered by hand. If a height is entered incorrectly, the formula will produce an absurd BMI that is easy to spot; if the BMI were typed in separately, the error would stay hidden. In Practical Session 1 you will see that a height of 1.72 entered in a column in centimeters gives, when BMI is recalculated with the formula, a value of more than 260,000 kg/m², which immediately catches the eye. Conversely, a derived variable kept in the file can help recover an incorrect value: height can be recalculated from BMI and weight.

Clinical scores (FINDRISC, the Apgar score, the Glasgow score, cardiovascular risk scores) are sums of points awarded for different characteristics. Formally they are discrete variables with a large number of possible values; in practice they are often treated as numerical, but with caution, because one extra point does not have the same clinical meaning in all parts of the scale. The FINDRISC score in mg2026 takes values between 0 and 25, with a mean of 10.37 and a median of 10.

2.1.4 Dichotomization and Its Cost

Physicians love thresholds: hypertensive or not (140 mmHg), obese or not (BMI 30 kg/m²), diabetic or not (fasting plasma glucose 126 mg/dL). Turning a continuous variable into a dichotomous one is called dichotomization. It is useful for clinical decisions and for communicating with the patient, but in statistical analysis it comes at a cost.

The first cost is loss of information. After dichotomization, a patient with a BMI of 30.0 and one with a BMI of 48.3 are equally "obese", while a patient with a BMI of 29.9 is "non-obese", although differing from the first by only 0.1 kg/m². In the mg2026 dataset, mean systolic blood pressure is 136.87 mmHg in the 54 obese patients and 128.21 mmHg in the 106 non-obese

patients, but this comparison ignores the fact that, among obese patients, SBP continues to rise with BMI. The regression in Chapter 9 will show that each additional kg/m² is associated, on average, with an additional 1.435 mmHg, information that dichotomization throws away. The second cost is a loss of statistical power: more patients are needed to detect the same association with a dichotomized variable. The third is the arbitrariness of the threshold: if the threshold is not set in advance, the researcher may be tempted to choose the one that gives the "nicest" result.

A good example of an analysis that keeps both perspectives comes from the prospective study of 921 students [5]. The change in BMI was first analyzed as a continuous variable (mean -0.45 ± 1.07 kg/m²) with linear regression; then, separately, clinically relevant weight gain was defined, which occurred in 7.2% of the cohort, and analyzed with logistic regression. The first analysis uses all the information; the second answers a clinical question ("who gains a significant amount of weight?"). The practical conclusion: keep the continuous variable for the main analysis and use established clinical thresholds, set in advance, for secondary analyses.

2.1.5 Coding and the Data Table

The data of a study are organized in a rectangular table according to a simple rule: one row for each statistical unit (patient) and one column for each variable. The first row contains the names of the variables. Each cell contains a single value. All statistical programs (JASP, MedCalc, Epi Info) assume this structure, sometimes called the "long" or "tidy" data format.

A few practical rules prevent most problems. Variable names are short, with no spaces, no diacritics and no special characters (tas0, not "initial systolic BP (mmHg)"); the full description goes in the data dictionary. A column contains a single type of data: in the blood glucose column only numbers are written, without the unit ("87", not "87 mg/dl"). Categories are always written the same way ("F", not "F" one time and "f" or "Feminin" another). Missing values are left blank or marked with a single code declared in the dictionary; 0 is never written for a missing value, because 0 is a real value for many variables. Repeated measurements (blood glucose at 0, 3 and 6 months) go in separate columns, with names that indicate the time point.

Table 2.2 shows an excerpt from the data dictionary of the mg2026 dataset. The full dictionary is in Appendix C.

Table 2.2. Excerpt from the data dictionary of the mg2026 dataset.

Name	Description	Type	Unit / codes	Plausible range
id	patient code	text	P001-P160	unique
sex	sex	nominal	F, M	-
varsta	age at enrollment	integer numeric	years	18-90
educatie	educational level	ordinal	1 lower secondary; 2 high school or vocational; 3 university; 4 postgraduate	1-4
inaltime	height	numeric	cm	130-210
tas0	systolic blood pressure at enrollment	integer numeric	mmHg	70-260
glicemie0	fasting plasma glucose at enrollment	numeric	mg/dL	40-400
hba1c0	glycated hemoglobin	numeric, one decimal place	%	3,5-15
diabet_ogtt	diabetes confirmed by OGTT	dichotomous	0 no, 1 yes	0, 1
satisfactie	satisfaction at 6	ordinal	1 very dissatisfied...	1-5

Name	Description	Type	Unit / codes	Plausible range
	months		5 very satisfied	

2.1.6 Common Data Entry Errors

The file mg2026_brut.csv deliberately contains some of the errors you will encounter in any real database. We present them here by type, and in Practical Session 1 you will find and correct them.

Table 2.3. Types of data entry errors, with examples from mg2026_brut.csv.

Type of error	Example from the raw file	How it is detected
Inconsistent coding	sex "f" and "Feminin" instead of "F"; mediu "Urban" instead of "urban"	frequency table of the variable (extra categories appear)
Invisible spaces	"Fumator " (with a trailing space)	frequencies; the TRIM function
Extra digit	SBP 1540 mmHg	minimum and maximum; plausible range
Wrong unit or decimal place	height 1.72 (meters) in a column in cm; weight 7.26 kg instead of 72.6	minimum and maximum; the derived variable BMI
Text in a numeric column	blood glucose "87 mg/dl"; HbA1c "5,2" written with a comma in a file that uses the decimal point, and therefore read as text	number of numeric values (n) smaller than the number of rows
Missing values	missing age; missing blood glucose at 6 months	COUNTBLANK, n reported by the software
Duplicate row	the same patient entered twice	counting unique codes; sorting by id

The recommended order of work is: first the frequencies of the qualitative variables (extra categories show up immediately), then the minimum, maximum and number of values of the numerical variables, then the check of unique codes. Half an hour of cleaning saves days of redoing the analysis.

> **Case study: the same variable, three ways of reporting**
>
> The way an article reports a variable shows how the authors understood its type. In a survey of 121 medical students, course ratings on a scale from 1 to 10 were reported as median and interquartile range: 9 (7-10) for emergency medicine, 7 (5-9) for biomedical informatics and 8 (5-9) for biostatistics [16]. The choice is correct: a mark given by a student is ordinal, and a median says "half of the students gave at least 8" without assuming that the distance between 7 and 8 equals the distance between 9 and 10. In a study on prescribing in older adults, the number of medicines per prescription, a discrete variable, was reported as mean ± standard deviation, 4.7 ± 1.51 [38]; for a count with many possible values and a relatively symmetric distribution, the mean is acceptable and easy to interpret. In a study on the acceptability of mobile technology in 103 patients with diabetes, age was reported as median and range: 37 years (18 to 65) [54]. The range shows the age spectrum of the participants, but it depends on only two extreme individuals; the interquartile range would have described better where most patients lie. None of these choices is "wrong", but each conveys something different, and the reader needs to know what exactly.

2.2 Frequency Tables

The first summary of a variable is the frequency table: the list of values (or classes of values) together with the number of times each occurs. It looks like a modest tool, but it is the safest way to see what the data look like, to discover coding errors and to prepare graphs.

2.2.1 Absolute, Relative and Cumulative Frequencies

The absolute frequency (ni) of a value is the number of observations equal to that value. The sum of the absolute frequencies is the sample size, n. The relative frequency (fi) is the absolute frequency divided by n, expressed as a fraction or as a percentage:

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

The cumulative frequency (absolute or relative) of a value is the sum of the frequencies of all values less than or equal to it. It makes sense only if the values have an order, so for ordinal and numerical variables, not for nominal ones. The cumulative relative frequency of the last class is always 100% (or 99.9-100.1%, because of rounding).

Table 2.4 shows satisfaction with the program at 6 months, an ordinal variable measured on a Likert scale from 1 (very dissatisfied) to 5 (very satisfied).

Table 2.4. Satisfaction with the program at 6 months, mg2026 dataset (n = 160).

Satisfaction	Absolute frequency	Relative frequency (%)	Cumulative absolute frequency	Cumulative relative frequency (%)
1 very dissatisfied	4	2,5	4	2,5
2 dissatisfied	19	11,9	23	14,4
3 neutral	54	33,8	77	48,1
4 satisfied	55	34,4	132	82,5
5 very satisfied	28	17,5	160	100,0
Total	160	100,0		

The table can be read in several ways. Most patients chose 4 (55 patients, 34.4%), almost tied with 3 (54). Only 14.4% of patients were dissatisfied or very dissatisfied (the cumulative frequency at value 2), and 51.9% were satisfied or very satisfied (34.4 + 17.5). The cumulative frequency passes 50% at value 4, so the median satisfaction is 4: at least half of the patients gave a rating of 4 or 5.

For nominal variables, the table is limited to absolute and relative frequencies. In mg2026, smoking status has the following distribution: never smokers 74 (46.3%), former smokers 33 (20.6%), current smokers 53 (33.1%). Area of residence: urban 95 (59.4%), rural 65 (40.6%). Sex: women 86 (53.8%), men 74 (46.3%); the rounded percentages add up to 100.1%, which is normal and may be mentioned in a note below the table.

2.2.2 Grouping Continuous Data into Classes

A continuous variable such as systolic blood pressure has dozens of distinct values in mg2026, from 102 to 169 mmHg. A table listing each value would have more than 50 rows, most of them with a frequency of 1 or 2. For this reason the values are grouped into classes (intervals) of equal width.

The number of classes is chosen so that the table is readable and preserves the shape of the distribution. A rule of thumb is Sturges' rule:

$$k = 1 + 3.322 \times \log_{10}(n)$$

For n = 160, log10(160) = 2.204, so k = 1 + 3.322 x 2.204 = 8.32, that is, about 8 classes. The class width is obtained by dividing the range by the number of classes:

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

For SBP, the range is 169 - 102 = 67 mmHg, and $h = \frac{67}{8} = 8.4$ mmHg. In practice, the width is rounded to a value convenient for the reader, here 10 mmHg, and the class limits are chosen at

round values. With classes of 10 mmHg starting from 100 we obtain 7 classes, which is perfectly acceptable: Sturges' rule is a starting point, not a law.

Class limits must be unambiguous: each value must belong to a single class. For integer values, the classes are written 100-109, 110-119 and so on. For values with decimals, intervals of the form [100, 110) are used, where the closing parenthesis shows that 110 does not belong to the class, or limits with one more decimal place than the data (99.5-109.5), as in the tables of the classic problem collections. The midpoint (center) of a class is the mean of its true limits: for the class 100-109, with true limits 99.5 and 109.5, the midpoint is 104.5.

Table 2.5. Systolic blood pressure at enrollment, grouped into 10 mmHg classes, mg2026 dataset (n = 160).

SBP (mmHg)	Class midpoint	ni	fi (%)	Cumulative frequency	Cumulative frequency (%)
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		

Clinically useful information can be read quickly from the table. The most frequent class (the modal class) is 130-139 mmHg. The proportion of patients with SBP below 140 mmHg is the cumulative frequency of the class 130-139, that is, 74.4%; consequently, 25.6% have SBP ≥ 140 mmHg, exactly the proportion calculated from the individual data (41 patients). The proportion of patients with SBP of at least 150 mmHg is 8.1 + 1.9 = 10.0%, that is, 16 patients.

Grouping comes at a price: the exact value of each observation is lost. If we have only the table, the mean can be estimated by assuming that all the values in a class are equal to its midpoint:

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

where ci is the midpoint of class i. For Table 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. The mean calculated from the individual data is 131.13 mmHg; the grouping error is small, a quarter of a mmHg.

2.2.3 Frequency Tables with Missing Values

A classic type of problem, found in all the problem collections of the discipline and in the exam, gives a frequency table with a few missing cells and asks for them to be reconstructed. No new formula is needed, only the three relationships above, used in the right order:

- absolute frequency = relative frequency x n (rounded to a whole number, because we are counting patients);
- relative frequency = absolute frequency / n x 100;
- cumulative frequency of a class = cumulative frequency of the previous class + frequency of the class; conversely, the frequency of a class = the difference between two consecutive cumulative frequencies;
- sum of absolute frequencies = n; sum of relative frequencies = 100%.

The working strategy is to start with the cells that can be calculated directly, then use them to obtain the others, and finally check that the totals add up. Worked Problems 2.1 and 2.2 show two such tables solved in full, and Practical Session 2 contains two more exercises.

2.3 Measures of Central Tendency

Frequency tables show the whole distribution. Often, however, we need a single value that summarizes the "center" of the data, the typical value. There are several such measures, and the choice between them depends on the type of variable and the shape of the distribution.

2.3.1 The Arithmetic Mean

The arithmetic mean is the sum of the values divided by their number:

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

It is the best-known and most widely used measure, has excellent mathematical properties and underlies most statistical tests. For systolic blood pressure in mg2026, m = 131.13 mmHg. The mean has a simple physical interpretation: it is the "center of gravity" of the data, the point at which the sum of the positive deviations equals the sum of the negative deviations, $\sum(x_i - m) = 0$.

The mean also has an important flaw: it is sensitive to extreme values. A single very large value pulls it up, and a single very small value pulls it down. This is why the mean is the appropriate measure for numerical variables with an approximately symmetric distribution, but it can be misleading for skewed distributions.

The weighted mean is used when values have different importance (weights), for example when we combine means calculated in groups of different sizes:

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

Mean SBP in the three arms of the mg2026 study is 131.53 mmHg (n = 53), 132.79 mmHg (n = 53) and 129.11 mmHg (n = 54). The overall mean is obtained by weighting by group size: $\frac{53\times131.53+53\times132.79+54\times129.11}{160} = \frac{6\,971.09+7\,037.87+6\,971.94}{160} = \frac{20\,980.90}{160} = 131.13$ mmHg, which is exactly the mean of the 160 patients. Here the groups are almost equal in size, and the simple average of the three means (131.14) differs very little. The difference becomes important when the groups are very unequal, as with the three practices in Worked Problem 1.5. The estimate of the mean from the frequency table in Section 2.2.2 is also a weighted mean, with weights ni.

2.3.2 The Median

The median is the value that divides the ordered series of observations into two halves of equal size: half of the observations are less than or equal to it, and half are greater than or equal to it. To calculate it, the values are sorted in ascending order. If n is odd, the median is the value in position (n + 1)/2. If n is even, the median is the mean of the two middle values, in positions n/2 and n/2 + 1.

For the ages of the first 20 patients in mg2026, sorted in ascending order (37, 40, 45, 47, 50, 52, 53, 54, 55, 56, 56, 57, 57, 60, 62, 62, 65, 66, 69, 70), n = 20 is even, and the values in positions 10 and 11 are 56 and 56; the median is 56 years. For the whole dataset, the median age is also 56 years, and the median SBP is 131 mmHg.

The median is robust: it is not influenced by the extreme values themselves, only by their position. If the oldest of the 20 patients were 99 instead of 70, the median would remain 56. This

is why the median is the preferred measure for skewed variables (triglycerides, length of hospital stay, levels of some biomarkers, waiting times) and for ordinal variables, for which the mean has no correct interpretation.

2.3.3 The Mode

The mode (modal value) is the value with the highest frequency. It is the only measure of central tendency that can be used for nominal variables: the mode of smoking status in mg2026 is "nefumator" (never smoker, 74 patients). For numerical variables, the mode is less useful, because it depends heavily on chance and on rounding. In mg2026, the most frequent SBP value is 138 mmHg (10 patients), the most frequent age is 51 years (10 patients), and the most frequent FINDRISC score is 10 (23 patients).

A distribution may have several modes. A bimodal distribution, with two peaks, often suggests that the sample mixes two different populations, for example height in a group of women and men (in mg2026, mean height is 162.6 cm in women and 175.97 cm in men). In such a case, the overall mean describes an "average patient" who does not exist, and a separate description by subgroup is more informative.

2.3.4 The Geometric Mean

The geometric mean of n positive values is the n-th root of their product. It is more conveniently calculated with logarithms: the mean of the logarithms is calculated, and then one returns to the original scale with the exponential function:

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

The geometric mean is suitable for positive, right-skewed variables whose distribution becomes approximately symmetric after a log transformation: the concentrations of some biomarkers, antibody titers, viral load, triglycerides. In mg2026, the arithmetic mean of triglycerides is 161.18 mg/dL, the median 151.5 mg/dL, and the mean of the natural logarithms is 4.97; the geometric mean is $\exp(4.97) \approx 144.0$ mg/dL. The geometric mean is smaller than the arithmetic mean (this is always the case, unless all values are equal) and close to the median, because it is not pulled up by the few very large values (the maximum is 442 mg/dL). Chapter 3 will return to the log transformation, and Chapter 4 will show how to calculate a confidence interval for the geometric mean.

A simple example shows the logic. For three antibody titers of 1/10, 1/100 and 1/1,000, the arithmetic mean of the denominators is 370, a value that resembles none of the three. The geometric mean is the cube root of 10 x 100 x 1,000 = 1,000,000, that is, 100, the middle value on the logarithmic scale on which the laboratory actually thinks.

2.3.5 Choosing the Measure and the Effect of Extreme Values

The relationship between the mean and the median says something about the shape of the distribution. In a symmetric distribution, the mean and the median are practically equal: for SBP, 131.13 and 131 mmHg; for age, 56.29 and 56 years. In a positively skewed distribution (a long tail to the right), the mean is greater than the median. For BMI, the mean is 29.04 kg/m² and the median 28.25 kg/m². Figure 2.2 shows why: a few patients with severe obesity, up to 48.3 kg/m², pull the mean to the right but do not move the median.

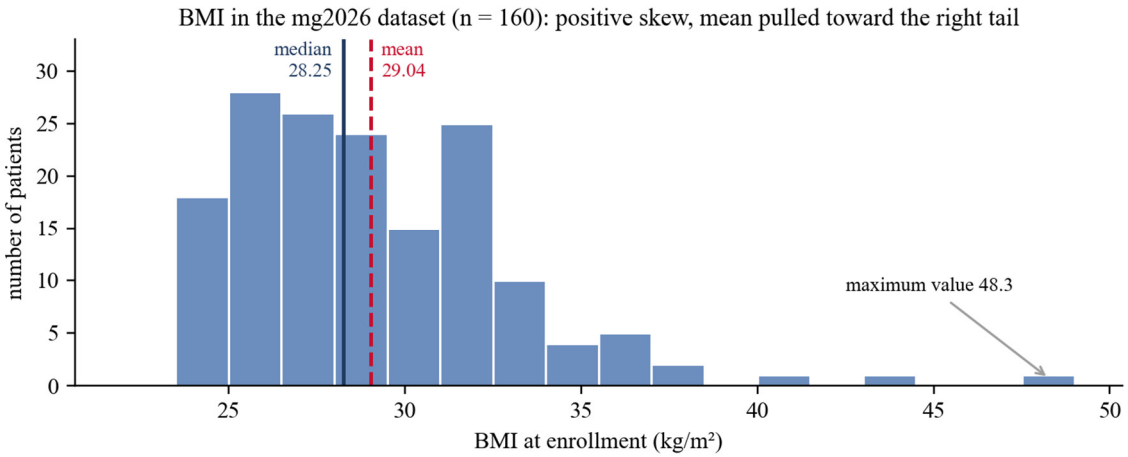


Figure 2.2. Distribution of BMI at enrollment in the mg2026 dataset (n = 160). The histogram has a tail to the right, and the mean (29.04 kg/m², dashed red line) is greater than the median (28.25 kg/m², blue line). The maximum value, 48.3 kg/m², affects the mean but not the median. Synthetic data (mg2026 dataset).

The experiment is easy to do. If we remove the patient with a BMI of 48.3 kg/m², the mean of the remaining 159 patients falls from 29.04 to 28.92 kg/m², and the standard deviation from 3.91 to 3.61 kg/m²; the median goes from 28.25 to 28.2 kg/m², a negligible difference. A single observation out of 160 changed the mean by 0.12 kg/m² and the standard deviation by almost 8%. In a sample of 20 patients, the same patient would have changed the mean by more than 1 kg/m².

This does not mean that the patient with a BMI of 48.3 kg/m² must be removed. The value is biologically plausible (class III obesity), it has been verified, and the patient belongs to the population studied. Extreme values are removed only if they are proven errors; otherwise, methods that tolerate them are chosen (the median, the interquartile range, nonparametric tests), or the analysis is reported both with and without them.

Table 2.6. Choosing the measure of central tendency.

Type of data	Recommended measure	Remarks
Nominal	mode; proportion in each category	the mean of the codes makes no sense
Ordinal	median (and mode)	the mean is sometimes reported for scores with many values, with caution
Numerical, approximately symmetric	mean (and median)	SBP, age, cholesterol, height in mg2026
Numerical, skewed	median; geometric mean for log-normal data	triglycerides, BMI, length of hospital stay, inflammatory markers
Data with plausible extreme values	median	their influence on the mean is also reported

2.4 Measures of Dispersion

Two groups of patients may have the same mean systolic blood pressure, 131 mmHg, and still be very different: in one, all patients have values between 125 and 137 mmHg, in the other the values range from 100 to 165 mmHg. Central tendency says nothing about this spread. For that we need measures of dispersion. To calculate them by hand, we will use a small example.

Working example. At a checkup, eight patients have the following fasting glucose values (mg/dL): 88, 94, 97, 102, 105, 110, 116, 128. Their sum is 840, so the mean is $m = \frac{840}{8} =$ 105 mg/dL. The values are already sorted; the median is the mean of the values in positions 4 and 5, (102 + 105)/2 = 103.5 mg/dL.

2.4.1 The Range

The range is the difference between the maximum and the minimum value: 128 - 88 = 40 mg/dL in our example. For SBP in mg2026, the range is 169 - 102 = 67 mmHg. It is simple and intuitive, but it depends on only two observations, the extreme ones, and it almost always increases with sample size: the more patients we study, the more likely we are to come across unusual values. In articles it is usually reported in the form "minimum-maximum", not as a difference.

2.4.2 The Mean Absolute Deviation

The deviation of an observation from the mean is xi - m. For our example, the deviations are -17, -11, -8, -3, 0, 5, 11, 23. Their sum is zero, as always, so the mean of the deviations cannot measure dispersion. One solution is to take absolute values. The mean absolute deviation is the mean of the distances from the mean:

$$\text{MAD} = \frac{\sum|x_i - m|}{n} = \frac{17 + 11 + 8 + 3 + 0 + 5 + 11 + 23}{8} = \frac{78}{8} = 9.75 \text{ mg/dL}$$

The patients lie, on average, almost 10 mg/dL from the group mean. This measure is easy to understand, but it has awkward mathematical properties, so in inferential statistics it is used much less often than the standard deviation. For SBP in mg2026, the mean absolute deviation is 10.91 mmHg.

2.4.3 Variance and Standard Deviation

The second solution to the problem of signs is to square the deviations. The sample variance is the sum of the squared deviations divided by n - 1:

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

The standard deviation (denoted s or SD) is the square root of the variance and has the same unit of measurement as the data:

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

Table 2.7 shows the complete calculation for the eight glucose values.

Table 2.7. Hand calculation of the variance and standard deviation for eight glucose values.

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

$$s^2 = \frac{1\,158}{8 - 1} = \frac{1\,158}{7} = 165.43 \text{ (mg/dL)}^2$$

$$s = \sqrt{165.43} = 12.86 \text{ mg/dL}$$

Why n - 1 and not **n**? The deviations are calculated from the sample mean, not from the unknown population mean, and the sample mean is, by construction, the point from which the sum of squares is smallest. Dividing by n would systematically underestimate the population variance. Dividing by n - 1 corrects this bias; the number n - 1 is called the number of degrees of freedom because, once the mean is known, only n - 1 deviations can vary freely (the last one is

determined by the condition that the sum be zero). In Google Sheets, the STDEV (or STDEV.S) function divides by n - 1, and STDEVP (STDEV.P) divides by n; for sample data the former is always used. For large n, the difference is negligible.

The standard deviation is the most widely used measure of dispersion. It has a very useful interpretation for approximately normal distributions, which we will discuss in detail in Chapter 3: about 68% of the values lie in the interval $m \pm s$ and about 95% in the interval $m \pm 2s$. For SBP in mg2026, s = 13.88 mmHg, and the interval $m \pm s$ is 117.3-145.0 mmHg. The variance, on the other hand, has a squared unit of measurement (mmHg², (mg/dL)²), which is hard to interpret clinically; for SBP, $s^2 = 192.66\ \text{mmHg}^2$. It is nonetheless indispensable in calculations, for example in the analysis of variance in Chapter 6.

2.4.4 The Coefficient of Variation

The standard deviation depends on the unit of measurement and on the order of magnitude of the data. A standard deviation of 13.88 mmHg for SBP and one of 0.45% for HbA1c cannot be compared directly. The coefficient of variation expresses the standard deviation as a percentage of the mean:

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

For the eight glucose values, $\text{CV} = \frac{12.86}{105} \times 100 = 12.2\%$. In mg2026, the CV is 10.6% for SBP, 15.6% for age, 13.5% for BMI, 8.0% for HbA1c, 17.6% for cholesterol and 47.3% for triglycerides. Relative to their mean, triglycerides are almost five times as variable as systolic blood pressure. A large CV (above 30-50%) in a variable that cannot be negative is also a sign of skewness: if the standard deviation were almost equal to the mean, the interval m - 2s would contain negative values, which are biologically impossible.

The coefficient of variation has two common uses in medicine: comparing the variability of different quantities and assessing the precision of laboratory methods (a small analytical CV indicates a reproducible method). It makes sense only for variables on a ratio scale, with an absolute zero and a positive mean. For temperature in degrees Celsius, for example, the CV would change if we switched to degrees Fahrenheit, so it has no interpretation.

2.4.5 Standard Deviation and Standard Error: A Common Confusion

The standard error of the mean (SE or SEM) is calculated from the standard deviation and the sample size:

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

For the eight glucose values, $\text{SE} = \frac{12.86}{\sqrt{8}} = \frac{12.86}{2.83} = 4.55$ mg/dL. For SBP in mg2026, $\text{SE} = \frac{13.88}{\sqrt{160}} = \frac{13.88}{12.65} = 1.097$ mmHg.

The two quantities answer different questions. The standard deviation describes how much patients vary from one another; it is a characteristic of the population and does not decrease when we study more patients. The standard error describes how precisely the sample mean estimates the population mean; it decreases as n increases. If we studied 640 patients instead of 160, the standard deviation of SBP would remain around 14 mmHg, but the standard error would be halved, to about 0.55 mmHg.

The confusion has practical consequences. Some articles report "mean ± SE" in descriptive tables without saying so. For SBP, "131.13 ± 13.88" and "131.13 ± 1.10" describe the same data, but the second form suggests, to an inattentive reader, an extremely homogeneous population. The

reporting rule is simple: the standard deviation is used to describe the patients; the confidence interval (Chapter 4), which is built with the standard error, is used to describe the precision of an estimate. Every table must state what the figure after the ± sign represents.

2.5 Quantiles, Percentiles, Quartiles and the Five-Number Summary

The median is a special case of a broader family of measures of position called quantiles. A quantile of order q (between 0 and 1) is the value below which a fraction q of the observations lies. Percentiles divide the ordered series into 100 parts: the 10th percentile (P10) is the value below which 10% of the observations lie, and P90 the value below which 90% lie. Quartiles divide the series into four equal parts: the first quartile Q1 = P25, the second Q2 = P50 = the median, the third Q3 = P75.

Percentiles are familiar to every pediatrician from growth charts: a child "on the 90th percentile for weight" weighs more than 90% of children of the same age and sex in the reference population. In mg2026, P10 and P90 are 112 and 148.2 mmHg for SBP and 86 and 126 mg/dL for fasting plasma glucose. In other words, 10% of patients have a blood glucose of at least 126 mg/dL, the diagnostic threshold for diabetes, which agrees with direct counting: 18 patients out of 160 (11.3%).

2.5.1 Calculating Quartiles

For small series, quartiles can be calculated by hand with the "medians of the halves" method: the median is found, then Q1 is the median of the lower half and Q3 the median of the upper half. For the eight glucose values (88, 94, 97, 102 | 105, 110, 116, 128), the lower half has a median of (94 + 97)/2 = 95.5, and the upper half (110 + 116)/2 = 113. Therefore $Q_1 = 95.5$ mg/dL and $Q_3 = 113$ mg/dL.

Software uses slightly different interpolation formulas. Google Sheets (the QUARTILE function) and JASP use by default a linear interpolation that, for the same eight values, gives Q1 = 96.25 and $Q_3 = 111.5$ mg/dL. In large samples the differences become negligible, and for the small series in the exam either method is accepted, provided it is applied correctly. Do not be alarmed if MedCalc and JASP give slightly different quartiles for the same data.

2.5.2 The Interquartile Range and the Five-Number Summary

The interquartile range (IQR) is the difference Q3 - Q1 and contains the central half of the observations. It is the measure of dispersion that naturally accompanies the median and, like the median, it is robust to extreme values. In articles it is usually reported as an interval, "median (Q1-Q3)", for example BMI 28.25 (26.2-31.4) kg/m², not as a single difference.

The five-number summary (minimum, Q1, median, Q3, maximum) describes a distribution compactly and is the basis of the box plot. For the main variables in mg2026:

Table 2.8. Five-number summary for some variables in mg2026 (n = 160).

Variable	Minimum	Q1	Median	Q3	Maximum	IQR
Age (years)	35	50,75	56	62	79	11,25
BMI $\left(\frac{\text{kg}}{m^2}\right)$	23,5	26,2	28,25	31,4	48,3	5,2
SBP (mmHg)	102	122	131	140	169	18
Fasting plasma glucose (mg/dL)	75	93	104,5	116	154	23
Triglycerides (mg/dL)	44	107	151,5	203,25	442	96,25

A five-number summary is read like a map. For SBP, the distances from the median to Q1 (9 mmHg) and to Q3 (9 mmHg) are equal, a sign of symmetry. For triglycerides, the distance from

the median to Q3 (51.75) is greater than that to Q1 (44.5), and the maximum (442) is very far from Q3: the distribution has a long tail to the right.

2.5.3 Extreme Values: The 1.5 x IQR Rule

A practical rule proposed by Tukey regards as "extreme values" (outliers) the observations lying more than 1.5 x IQR below Q1 or above Q3. The limits (sometimes called "fences") are:

lower limit = Q1 - 1.5 x IQR; upper limit = Q3 + 1.5 x IQR

For SBP: IQR = 18, so the limits are 122 - 27 = 95 and 140 + 27 = 167 mmHg. Two values (168 and 169 mmHg) exceed the upper limit. For BMI: IQR = 5.2, the limits are 26.2 - 7.8 = 18.4 and 31.4 + 7.8 = 39.2 kg/m²; three values exceed the upper limit (40.0, 43.4 and 48.3). For triglycerides: the upper limit is $203.25 + 1.5 \times 96.25 = 347.6$ mg/dL, exceeded by five values (364, 364, 373, 405 and 442 mg/dL).

The rule identifies values that deserve a closer look, not values that must be deleted. A blood pressure of 169 mmHg or a triglyceride level of 442 mg/dL is seen daily in family practice. The right question is always "Is the value possible, and has it been verified?", not "Does it spoil the mean?" By contrast, a blood pressure of 1540 mmHg, the one in the raw file, is not an extreme value but an error: it is biologically impossible.

2.6 The Shape of the Distribution: A First Look at Skewness and Kurtosis

Two numerical measures describe the shape of a distribution and will be studied in detail in Chapter 3. The coefficient of skewness is approximately zero for a symmetric distribution, positive for a distribution with a tail to the right (rare large values) and negative for a tail to the left. The coefficient of excess kurtosis compares the "heaviness of the tails" with that of the normal distribution: it is zero for the normal distribution, positive for distributions with heavy tails and a sharp peak, and negative for flatter distributions.

In mg2026, SBP has a skewness of 0.01 and an excess kurtosis of -0.10, practically the values of a normal distribution. BMI has a skewness of 1.42 and an excess kurtosis of 3.68, hence a pronounced right tail and a few very large values. Triglycerides have a skewness of 0.95 and an excess kurtosis of 1.12; after a log transformation, the skewness becomes -0.28, close to symmetry. Fasting plasma glucose is slightly skewed (0.37). In practice, a skewness between -0.5 and 0.5 is considered small, and one above 1 in absolute value is considered important. For biological variables that cannot be negative and have small means relative to their standard deviation, positive skewness is the rule, not the exception.

A simple clue, without any calculation: if the mean is appreciably larger than the median and the standard deviation is comparable to the mean, the distribution is almost certainly skewed to the right.

2.7 Graphical Displays

A good graph shows in a few seconds what a table shows in a few minutes. The choice of graph depends, like the choice of measures, on the type of variable. Figure 2.3 shows four of the most widely used graphs, built on the mg2026 dataset.

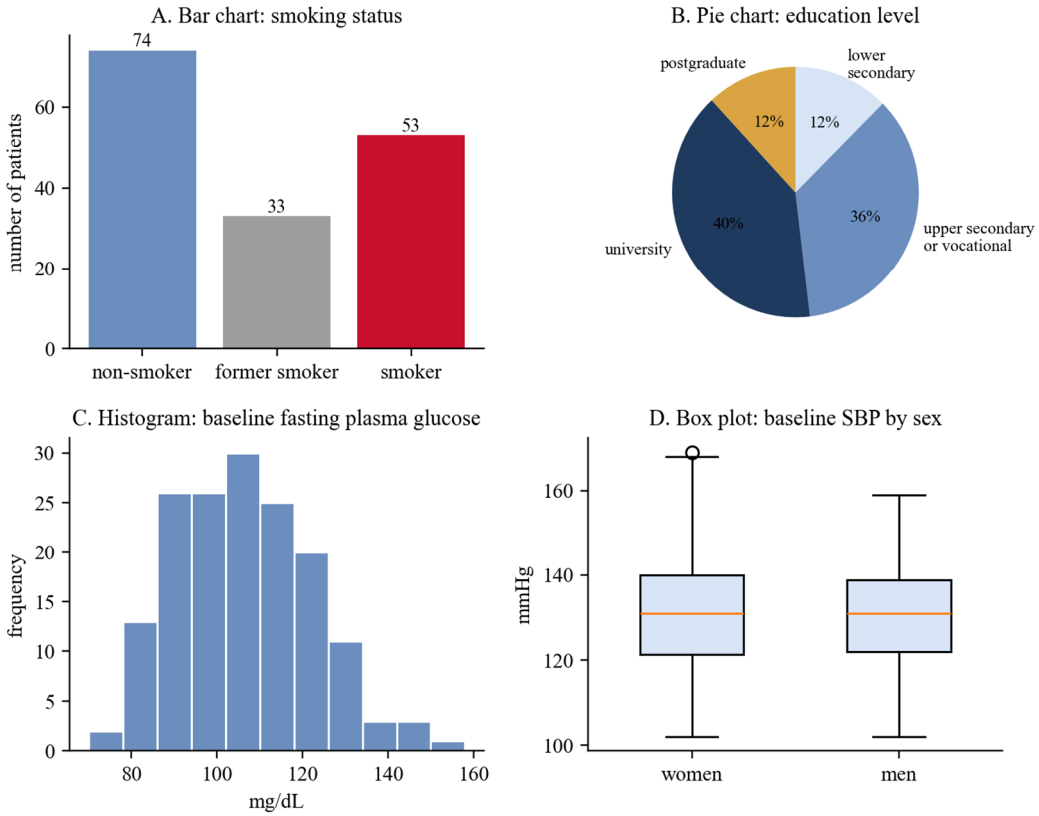


Figure 2.3. Common graphs for biomedical data, mg2026 dataset (n = 160). A: bar chart of smoking status (nominal variable). B: pie chart of educational level (ordinal variable with four categories). C: histogram of fasting plasma glucose at enrollment (continuous variable). D: box plots of systolic blood pressure in women and men. Synthetic data.

2.7.1 Graphs for Qualitative Variables

A bar (column) chart represents each category by a bar whose height is proportional to its frequency (absolute or relative). The bars are separated from one another, because the categories are distinct. For nominal variables, the order of the bars can be chosen freely (for example, in descending order of frequency); for ordinal variables, the natural order of the categories is kept. Figure 2.3A shows that never smokers are the largest group (74), followed by current smokers (53) and former smokers (33). Grouped or stacked bar charts make it possible to compare a variable between groups, for example smoking status by sex.

A pie chart divides a circle into sectors proportional to the relative frequencies; the angle of each sector is fi x 360°. It shows the structure of a whole with few categories (two to five) well. For educational level in mg2026 (Figure 2.3B), the sectors correspond to proportions of 12.5% (lower secondary school), 35.6% (high school or vocational school), 40.0% (university) and 11.9% (postgraduate); the university sector has an angle of 0.40 x 360 = 144°. The human eye, however, has difficulty comparing similar angles, so for more than five categories or for comparisons between groups, a bar chart is almost always better.

2.7.2 Graphs for Quantitative Variables

The histogram is the graphical representation of the frequency table for data grouped into classes. The classes are on the horizontal axis and the frequencies on the vertical axis, and the bars touch one another, because the variable is continuous. The histogram shows the shape of the distribution: symmetric or skewed, with one peak or two, with isolated values. Figure 2.3C shows a glucose distribution with a peak around 95-115 mg/dL and a slight tail to the right. The appearance of a histogram depends on the class width: classes that are too wide hide the shape of

the distribution, while classes that are too narrow produce a jagged graph. It is worth trying two or three versions.

The box plot (box-and-whisker plot) draws the five-number summary. The box extends from Q1 to Q3, a line inside it marks the median, and the whiskers extend to the smallest and largest values lying within the 1.5 x IQR limits; values outside these limits appear as separate points. The box plot is ideal for comparing several groups. In Figure 2.3D, the median SBP is 131 mmHg in both women and men, and the boxes overlap almost completely (women: Q1 121.25, Q3 140; men: Q1 122, Q3 138.75). The graph anticipates the result of the test in Chapter 6: there is no difference in SBP between the sexes in this sample.

The stem-and-leaf plot is a "written" histogram that preserves the individual values. Each value is split into a stem (the leading digits) and a leaf (the last digit). For the ages of the first 20 patients in mg2026:

Table 2.9. Stem-and-leaf plot of the ages of the first 20 patients (stem = tens, leaf = units).

Stem	Leaves
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

The plot shows immediately that most patients are between 50 and 59 years old and that the distribution is approximately symmetric. The minimum (37), the maximum (70) and the median (the mean of the values in positions 10 and 11, both 56) can be read directly from it. It is useful for small datasets written by hand.

The dot plot represents each observation by a point on an axis; equal values are stacked on top of one another. For small samples (under 30-40 observations) it is more honest than a bar chart showing only the mean, because it lets the reader see every patient. For small groups, many journals now recommend graphs that show the individual points over a box plot.

The scatter plot represents the relationship between two numerical variables: each patient is a point with coordinates (x, y). We will use it extensively in Chapter 9 for correlation and regression; for example, between BMI and SBP in mg2026 the cloud of points rises moderately, with a correlation coefficient r = 0.405. The scatter plot also helps in judging published results. In a study of 867 patients with colon cancer, the correlation between age and duration of surgery was r = -0.088, statistically significant (p = 0.01) [55]. A scatter plot with r = -0.088 looks practically like a blob with no direction at all; the statistical significance comes from the large sample size, not from a clinically important relationship.

The line plot connects values ordered in time with lines: mean blood glucose at 0, 3 and 6 months, the monthly incidence of a disease, the course of a patient's temperature on the chart. Lines make sense only when the horizontal axis has a natural order, usually time.

Table 2.10. Choosing a graph according to the type of variable.

Variable(s)	Recommended graph	To avoid
One nominal variable	bar chart; pie chart for 2-5 categories	pie chart with many categories or in 3D
One ordinal variable	bars in the order of the categories; stacked bars for Likert scales	pie chart that scrambles the order
One numerical variable	histogram; box plot; stem-and-leaf or dot plot for small n	bars showing only the mean
Numerical, by group	side-by-side box plots; individual points for small n	bars with "mean ± SE" without saying so

Variable(s)	Recommended graph	To avoid
Two numerical variables	scatter plot	lines between points with no time order
Change over time	line plot	3D bars

2.7.3 Rules for a Correct Graph

A graph can be numerically correct and still misleading. Figure 2.4 shows the most common pitfall: the truncated axis. In mg2026, mean blood glucose fell from 105.4 mg/dL at enrollment to 99.4 mg/dL at 3 months and 95.2 mg/dL at 6 months, that is, by about 10%. If the vertical axis starts at 94 mg/dL (panel A), the bar at 6 months looks almost ten times shorter than the one at enrollment, as if blood glucose had collapsed. With the axis starting at zero (panel B), the proportions between the bars are the real ones.

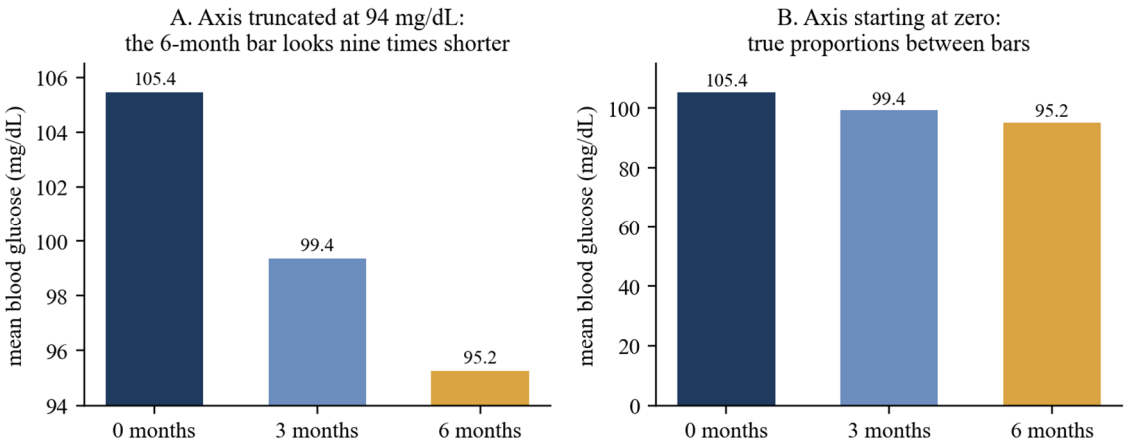


Figure 2.4. The same result, two different impressions. Mean blood glucose at 0, 3 and 6 months in the mg2026 dataset (105.4, 99.4 and 95.2 mg/dL). A: the vertical axis truncated at 94 mg/dL visually exaggerates the decrease. B: the axis starting at zero preserves the real proportions between the bars. Synthetic data (mg2026 dataset).

The rule is simple: in bar charts, the value axis starts at zero, because it is the length of the bar that conveys the magnitude. In line plots and scatter plots, the axis may start at a value other than zero, because what matters there is the position of the points, not the length of bars; in this case, the axis break must be clearly marked.

Other rules, gathered from the experience of grading students' work:

- No three-dimensional effects. A tilted 3D pie chart makes the front sectors look larger than they are.
- No two vertical axes with different scales on the same graph unless absolutely necessary; the choice of scales can make two curves appear to track each other or to diverge, as the author wishes.
- Axes have a title and a unit of measurement (SBP, mmHg; blood glucose, mg/dL).
- The legend or title of the figure states n, the period and the data source; the figure must be understandable without the text.
- Error bars are explained: standard deviation, standard error or 95% confidence interval. Without this information, the bars cannot be interpreted.
- Colors are chosen so that the graph remains readable when printed in black and white and by people with color vision deficiency.
- One graph, one message. If a graph needs a paragraph of explanation to be understood, it probably needs to be split in two.

2.8 The Descriptive Table of an Article: "Table 1"

Almost every clinical research article opens its results section with a table describing the participants, informally called "Table 1". It presents the demographic and clinical characteristics of the sample, often separately for the groups being compared. The conventions are the same in all journals: symmetric numerical variables are reported as mean ± standard deviation; skewed numerical or ordinal variables as median (Q1-Q3); qualitative variables as number of patients and percentage, n (%). Reporting guidelines require the table to state, for each variable, how many values are missing [25]. Table 2.11 is the "Table 1" of the simulated mg2026 study.

Table 2.11. Characteristics of the patients at enrollment, by treatment arm, mg2026 dataset.

Characteristic	Arm A (n = 53)	Arm B (n = 53)	Arm C (n = 54)	Total (n = 160)
Age, years, mean ± SD	57,53 ± 9,19	56,87 ± 9,42	54,52 ± 7,47	56,29 ± 8,77
Female sex, n (%)	32 (60,4)	26 (49,1)	28 (51,9)	86 (53,8)
Urban residence, n (%)	25 (47,2)	36 (67,9)	34 (63,0)	95 (59,4)
Current smokers, n (%)	18 (34,0)	18 (34,0)	17 (31,5)	53 (33,1)
BMI, kg/m², median (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)
SBP, mmHg, mean ± SD	131,53 ± 13,84	132,79 ± 13,79	129,11 ± 14,01	131,13 ± 13,88
Fasting plasma glucose, mg/dL, mean ± SD	107,79 ± 15,65	106,28 ± 14,70	102,30 ± 16,64	105,44 ± 15,76
Triglycerides, mg/dL, median (Q1-Q3)	143 (95-204)	144 (112-185)	166,5 (107-213)	151,5 (107-203,25)
FINDRISC score, median (Q1-Q3)	10	10	10	10 (8-13)
Diabetes confirmed by OGTT, n (%)	10 (18,9)	9 (17,0)	10 (18,5)	29 (18,1)

Note: SD, standard deviation; Q1, Q3, first and third quartiles. There are no missing values. Percentages are calculated from the number of patients in each arm.

A few remarks on reading such a table. The first thing to check is whether the numbers add up: 32 + 26 + 28 = 86 women; 53 + 53 + 54 = 160. The second is whether the format of each variable matches its distribution: BMI and triglycerides, which are skewed, are reported as median (Q1-Q3), and SBP and age, which are symmetric, as mean ± SD. The third, specific to randomized trials, is the comparison of the arms. Arm A has more women (60.4%) and fewer urban patients (47.2%) than the others. These differences are the effect of chance, because allocation was random; this is why the guidelines for randomized trials recommend not calculating p values for the comparisons in Table 1, but judging instead whether the imbalances are large enough to matter clinically [56].

Articles by colleagues from our discipline offer varied examples of descriptive reporting. A diagnostic accuracy study of contrast-enhanced ultrasound for cirrhosis reported age as mean ± SD, 58.98 ± 8.90 years, and sex as a percentage, 56.3% men, for 64 participants (37 with cirrhosis, 27 healthy volunteers); the ultrasound parameter was 0.83 ± 0.09 in patients with cirrhosis versus 0.49 ± 0.11 in volunteers [2]. A study on the diagnosis of malignant biliary obstruction, in 160 patients, reported 85 men (53.1%) and a mean age of 69.31 ± 10.96 years [57]. In the hemodialysis study from Chapter 1, the ages of the two cohorts were reported as 59.7 ±

12.92 and 61 ± 12.93 years [39]. All these reports follow the "mean ± SD" convention for an approximately symmetric age distribution.

It is always worth checking the internal consistency of the figures as well. In a study on the ultrasound quantification of hepatic steatosis, steatosis grades were reported as S0 48/176 (27.2%), S1 56/176 (31.6%), S2 14/176 (7.3%) and S3 59/176 (33.9%), with mean values of the UGAP parameter rising from $198.3 \pm \frac{25.7 d_B}{m}$ in S0 to $270.8 \pm \frac{31.62 d_B}{m}$ in S3 [58]. If you add up the numerators, you get 177, not 176, and 14/176 means 8.0%, not 7.3%. Most likely this is an editing error in the abstract, with no consequences for the conclusions, but the exercise shows why a careful reader always redoes the sums. Note also the unequal number of decimal places (198.3 versus 216.86): a good rule is that all values of the same variable should be reported with the same number of decimal places, and that this number should not exceed the precision of the measurement.

For qualitative variables from questionnaires, reporting consists almost exclusively of percentages. In the osteoporosis survey conducted in pharmacies, of 189 respondents, 78.8% were women, the largest age groups were 60-69 years (31.7%) and 50-59 years (30.7%), and 69.3% came from urban areas [27]. A complete frequency table would add the absolute number for each percentage, so that the reader could redo the calculations.

Case study: means, medians and scores with a floor effect

A retrospective study of 200 patients with type 2 diabetes treated with SGLT-2 inhibitors reported values before and after one year of treatment [59]. Blood glucose fell from 180.00 mg% (IQR 154.50-207.00) to 130.00 mg% (IQR 117.50-150.00), HbA1c from 8.40% (IQR 7.98-9.15%) to 7.30% (IQR 6.90-7.95%), weight from 90.50 kg (82.00-106.50) to 89.00 kg (77.50-100.00), and systolic blood pressure from 140.00 mmHg (130.00-160.00) to 120.00 mmHg (115.50-130.00), all with $p < 0.001$ except weight ($p = 0.018$). Waist circumference was reported differently, as 107.05 (± 16.39) and $102.50 (\pm 15.11)$ $(p = 0.042)$. The presence of the IQR shows that the first values are medians, although the abstract calls them "means"; this is a common terminological inconsistency, which the reader must recognize from the form of the reporting. Reporting some variables as median (IQR) and others as mean ± SD is correct if the choice followed the distribution of each variable, but the methods should say so explicitly, together with which paired test was used. A second example: in a cross-sectional study of 304 adults with diabetes, a digital engagement score was reported as median [Q1-Q3], 1.71 [1.71-2.39] versus 3.86 [3.86-4.43] between groups, and 35.5% of participants had high digital engagement [7]. Note that in both groups the median coincides with the first quartile. This happens when a score has few possible values and many observations equal to the lowest value in a group: at least a quarter of the patients, and possibly up to half, have exactly the same value. The phenomenon is related to the floor effect, described explicitly in a validation study of a questionnaire on menopause burden in 308 women, where 37.0% of respondents had the minimum score on one subscale and 23.4% on another [60]. When a large proportion of patients have the minimum score, the mean and standard deviation describe the distribution poorly, and the instrument cannot distinguish between patients at the bottom of the scale. What should additionally be reported: the proportion of observations at the minimum and maximum values and, for bounded scores, a histogram.

2.9 Complete Descriptive Analysis of the mg2026 Dataset

We close the theoretical part with a complete description of the dataset you will use throughout the semester. The values in Tables 2.12 and 2.13 are the ones you should obtain in JASP and

Google Sheets in Practical Sessions 1 and 2; if your result differs, look for the cause in the import or the cleaning.

Table 2.12. Descriptive statistics for the numerical variables at enrollment, mg2026 dataset (n = 160).

Variable	Mean	SD	Median	Q1-Q3	Min-max	CV (%)	Skewness
Age (years)	56,29	8,77	56	50,75-62	35-79	15,6	0,01
Height (cm)	168,79	9,05	168	162-177	145-192	5,4	0,10
Weight (kg)	83,05	15,09	80,2	72,05-90,93	58,1-159,9	18,2	1,29
BMI $\left(\frac{\text{kg}}{m^2}\right)$	29,04	3,91	28,25	26,2-31,4	23,5-48,3	13,5	1,42
Waist (cm)	88,54	12,11	86	80,75-95	65-140	13,7	0,86
SBP (mmHg)	131,13	13,88	131	122-140	102-169	10,6	0,01
DBP (mmHg)	79,23	8,63	80	74-85	58-103	10,9	-0,09
Heart rate (beats/min)	74,71	9,30	74,5	69-80	51-102	12,4	0,05
Total cholesterol (mg/dL)	209,39	36,91	207	183-235	119-306	17,6	0,02
HDL cholesterol (mg/dL)	49,52	11,78	49	41-59	20-78	23,8	0,04
Triglycerides (mg/dL)	161,18	76,30	151,5	107-203,25	44-442	47,3	0,95
Fasting plasma glucose (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
FINDRISC score	10,37	4,27	10	8-13	0-25	41,2	0,39

Table 2.13. Distribution of the qualitative variables, mg2026 dataset (n = 160).

Variable	Category	n	%
Sex	women / men	86 / 74	53,8 / 46,3
Residence	$\frac{\text{urban}}{\text{rural}}$	95 / 65	59,4 / 40,6
Smoking	never smoker / former smoker / current smoker	74 / 33 / 53	46,3 / 20,6 / 33,1
Education	lower secondary / high school or vocational / university / postgraduate	20 / 57 / 64 / 19	12,5 / 35,6 / 40,0 / 11,9
Physical activity (1-5)	1 / 2 / 3 / 4 / 5	21 / 42 / 60 / 34 / 3	13,1 / 26,3 / 37,5 / 21,3 / 1,9
Satisfaction at 6 months (1-5)	1 / 2 / 3 / 4 / 5	4 / 19 / 54 / 55 / 28	2,5 / 11,9 / 33,8 / 34,4 / 17,5
Treatment arm	$\frac{A}{\frac{B}{C}}$	53 / 53 / 54	33,1 / 33,1 / 33,8

Variable	Category	n	%
Diabetes confirmed by OGTT	yes / no	29 / 131	18,1 / 81,9

Note: percentages may add up to 100.1 because of rounding.

How should the interpretation of these tables be written? Here is a model, in the style of the results section of an article. "A total of 160 patients were enrolled (86 women, 53.8%), with a mean age of 56.29 ± 8.77 years (range 35-79). Most came from urban areas (59.4%), and one third were current smokers (33.1%). Median BMI was 28.25 kg/m² (Q1-Q3 26.2-31.4), with 54 patients (33.8%) obese. Mean systolic blood pressure was 131.13 ± 13.88 mmHg, and 41 patients (25.6%) had SBP ≥ 140 mmHg at enrollment. Mean fasting plasma glucose was 105.44 ± 15.76 mg/dL, mean HbA1c 5.58 ± 0.45%, and diabetes was confirmed by OGTT in 29 patients (18.1%). Triglycerides had a skewed distribution, with a median of 151.5 mg/dL (Q1-Q3 107-203.25)."

A few remarks on the choices made in this text. Age, SBP, blood glucose, HbA1c, cholesterol and HDL have little skewness and means practically equal to their medians, so they are reported as mean ± SD. BMI, weight, waist and triglycerides have marked positive skewness (above 0.8) and are more correctly reported as median (Q1-Q3). The FINDRISC score is a discrete variable with many possible values; it can be reported either way, but the median with IQR is the safer choice. Education, physical activity and satisfaction are ordinal: they are reported as frequencies and, possibly, as a median (3, 3 and 4, respectively), never as mean ± SD. For dichotomous variables, the percentage fully describes the variable; the standard deviation is not reported.

Case study: when the standard deviation is almost as large as the mean

The standard deviation reported in an article can be used as a quick test of the shape of the distribution. In a retrospective study of 486 patients with COVID-19 (200 without diabetes, 62 with type 1 diabetes, 224 with type 2 diabetes), length of hospital stay was 11.6 ± 7.0 days in type 2 diabetes and 9.1 ± 5.8 days in type 1 diabetes, and blood glucose at admission was 178.3 ± 34.7 mg/dL versus 164.8 ± 39.6 mg/dL [8]. For blood glucose, the coefficient of variation is about 20%, compatible with a reasonably symmetric distribution. For length of stay, the CV is about 60% (7.0/11.6), and m - 2s = 11.6 - 14.0 = -2.4 days, an impossible value. The distribution is almost certainly skewed to the right, with a few patients hospitalized for a very long time, and the median with IQR would have described it better. The situation is even clearer in a study on the diagnosis of stroke in the emergency department, in 85 patients: the GFAP protein was 142.91 ± 102.19 pg/mL in patients with stroke and 37.76 ± 19.92 pg/mL in those without, and UCH-L1 was 1,307.68 ± 967.54 versus 189.81 ± 92.69 pg/mL[61]. A standard deviation of more than 70% of the mean signals marked skewness; for such a variable, the comparison should be made with a nonparametric test (Chapter 7) or after a log transformation (Chapter 3). By contrast, another study from the same emergency department, on thrombolysis in ischemic stroke, correctly reported times, typically skewed variables, as medians: a median of 100 minutes from onset to arrival in the emergency department and 15.5 minutes from arrival to CT scan, in 258 eligible patients out of the 355 initially assessed, of whom 46 received thrombolysis (17.83%) [62]. The conclusion for the reader: before accepting a comparison of means, compare the standard deviation with the mean; if s is greater than about half of m for a variable that cannot be negative, the distribution is skewed and the mean is not the typical value.

2.10 Worked Problems

Worked Problem 2.1. Frequency Table with Missing Values: Systolic Blood Pressure in a Cardiology Outpatient Clinic

Statement. In a cardiology outpatient clinic, systolic blood pressure was measured in 50 hypertensive patients at their first visit. The results were grouped in the table below, from which a few values are missing. (a) Fill in the values A, B, C and D. (b) What is the probability that a patient chosen at random from this group has an SBP of at least 160 mmHg? (c) In which class does the median lie? Estimate the median and the mean from the table.

Table 2.14. SBP in 50 hypertensive patients (grouped data).

SBP (mmHg)	Absolute frequency	Relative frequency (%)	Cumulative frequency (%)
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 is obtained from the relative frequency: A = 34.0% x 50 = 0.34 x 50 = 17 patients. B is obtained from the absolute frequency: $B = \frac{12}{50} \times 100 = 24.0\%$. C is the cumulative frequency of the third class: C = 30.0 + 34.0 = 64.0%. Check: 64.0 + 24.0 = 88.0, exactly the value in the next row. D is obtained from the relative frequency: D = 0.08 x 50 = 4 patients. Final check: 4 + 11 + 17 + 12 + 4 + 2 = 50, and 8 + 22 + 34 + 24 + 8 + 4 = 100%. The table adds up.

(b) SBP ≥ 160 mmHg means the last three classes: 24.0 + 8.0 + 4.0 = 36.0%. Equivalently, 100 - 64.0 = 36.0% (the complement of the cumulative frequency up to 159 mmHg). The probability is 0.36, that is, 18 patients out of 50.

(c) The median is the value that divides the 50 patients into two halves, so it lies between the 25th and the 26th ordered values. The cumulative frequency is 30.0% (15 patients) at the end of the class 120-139 and 64.0% (32 patients) at the end of the class 140-159; the 25th value is therefore in the class 140-159, which is the median class. If we assume that the 17 patients in this class are uniformly distributed, the median is estimated by interpolation, using the true limits of the class (139.5-159.5, width h = 20):

$$\text{Median} \approx 139.5 + \frac{25-15}{17} \times 20 = 139.5 + 11.8 = 151.3 \text{ mmHg}$$

The mean is estimated using the class midpoints (109.5, 129.5, 149.5, 169.5, 189.5 and 209.5):

$$m \approx \frac{4 \times 109.5 + 11 \times 129.5 + 17 \times 149.5 + 12 \times 169.5 + 4 \times 189.5 + 2 \times 209.5}{50}$$

$$= \frac{438 + 1\,424.5 + 2\,541.5 + 2\,034 + 758 + 419}{50} = \frac{7\,615}{50} = 152.3 \text{ mmHg}$$

Interpretation. More than one third of the hypertensive patients (36%) present with values of at least 160 mmHg, that is, in the range of grade 2 hypertension or higher. The estimated mean and median are close (152.3 versus 151.3 mmHg), so the distribution is approximately symmetric, with a slight tail toward high values. Caution: both values are estimates from grouped data; with the individual data we would obtain slightly different values.

Worked Problem 2.2. Frequency Table for a Discrete Variable: Consultations per Year

Statement. A family doctor counted the consultations in the past year for 80 patients with type 2 diabetes chosen at random from the practice list. (a) Fill in the values A, B, C and D. (b) Calculate the mode, median and mean of the number of consultations. (c) What proportion of patients had at least 4 consultations?

Table 2.15. Number of consultations per year in 80 patients with type 2 diabetes.

Consultations per year	Absolute frequency	Relative frequency (%)	Cumulative frequency (%)
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 (or from the cumulative frequencies: 50.0 - 25.0 = 25.0%, which again gives 20). $B = \frac{18}{80} \times 100 = 22.5\%$ (check: 50.0 + 22.5 = 72.5). C = 0.125 x 80 = 10 patients. D = 72.5 + 12.5 = 85.0% (check: 85.0 + 10.0 = 95.0). Sum of the absolute frequencies: 6 + 14 + 20 + 18 + 10 + 8 + 4 = 80.

(b) The mode is the value with the highest frequency: 2 consultations (20 patients). For the median, n = 80 is even, so the median is the mean of the values in positions 40 and 41. The cumulative frequency is exactly 50.0% (40 patients) at the value 2: the 40th value is 2, and the 41st is the first value of 3. The median is (2 + 3)/2 = 2.5 consultations. The mean is a weighted mean with the frequencies as weights:

$$m = \frac{0\times6+1\times14+2\times20+3\times18+4\times10+5\times8+6\times4}{80} = \frac{0+14+40+54+40+40+24}{80} = \frac{212}{80} = 2.65 \text{ consultations}$$

(c) At least 4 consultations: 12.5 + 10.0 + 5.0 = 27.5%, that is, 22 patients. Equivalently, 100 - 72.5 = 27.5%.

Interpretation. The "typical" patient has two or three consultations a year, and a quarter of the patients (25.0%, the cumulative frequency at the value 1) have at most one consultation. The trap in this problem is the median: when the cumulative frequency reaches exactly 50% at a value, the median is not that value but the mean of that value and the next one. In the exam, many students answer "2". From a clinical point of view, the 6 patients (7.5%) with no consultation at all in a year are the ones who should be actively called in for a checkup.

Worked Problem 2.3. Classifying the Variables on a Case Record Form

Statement. The record form of a study of patients hospitalized with pneumonia contains the following fields: (1) patient code; (2) age (years); (3) sex; (4) CURB-65 score (0-5); (5) oxygen saturation (%); (6) number of days in hospital; (7) perceived severity of dyspnea (absent, mild, moderate, severe); (8) organism isolated (pneumococcus, Haemophilus, Mycoplasma, other, not isolated); (9) in-hospital death (yes/no); (10) temperature at admission (°C). For each variable, state its type, an appropriate descriptive measure and an appropriate graph. Solution. Table 2.16 summarizes the answer.

Table 2.16. Classification of the variables on the pneumonia study form.

Variable	Type	Descriptive measure	Graph
Patient code	identifier, not analyzed	-	-
Age	continuous numerical, ratio scale	mean ± SD (or median, if skewed)	histogram, box plot

Variable	Type	Descriptive measure	Graph
Sex	nominal dichotomous	n (%)	bar chart
CURB-65 score	ordinal (score with 6 values)	median (Q1-Q3), frequencies	bars in score order
O2 saturation	continuous numerical, bounded above at 100%	median (Q1-Q3), because it is usually skewed to the left	histogram, box plot
Days in hospital	discrete numerical, skewed	median (Q1-Q3)	histogram, box plot
Severity of dyspnea	ordinal	frequencies, median	bars in the order of the categories
Organism isolated	nominal with 5 categories	n (%)	bar chart (descending order)
In-hospital death	nominal dichotomous	n (%)	bar chart
Temperature	continuous numerical, interval scale	mean ± SD	histogram, box plot

Interpretation. Two details deserve attention. The CURB-65 score is numerical in form but ordinal in nature: the difference between 0 and 1 point does not have the same clinical meaning as that between 3 and 4. Oxygen saturation cannot exceed 100%, and in patients with pneumonia the values crowd toward the top of the scale, with a tail toward low values; the median is more appropriate than the mean. Temperature is the classic example of an interval scale: it makes no sense to say that 40 °C is twice 20 °C, so the coefficient of variation makes no sense either.

Worked Problem 2.4. Complete Description of a Small Series: HbA1c in Ten Patients

Statement. The HbA1c values (%) of the 10 patients with type 2 diabetes seen in one day are: 6.2, 6.8, 7.1, 7.4, 7.5, 7.9, 8.3, 8.8, 9.6, 11.4. Calculate the mean, median, standard deviation, coefficient of variation, standard error, quartiles and interquartile range, and check whether there are extreme values according to the 1.5 x IQR rule. What proportion of patients have HbA1c ≥ 7%?

Solution. The values are already sorted, n = 10.

Mean: the sum of the values is 81.0, so $m = \frac{81.0}{10} = 8.1\%$.

Median: n is even, and the values in positions 5 and 6 are 7.5 and 7.9; median = (7.5 + 7.9)/2 = 7.7%.

Standard deviation: the deviations from the mean are -1.9, -1.3, -1.0, -0.7, -0.6, -0.2, 0.2, 0.7, 1.5 and 3.3. Their squares are 3.61, 1.69, 1.00, 0.49, 0.36, 0.04, 0.04, 0.49, 2.25 and 10.89, with a sum of 20.86.

$s^2 = \frac{20.86}{9} = 2.318$; $s = \sqrt{2.318} = 1.52\%$

Coefficient of variation: $\mathrm{CV} = \frac{1.52}{8.1} \times 100 = 18.8\%$. Standard error: $\mathrm{SE} = \frac{1.52}{\sqrt{10}} = \frac{1.52}{3.162} = 0.48\%$.

Quartiles (medians of the halves method): the lower half (6.2, 6.8, 7.1, 7.4, 7.5) has the median Q1 = 7.1%; the upper half (7.9, 8.3, 8.8, 9.6, 11.4) has the median Q3 = 8.8%. IQR = 8.8 - 7.1 = 1.7%. (Google Sheets and JASP give, by interpolation, Q1 = 7.175 and Q3 = 8.675.)

Extreme values: the limits are 7.1 - 1.5 x 1.7 = 7.1 - 2.55 = 4.55% and 8.8 + 2.55 = 11.35%. The value 11.4% slightly exceeds the upper limit and appears as an extreme value.

Proportion with HbA1c ≥ 7%: 8 patients out of 10, that is, 80%.

Interpretation. The mean (8.1%) is greater than the median (7.7%), a sign that one large value, 11.4%, pulls the mean up. Without this value, the mean of the nine patients $\frac{\text{ar fi}\quad 69.6}{9} =$ 7.73%, practically equal to the median. The patient with 11.4% is not an error but a patient with

very poorly controlled diabetes, who urgently needs treatment intensification. For such a small series, the correct way to report is the median 7.7% (Q1-Q3 7.1-8.8), possibly supplemented by mean ± SD, 8.1 ± 1.52%. From a clinical point of view, 80% of the patients do not reach the HbA1c target below 7%, and this is, in fact, the main message of the series.

Worked Problem 2.5. Standard Deviation or Standard Error?

Statement.

(a) An article reports "SBP 158 ± 2.1 mmHg" for 100 hypertensive patients, without stating what 2.1 represents. Which is more plausible: a standard deviation or a standard error? (b) Another article reports length of hospital stay for 81 patients as 6.5 ± 5.9 days (mean ± SD). What can you say about the shape of the distribution? Calculate the standard error.

Solution.

(a) If 2.1 were the standard deviation, about 95% of the patients would have an SBP between 158 - 4.2 = 153.8 and 158 + 4.2 = 162.2 mmHg, that is, all the hypertensive patients would have almost identical blood pressures. This is biologically implausible: in mg2026, the standard deviation of SBP is 13.88 mmHg. If 2.1 is the standard error, the standard deviation ar fi $s = \text{SE} \times \sqrt{n} = 2.1 \times \sqrt{100} = 2.1 \times 10 = 21$ mmHg, a perfectly plausible value for a group of hypertensive patients. Conclusion: it is almost certainly the standard error, and the author should have said so.

(b) The coefficient of variation is $\frac{5.9}{6.5} \times 100 = 91\%$, and m - 2s = 6.5 - 11.8 = -5.3 days, which is impossible. The distribution is strongly skewed to the right: most patients stay a few days, and a few stay a very long time. The median and IQR would have been more informative. The standard error is $\text{SE} = \frac{5.9}{\sqrt{81}} = \frac{5.9}{9} = 0.66$ days.

Interpretation. The first question to ask when you see the ± sign is "what does it represent?" The second is "is it plausible?" A very small s/m ratio for a biological variable suggests a standard error; an s/m ratio close to 1 for a positive variable signals skewness. The standard error of 0.66 days shows that the mean of 6.5 days is estimated fairly precisely, even though individual patients vary a great deal; these are two different things.

Worked Problem 2.6. The Weighted Mean and the Geometric Mean

Statement. (a) Three practices report the mean HbA1c of their patients with diabetes: 7.2% (40 patients), 7.8% (25 patients) and 6.9% (85 patients). What is the mean HbA1c for all 150 patients? (b) In one patient, C-reactive protein (CRP) had the values 2, 8 and 32 mg/L at three successive measurements. Calculate the arithmetic mean and the geometric mean. Which one better describes the typical value?

Solution. (a) The mean weighted by the number of patients:

$$m = \frac{40 \times 7.2 + 25 \times 7.8 + 85 \times 6.9}{40 + 25 + 85} = \frac{288 + 195 + 586.5}{150} = \frac{1\,069.5}{150} = 7.13\%$$

The simple average of the three means ar fi $\frac{(7.2+7.8+6.9)}{3} = 7.3\%$, an incorrect value, because it gives the practice with 25 patients the same importance as the one with 85.

(b) Arithmetic mean: $\frac{2+8+32}{3} = \frac{42}{3} = \frac{14\text{mg}}{L}$. Geometric mean: the cube root of 2 x 8 x 32 = 512, that is, 8 mg/L. Equivalently, with common logarithms: log 2 = 0.301; log 8 = 0.903; log 32 = 1.505; their mean = 2.709/3 = 0.903; 10 to the power 0.903 = 8.

Interpretation. The CRP values increase fourfold from one measurement to the next; on the logarithmic scale they are equally spaced, and the middle value, 8 mg/L, is the typical one. The

arithmetic mean, 14 mg/L, is pulled up by the last value. For markers that vary multiplicatively (CRP, titers, viral load, triglycerides), the geometric mean is the natural choice.

2.11 Proposed Exercises

1. Classify the following variables (nominal, ordinal, discrete numerical, continuous numerical) and state the measurement scale: blood group; Apgar score at 5 minutes; number of previous pregnancies; body temperature; NYHA class; HbA1c; agreement with the statement "I am satisfied with the treatment" (5-point Likert scale); the patient's numeric code; length of hospital stay in days; smoking status (yes/no).
2. Complete the frequency table of fasting plasma glucose in 40 patients: class 70-89 mg/dL, 6 patients, 15.0%, cumulative 15.0%; class 90-109, A patients, 35.0%, cumulative B; class 110-125, 10 patients, C%, cumulative 75.0%; class 126-139, D patients, 12.5%, cumulative 87.5%; class ≥ 140, E patients, cumulative 100.0%. What proportion of patients have blood glucose in the diagnostic range for diabetes (≥ 126 mg/dL)?
3. For the SBP values 118, 124, 130, 136, 142 and 150 mmHg, calculate by hand the mean, median, mean absolute deviation, variance, standard deviation, coefficient of variation and standard error.
4. The length of hospital stay (days) in 10 patients operated on for cholecystitis was: 3, 4, 4, 5, 5, 6, 7, 8, 12, 31. Calculate the mean, median and mode. Determine the quartiles with the medians of the halves method and check for extreme values with the 1.5 x IQR rule. How does the mean change without the last value? Which measures would you report?
5. A study has 250 triglyceride values, between 68 and 412 mg/dL. How many classes does Sturges' rule recommend? What class width results, and to what convenient value would you round it?
6. An article reports fasting plasma glucose for 64 patients as 112 ± 3 mg/dL. If 3 is the standard error, what is the standard deviation? Which interpretation (SD or SE) is more plausible, and why? Calculate the standard error of SBP for the women in mg2026 (SD 14.47 mmHg, n = 86).
7. The mean age of the patients in three wards of a hospital is 45.2 years (120 patients), 61.8 years (60 patients) and 70.4 years (20 patients). Calculate the mean age of all 200 patients and compare it with the simple average of the three means.
8. Using Table 2.11: (a) check the percentage of women in arm B; (b) identify the variables reported as median (Q1-Q3) and explain the choice; (c) explain why p values are not calculated for the comparisons in Table 1 of a randomized trial.
9. Choose the appropriate graph for: (a) the distribution of blood groups in 500 donors; (b) the comparison of blood glucose between three treatment arms; (c) the relationship between age and systolic blood pressure; (d) the monthly course of influenza cases in a county over two years; (e) the distribution of birth weight in 1,200 newborns; (f) the responses to a 5-point Likert question, compared between women and men.
10. The HIV viral load in four patients is 1,000, 10,000, 100,000 and 1,000,000 copies/mL. Calculate the arithmetic mean and the geometric mean (use common logarithms). Why do laboratories and clinicians express these values on a logarithmic scale?

Common Pitfalls

- Calculating the mean of nominal variables coded with numbers (for example, "mean sex = 1.46").
- Reporting mean ± SD for ordinal variables (Likert scores, stages) or for highly skewed variables (length of hospital stay, inflammatory markers).
- Confusing the standard deviation with the standard error; "mean ± SE" makes the data look much more homogeneous.
- Dividing by n instead of n - 1 when calculating the variance of a sample (STDEVP instead of STDEV).
- Taking a value as the median when the cumulative frequency reaches exactly 50% at that value; the correct median is the mean of that value and the next one.
- Removing extreme values just because they "spoil the mean", without showing that they are errors.
- Bar charts with a truncated axis, 3D pie charts or pie charts with too many categories.
- Calculating the coefficient of variation for variables on an interval scale (temperature in °C).
- Percentages that cannot be checked because the absolute numbers are missing.

Key Points

- The type of variable (nominal, ordinal, discrete, continuous) determines the permitted measures, graphs and tests.
- Data are organized with one row per patient and one column per variable, with a data dictionary written before collection; errors are checked against the source document.
- Relative and cumulative frequencies can be reconstructed from one another; cumulative frequency makes sense only for ordered data.
- The mean is appropriate for symmetric numerical data; the median for skewed or ordinal data; the mode for nominal data; the geometric mean for positive data with multiplicative variation.
- The standard deviation describes the variability of the patients; the standard error $\left(\frac{s}{\sqrt{n}}\right)$ describes the precision of the mean and decreases as n increases.
- The five-number summary and the box plot show the center, dispersion, skewness and extreme values; the 1.5 x IQR rule flags values to be checked, not deleted.
- An s close to m for a positive variable signals skewness; in this case the median (Q1-Q3) is reported.
- In "Table 1": mean ± SD for symmetric variables, median (Q1-Q3) for skewed ones, n (%) for qualitative ones; the numbers must add up.
- Bar charts start at zero, with no 3D effects; every graph has labeled axes, units and n.

Self-Assessment Questions

1. (SC) The stage of liver fibrosis (F0, F1, F2, F3, F4) is a variable that is:
    a) nominal
    b) ordinal
    c) discrete numerical on a ratio scale
    d) continuous numerical
    e) dichotomous
2. (SC) For the variable "length of hospital stay", with a mean of 11.6 days and a standard deviation of 7.0 days, the most appropriate measure of central tendency is:
    a) the arithmetic mean
    b) the mode
    c) the median

   d) the range
   e) the coefficient of variation

3. (SC) In a frequency table for 40 patients, a class has a relative frequency of 22.5%. Its absolute frequency is:
   a) 22
   b) 9
   c) 18
   d) 10
   e) 4

4. (SC) The standard deviation of SBP in a sample of 100 patients is 15 mmHg. The standard error of the mean is:
   a) 0.15 mmHg
   b) 1.5 mmHg
   c) 15 mmHg
   d) 150 mmHg
   e) 3.87 mmHg

5. (SC) For the series 2, 3, 3, 4, 5, 6, 7, 8, the median is:
   a) 4
   b) 4,5
   c) 4,75
   d) 3
   e) 5

6. (MC) Which of the following statements about the median are correct?
   a) it is equal to the second quartile and to the 50th percentile
   b) it is little influenced by extreme values
   c) it is the appropriate measure for ordinal variables
   d) it is always greater than the mean
   e) it can be calculated for nominal variables

7. (MC) Which of the following statements about the standard deviation are correct?
   a) it has the same unit of measurement as the data
   b) it is the square root of the variance
   c) it decreases as the sample size increases, just like the standard error
   d) in a sample it is calculated with n - 1 in the denominator
   e) it cannot be negative

8. (MC) In a positively skewed distribution (tail to the right):
   a) the mean is, as a rule, greater than the median
   b) the coefficient of skewness is positive
   c) rare large values pull the mean up
   d) the mean and the median are equal
   e) the median with IQR describes the data better than the mean ± SD

9. (MC) Which of the following graphs are appropriate for a continuous numerical variable?
   a) the histogram

b) the box plot

c) the stem-and-leaf plot

d) the pie chart

e) the dot plot for small samples

10. (MC) In the mg2026 dataset, BMI has a mean of 29.04 kg/m², a median of 28.25 kg/m², Q1 26.2, Q3 31.4 and a maximum of 48.3 kg/m². Which of the following statements are correct?

a) the interquartile range is 5.2 kg/m²

b) the upper limit according to the 1.5 x IQR rule is 39.2 kg/m²

c) the value 48.3 kg/m² must be removed from the analysis because it is extreme

d) the distribution is skewed to the right

e) the correct way to report it in Table 1 is 28.25 (26.2-31.4) kg/m²

Chapter 3. Statistical Distributions: The Normal Distribution, Checking Normality, Skewness, Kurtosis, the z-Score and Data Transformation

Objectives

By the end of this chapter, the student will be able to:

- distinguish a discrete from a continuous random variable and explain what the probability function, the density and the cumulative distribution function describe;
- calculate simple probabilities with the binomial and Poisson distributions, starting from examples from family practice;
- list the properties of the normal distribution and apply the 68.3-95.4-99.7 rule to real data;
- calculate a z-score, find the corresponding percentile in the table in Appendix E and in Google Sheets, and interpret the result clinically;
- construct a reference interval of the form $m \pm 1.96s$ and explain why 5% of healthy people fall "outside the normal range";
- check the normality of a variable with graphs (histogram, Q-Q plot, box plot), with indicators (skewness, kurtosis and their standard errors) and with tests (Shapiro-Wilk, Kolmogorov-Smirnov with the Lilliefors correction, Anderson-Darling), knowing the limitations of each method;
- choose an appropriate transformation (logarithmic, square root, reciprocal, Box-Cox) and back-transform the results correctly, including the geometric mean and its confidence interval;
- recognize extreme values with the z-score and the 1.5 x IQR rule and make a reasoned decision about what to do with them;
- critically read the way published articles report symmetric and skewed variables.

3.1 Random Variables and Probability Distributions

In Chapter 2 we described the data as they are: frequency tables, means, medians, histograms. Now we take a step back and ask where these shapes come from. Why does the histogram of systolic blood pressure in the mg2026 dataset look like a bell, while that of triglycerides has a long "tail" to the right? The answer lies in the notion of a distribution, that is, the law according to which the values of a variable are spread in a population. Almost everything that follows in this book, from confidence intervals to statistical tests, rests on assumptions about distributions, so it is worth understanding them well from the outset [19,63].

3.1.1 From Observed Frequency to Probability

When we say that 29 of the 160 patients in mg2026 have diabetes confirmed by OGTT, we are describing the sample: the relative frequency is 29/160 = 18.1%. If we imagine choosing at random a patient from the same population of adults at cardiometabolic risk, the probability that this patient has diabetes is about the same, around 0.18. Probability is, in a practical sense, the relative frequency we approach as the number of observations becomes very large. A coin tossed 10 times may come up heads 7 times; tossed 10,000 times, it will come up heads very close to 50% of the time.

Probability takes values between 0 (impossible event) and 1 (certain event). Two rules are enough for this chapter. The probability that an event does NOT occur is 1 minus the probability that it occurs. For independent events, the probability that both occur is the product of their

probabilities: if two patients are chosen independently, the probability that both have diabetes is 0.18 x 0.18 = 0.0324.

3.1.2 Discrete and Continuous Random Variables

A random variable is a quantity whose value we do not know in advance, but for which we know what values it can take and with what probabilities. The number of patients with diabetes in a group of 10 is a discrete random variable: it can be 0, 1, 2 and so on up to 10, but it cannot be 2.7. Likewise, the number of adverse reactions reported in a month in a practice, the number of births per day in a maternity hospital or the number of colonies on a Petri dish are discrete variables (counts).

Blood pressure, blood glucose, weight or time to thrombolysis are continuous variables: between any two possible values there is, in theory, an infinity of other values. The fact that the blood pressure monitor displays whole numbers is a property of the instrument, not of the phenomenon. This distinction matters because the way we calculate probabilities differs: for discrete variables we add up the probabilities of individual values, for continuous ones we calculate areas under a curve.

3.1.3 The Probability Function, the Density and the Cumulative Distribution Function

For a discrete variable, the distribution is described by a probability mass function: a list that gives, for each possible value, its probability. The sum of all these probabilities is 1. Its graph is a bar chart, as in Figure 3.1.

For a continuous variable, the probability of an exact value is zero. It makes no sense to ask what the probability is that a patient has an SBP of exactly 131.000... mmHg; it does make sense to ask what the probability is of an SBP between 130 and 140 mmHg. This is why we use the probability density function, a curve f(x) for which the area under the curve between two values a and b is the probability that the variable falls in the interval [a, b]. The total area under the curve is 1. The histogram of a large sample, drawn with relative frequencies and ever narrower classes, approaches the density.

The cumulative distribution function, denoted F(x), gives the probability that the variable is less than or equal to x:

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

It increases from 0 to 1 and corresponds to the cumulative frequencies in the frequency tables of Chapter 2. When we read the table of the normal distribution in Appendix E, we will in fact be reading values of the cumulative distribution function. The notion of percentile also comes from here: the 90th percentile of a variable is the value x for which F(x) = 0.90.

3.2 Two Discrete Distributions Useful in Medicine

3.2.1 The Binomial Distribution

Suppose that in the population of adults at cardiometabolic risk from which the mg2026 dataset comes, the prevalence of undiagnosed diabetes is 18%. A family doctor refers 10 patients chosen at random from this population for an OGTT. How many will have diabetes? We cannot know for sure, but we can calculate the probability of each possible result.

The situation meets the conditions of the binomial distribution: a fixed number of trials (n = 10 patients), each with two possible outcomes (diabetes or not), the same probability of "success" in each trial (p = 0.18), and independence between trials. The probability of obtaining exactly k cases is:

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

where $C(n,\ k) = \frac{n!}{k! \times (n-k)!}$ is the number of ways in which we can choose the k patients out of n. The mean and variance of the binomial distribution have very simple formulas:

mean = n x p; variance = n x p x (1 - p)

For our example, the mean is 10 x 0.18 = 1.8 patients, the variance 10 x 0.18 x 0.82 = 1.476, and the standard deviation √1.476 = 1.21 patients. The individual probabilities are given in Table 3.1.

Table 3.1. Binomial distribution for n = 10 patients and a prevalence p = 0.18.

k (patients with diabetes)	**0**	**1**	**2**	**3**	**4**	**5**	**6 or more**
$P(X = k)$	0,1374	0,3017	0,2980	0,1745	0,0670	0,0177	0,0037
P(X ≤ k), cumulative	0,1374	0,4392	0,7372	0,9117	0,9787	0,9963	1,0000

Let us check the first value by hand: $P(X = 0) = 0.82^{10} = 0.1374$. In other words, in almost one in seven groups of 10 patients we will find no patient with diabetes at all, even though the prevalence is 18%. The probability of finding at least one case is 1 - 0.1374 = 0.8626, and the probability of finding three or more cases is 1 - 0.7372 = 0.2628. These figures explain an observation every young doctor makes: in small groups, observed frequencies "jump" a great deal around the expected value, without this meaning anything special.

Let us apply the same logic to the whole mg2026 sample. If the true prevalence were 18%, the expected number of patients with diabetes among 160 would be 160 x 0.18 = 28.8, with a standard deviation of $\sqrt{160 \times 0.18 \times 0.82} = 4.86$. We observed 29, practically exactly the expected value. The probability of observing 35 or more would be 0.122, so even such a value would not have been surprising. Note something important for the following chapters: when n is large, the binomial distribution looks more and more like a normal distribution with mean n x p and standard deviation $\sqrt{n \times p \times (1 - p)}$. The approximate calculation with the normal distribution (with a small continuity correction) gives 0.120 for the same question, very close to the exact value. The Wald confidence interval for a proportion, which we will discuss in Chapter 4, is based on this approximation.

3.2.2 The Poisson Distribution

In a family practice, on average two adverse drug reactions are reported per month. In any given month there may be zero, one, three or six. Here we do not have a fixed number of trials: the number of patients exposed is large, and the probability that a particular patient has an adverse reaction in a particular month is very small. For such rare events, counted over an interval of time or space, the Poisson distribution is used:

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

where λ (lambda) is the mean number of events per interval, and e = 2.718... is the base of the natural logarithm. The Poisson distribution has an unusual property: its mean and variance are equal, both being λ. For λ = 2, the standard deviation is √2 = 1.41.

Table 3.2. Poisson distribution for λ = 2 adverse reactions per month.

k (reactions in a month)	**0**	**1**	**2**	**3**	**4**	**5 or more**
$P(X = k)$	0,1353	0,2707	0,2707	0,1804	0,0902	0,0527

The probability of a month with no adverse reaction at all is $e^{-2} = 0.1353$, that is, about one month in seven. The probability of four or more reactions is 0.0902 + 0.0527 = 0.1429, and five or more occur in 5.3% of months. If the doctor raised an alarm every time four reactions were counted in a month, false alarms would be triggered in almost one month in seven, even if nothing had changed. A threshold of five reactions gives false alarms in about one month in twenty. This is exactly the kind of reasoning used by pharmacovigilance and epidemiological surveillance systems: an observed count becomes a "signal" only when it is improbable under the usual regime.

The Poisson distribution is, in fact, the limiting case of the binomial distribution when n is very large and p very small, with n x p = λ. If 500 patients receive a drug, each with a probability of 0.004 of having a particular reaction in a month, the binomial distribution gives P(X = 0) = 0.1348, and the Poisson distribution with λ = 2 gives 0.1353. The difference is negligible. Over a whole year, with λ = 24, the probability of counting 30 or more reactions is 0.132; note that, for large λ, the Poisson distribution also becomes almost symmetric and approaches a normal distribution. Figure 3.1 compares the two discrete distributions.

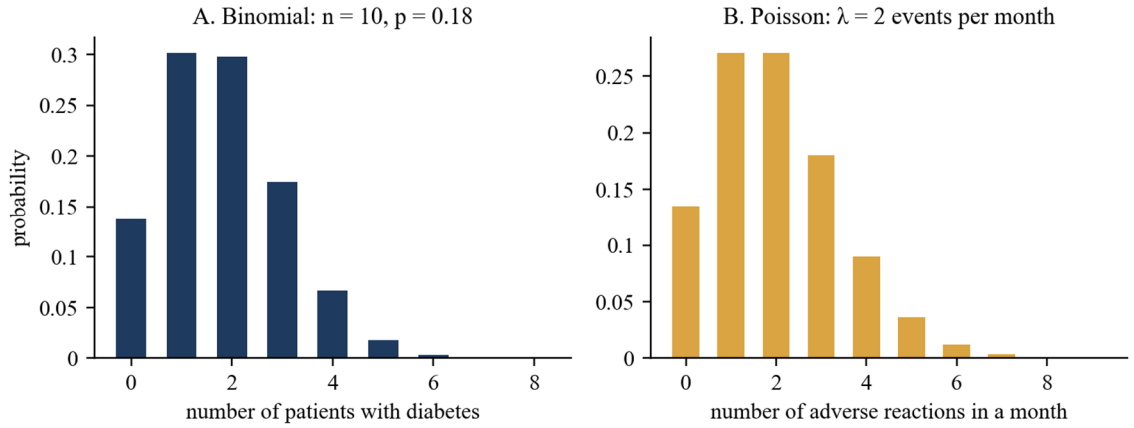


Figure 3.1. Discrete distributions. A: binomial distribution of the number of patients with diabetes in a group of 10, at a prevalence of 18%; the mean is 1.8. B: Poisson distribution of the monthly number of adverse reactions in a practice, with λ = 2. Both are skewed to the right. Theoretical values.

3.3 The Normal Distribution

3.3.1 Why It Appears So Often

The normal distribution, also called the Gaussian or Laplace-Gauss distribution, describes continuous variables that result from the sum of many small, independent influences. An adult's height depends on dozens of genes, on childhood nutrition, on illnesses, on the social environment; blood pressure depends on genetic factors, weight, salt intake, stress, the time of measurement and the nurse's technique. When many small causes add up, the result tends to be distributed symmetrically, in the shape of a bell. The mathematical explanation is the central limit theorem, which we will meet in Section 3.7.

It must be said from the outset that many biological variables are NOT normal. Serum concentrations (triglycerides, liver enzymes, CRP, insulin), durations (hospital stay, procedure times), antibody titers and the number of days of sick leave usually show positive skewness. There is no law of nature that forces medical data to be normal. The normal distribution is a model, often a good one, sometimes a poor one, and that is precisely why it must be checked every time [42,53].

3.3.2 Properties of the Normal Distribution

The density of the normal distribution has the formula:

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

You do not need to memorize it. You do, however, need to remember its consequences:

- The curve is symmetric about the mean μ. Half of the area lies to the left of the mean, half to the right.
- The mean, median and mode coincide. In a real sample they will only be close: for SBP in mg2026, the mean is 131.13 mmHg and the median 131 mmHg.
- The distribution is completely determined by two parameters: the mean μ, which gives its position, and the standard deviation σ, which gives its width. Two normal distributions with the same mean and different σ have the same center, but one is tall and narrow, the other low and wide.
- The total area under the curve is 1.
- The "tails" of the curve approach the horizontal axis as we move away from the mean but never touch it. In theory, any value is possible, however large or small, except that values very far from the mean have extremely small probabilities.
- The inflection points of the curve lie at μ - σ and μ + σ.

This last property of the tails has a practical consequence that students discover for themselves when they apply the normal model to variables that cannot be negative. A normal distribution with a mean of 11.6 days and a standard deviation of 7.0 days would predict that about 5% of patients stay in hospital a negative number of days, and 2.3% even fewer than -2.4 days, which is absurd. We will return to this example, which comes from a published article, in Section 3.8.

3.3.3 The 68.3%-95.4%-99.7% Rule

For any normal distribution, whatever μ and σ, the proportion of values lying within a given distance of the mean, measured in standard deviations, is the same (Figure 3.2):

- 68.3% of the values lie between μ - σ and μ + σ;
- 95.4% of the values lie between μ - 2σ and μ + 2σ;
- 99.7% of the values lie between μ - 3σ and μ + 3σ.

Two derived values are used just as often: exactly 95% of the values lie between μ - 1.96σ and μ + 1.96σ, and 99% between μ - 2.576σ and μ + 2.576σ. The difference between 2 and 1.96 looks like a subtlety, but you will meet it in every 95% confidence interval.

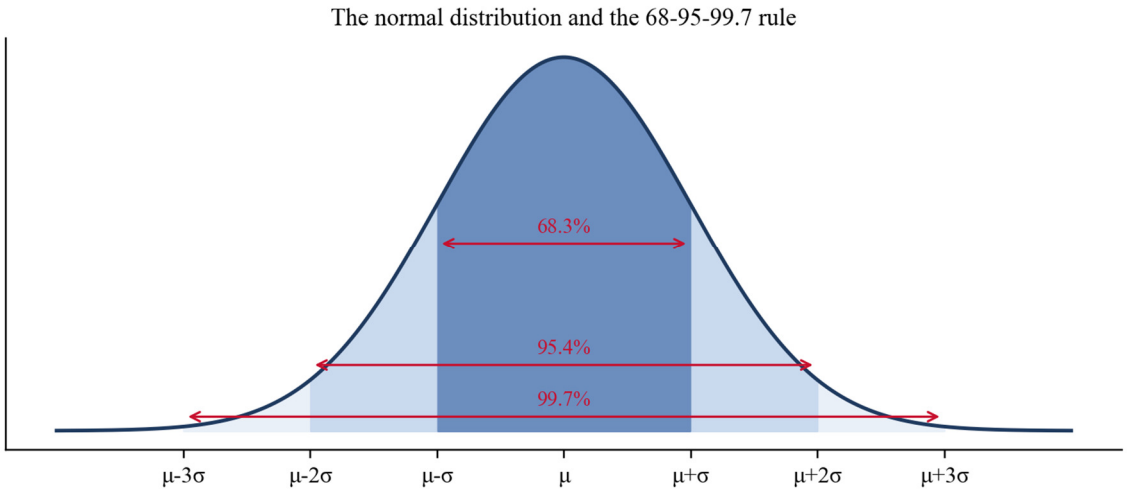


Figure 3.2. The normal distribution and the 68-95-99.7 rule: the area between μ - σ and μ + σ is 68.3% of the total area, between μ - 2σ and μ + 2σ it is 95.4%, and between μ - 3σ and μ + 3σ it is 99.7%. Theoretical curve.

Let us check the rule on the mg2026 dataset. Baseline systolic blood pressure has a mean of 131.13 mmHg and a standard deviation of 13.88 mmHg. The interval $m \pm s$ is 117.3-145.0 mmHg and contains 70.0% of the 160 patients (68.3% in theory). The interval $m \pm 2s$, that is, 103.4-158.9 mmHg, contains 93.8% (95.4% in theory). The interval $m \pm 3s$, 89.5-172.8 mmHg, contains all of them (99.7% in theory). The fit is not perfect, nor could it be, because 160 patients are a sample, not the population. It is, however, good enough for us to use the normal model for SBP. Watch the notation: μ and σ are the population parameters, which we do not know; m and s are their estimates from the sample. In practice, we replace the parameters with the estimates.

3.3.4 Observed Proportion and the Proportion Predicted by the Normal Model

A simple and very instructive test of the normal model is to compare the observed proportion of patients above a clinical threshold with the proportion predicted by a normal distribution with the mean and standard deviation of the sample. For SBP ≥ 140 mmHg, the classic hypertension threshold, the z-score is (140 - 131.13) / 13.88 = 0.64. From Table E.1, Φ(0.64) = 0.7389, so the theoretical proportion above the threshold is 1 - 0.7389 = 0.261, that is, 26.1%. In the sample, 41 of 160 patients have SBP ≥ 140, that is, 25.6%. The difference of half a percentage point is negligible.

The same calculation, repeated for other variables, tells us a great deal about the shape of the distribution (Table 3.3).

Table 3.3. Proportions observed and predicted by the normal model (mg2026 dataset, n = 160).

Variable and threshold	z-score of the threshold	Theoretical proportion	Observed proportion	Comment
SBP ≥ 140 mmHg	0,64	26,1%	25,6%	very good fit
SBP < 120 mmHg	-0,80	21,1%	20,0%	very good fit
Age ≥ 65 years	0,99	16,0%	16,3%	very good fit
Cholesterol ≥ 240 mg/dL	0,83	20,3%	23,8%	acceptable
Glucose ≥ 126 mg/dL	1,30	9,6%	11,3%	right tail slightly heavier
BMI ≥ $\frac{30\ \text{kg}}{m^2}$	0,25	40,3%	33,8%	model unsuitable

Variable and threshold	z-score of the threshold	Theoretical proportion	Observed proportion	Comment
BMI $\geq \frac{40\ \text{kg}}{m^2}$	2,80	0,3%	1,9%	model unsuitable

For BMI, the normal model errs in both directions: it overestimates the proportion of obese patients (40.3% versus 33.8%), because the mean is pulled up by a few very large values, and it severely underestimates the proportion with class III obesity (0.3% versus 1.9%, that is, 3 patients observed versus fewer than 1 predicted). This is the typical signature of a right-skewed distribution with a long tail. For blood pressure and age, the model works excellently. We will see in Section 3.8 that formal normality tests reach exactly the same conclusion.

3.4 The Standard Normal Distribution and the z-Score

3.4.1 Standardization

There is an infinity of normal distributions, one for each pair (μ, σ). It would be impossible to have a table for each of them. The solution is to reduce them all to a single one, the standard normal distribution, with mean 0 and standard deviation 1, denoted N(0, 1). The transformation is called standardization and produces the z-score:

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

or, with the sample estimates, $z = \frac{x-m}{s}$. The z-score tells us how many standard deviations separate a value from the mean, and in which direction. A positive z means a value above the mean, a negative z a value below it. The z-score has no unit of measurement: mmHg cancels out with mmHg. This is why we can compare the z-scores of completely different variables. A patient with z = 2.1 for blood glucose and z = 0.3 for cholesterol is "more unusual" for blood glucose than for cholesterol, even though the two are measured on different scales.

If a variable is normal, its z-scores follow the standard normal distribution. The mean of the z-scores of a sample is always 0 and their standard deviation is 1, even if the variable is not normal. Standardization does not, however, change the shape of the distribution: a skewed variable remains just as skewed after transformation into z-scores. It is a common mistake to believe that "standardization normalizes the data".

3.4.2 A Complete Example: An SBP of 160 mmHg

A patient in mg2026 has an SBP of 160 mmHg. How unusual is this value in the sample?

$$z = \frac{160 - 131.13}{13.88} = \frac{28.87}{13.88} = 2.08$$

The value lies 2.08 standard deviations above the mean (Figure 3.3). In Table E.1 in Appendix E we look up row 2.0 and column 0.08 and read Φ(2.08) = 0.9812. This means that, under the normal model, 98.1% of patients have an SBP less than or equal to 160 mmHg: the patient is at the 98th percentile. The proportion with higher values is 1 - 0.9812 = 0.0188, that is, 1.9%, and the area between the mean and 160 mmHg is 0.9812 - 0.5 = 0.4812.

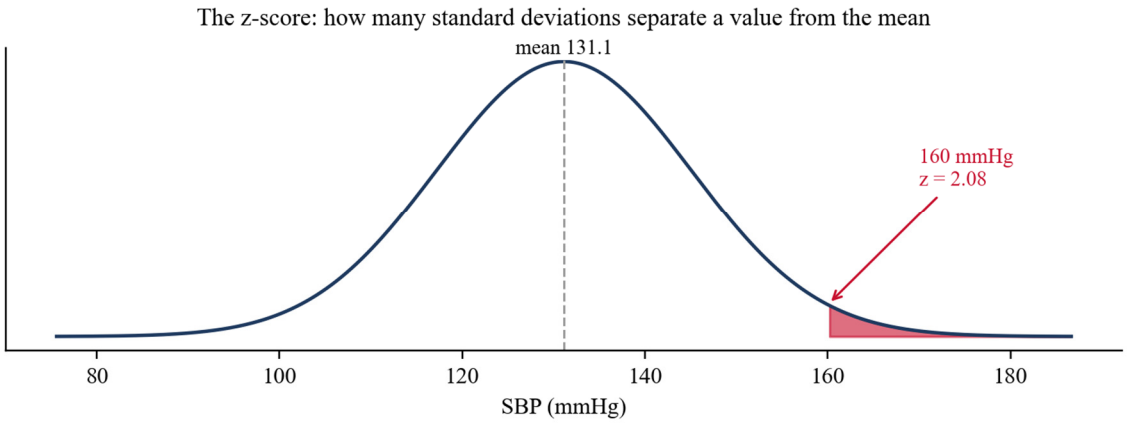


Figure 3.3. The z-score of a systolic blood pressure of 160 mmHg in the mg2026 dataset (mean 131.1 mmHg, standard deviation 13.9 mmHg) is 2.08. The shaded area to the right of the value represents the theoretical proportion of patients with a higher SBP, about 1.9%. Synthetic data (mg2026 dataset).

For values below the mean, we use the symmetry of the curve. A patient with an SBP of 108 mmHg has $z = \frac{108-131.13}{13.88} = -1.67$. The table gives Φ(1.67) = 0.9525, and by symmetry Φ(-1.67) = 1 - 0.9525 = 0.0475. The patient is at about the 5th percentile. In the sample, 6.9% of patients have SBP ≤ 108 mmHg, again close to the theoretical value.

3.4.3 The Inverse Problem: From Percentage to Value

Sometimes the question is the reverse: what is the SBP value below which 90% of patients lie? We look up in the table the z for which Φ(z) = 0.90 and find z = 1.28 (more precisely 1.2816, the value given in Table E.2). Then we work backward:

$$x = m + z \times s = 131.13 + 1.28 \times 13.88 = 148.9 \text{ mmHg}$$

The 90th percentile observed in the sample is 148.2 mmHg, and the theoretical one is 148.9 mmHg. The z values worth knowing by heart are few: 1.28 (90th percentile), 1.645 (95th percentile), 1.96 (97.5th percentile), 2.33 (99th percentile) and 2.576 (99.5th percentile). You will use them for confidence intervals, for one-sided and two-sided tests and for sample size calculation.

3.4.4 The Normal Distribution in Google Sheets

With a spreadsheet there is no longer any need for the table. Google Sheets offers four functions for the normal distribution, plus the standardization function [14,43]. With an English locale, as used in this book, the argument separator is "," and the decimal separator is the point (with the Romanian locale, ";" and the decimal comma are used).

Table 3.4. Google Sheets functions for the normal distribution, with examples using SBP from mg2026.

Function	What it calculates	Example	Result
=STANDARDIZE(x, mean, SD)	the z-score	=STANDARDIZE(160, 131.13, 13.88)	2,08
=NORM.DIST(x, mean, SD, TRUE)	area to the left of x (the cumulative distribution function)	=NORM.DIST(160, 131.13, 13.88, TRUE)	0,981
=NORM.DIST(x, mean, SD, FALSE)	height of the curve (the density), rarely used	=NORM.DIST(131, 131.13, 13.88, FALSE)	0,029
=NORM.S.DIST(z, TRUE)	area to the left of z under the standard normal curve	=NORM.S.DIST(2.08, TRUE)	0,9812
=NORM.INV(p, mean,	the value x below which	=NORM.INV(0.9, 131.13,	148,9

Function	What it calculates	Example	Result
SD)	the proportion p lies	13.88)	
=NORM.S.INV(p)	the value z below which the proportion p lies	=NORM.S.INV(0.975)	1,96

In practice, cell references are used instead of numerical values: =STANDARDIZE(M2, AVERAGE($M$2:$M$161), STDEV($M$2:$M$161)) calculates the z-score of the first patient and can be copied down the whole column. The last argument of the NORM.DIST function is a classic trap in the laboratory: with TRUE we obtain the cumulative probability (which is almost always what interests us), with FALSE we obtain the density, a number that is not a probability and has no direct interpretation. The NORMDIST and NORMSDIST functions from older versions of Excel also exist, with the same logic.

3.4.5 The z-Score in Clinical Practice

z-scores are not just a classroom exercise. Pediatric growth charts express a child's weight and height as a z-score (or percentile) relative to a reference population of the same age and sex; a child whose height is at z = -2 is at the 2.3rd percentile. Bone densitometry reports the T-score, which is a z-score calculated relative to the mean of young adults, and the Z-score, calculated relative to people of the same age. Pediatric echocardiography expresses the dimensions of the cardiac chambers and vessels as z-scores relative to body surface area. In all these situations, the physician compares the patient with a reference distribution, and the question is always the same: how far from the center does the patient's value lie, measured in standard deviations?

3.5 Reference Intervals in Laboratory Medicine

3.5.1 Where "Normal Values" Come From

On the laboratory report, next to each result, there is a reference interval, popularly called "normal values". It is usually built on a sample of healthy people, as the interval that contains the central 95% of their values [64]. If the variable is approximately normal, the interval is:

$$m - 1.96 \times s;\ m + 1.96 \times s$$

If the variable is not normal, the 2.5th and 97.5th percentiles of the reference sample are used directly (the nonparametric method), or the data are transformed, the interval is calculated on the transformed scale, and the limits are back-transformed [42,53].

The logical consequence of this construction surprises many people: by definition, 5% of perfectly healthy people have a value "outside the normal range", 2.5% below the lower limit and 2.5% above the upper one. A result slightly outside the interval does not mean disease. It only means that the value is unusual compared with the reference population.

The problem gets worse when many tests are ordered. If the k tests were independent, the probability that a healthy person has all results within the interval ar fi 0.95^k. For 5 tests, 0.95^5 = 0.774, so 22.6% of healthy people would have at least one "abnormal" result. For 12 tests, the proportion rises to 46%, and for 20 tests to 64%. Real tests are correlated with one another, so the exact figures differ, but the idea remains: a "screening" with dozens of parameters will find something to investigate in almost every healthy person. This is one of the reasons why tests should be ordered with a clinical question in mind, not "in bulk".

3.5.2 A 95% Interval for the mg2026 Dataset

Let us construct the interval $m \pm 1.96s$ for SBP in mg2026:

lower limit = 131.13 - 1.96 x 13.88 = 131.13 - 27.20 = 103.9 mmHg; upper limit = 131.13 + 1.96 x 13.88 = 131.13 + 27.20 = 158.3 mmHg

The interval 103.9-158.3 mmHg contains 150 of the 160 patients, that is, 93.8%, close to 95%. The 2.5th and 97.5th percentiles calculated directly from the data are 103.0 and 156.1 mmHg, very close to the limits obtained with the normal model. When the variable is normal, the two methods give similar results, and the parametric method has the advantage of using all the data and not just the few extreme values.

A substantive clarification is needed: the patients in mg2026 are at cardiometabolic risk, so they do not form a healthy reference population. The interval calculated describes this sample and serves us as an exercise, not as a reference interval for blood pressure. Blood pressure targets are not defined statistically but on the basis of the associated cardiovascular risk, through guidelines. The distinction between "statistically normal" (frequent) and "clinically normal" (without increased risk) is worth remembering: a cholesterol level of 220 mg/dL may be very common in a population and still be associated with increased risk.

3.5.3 What Happens with a Skewed Variable

For triglycerides (mean 161.18 mg/dL, standard deviation 76.30 mg/dL), the formula $m \pm 1.96s$ gives 11.6-310.7 mg/dL. The lower limit makes no clinical sense: nobody in the sample has triglycerides below 44 mg/dL, and 11.6 mg/dL is a practically nonexistent value in adults. On the right-hand side, 6 patients exceed the upper limit, all in the long tail. The interval contains 96.3% of the data, but it is shifted: too wide at the bottom, too narrow at the top.

If we log-transform the values (Section 3.9), the mean of the logarithms is 4.970 and their standard deviation is 0.488. The interval on the logarithmic scale is $4.970 \pm 1.96 \times 0.488$, and after back-transformation with the exponential function we obtain 55.3-374.9 mg/dL. The observed 2.5th and 97.5th percentiles are 53.8 and 364.0 mg/dL, almost identical to the limits obtained on the logarithmic scale. The back-transformed interval is no longer symmetric around the mean, and this is correct: it reflects the real shape of the distribution.

> **Case study: cut-offs on continuous distributions in liver elastography**
>
> In the noninvasive diagnosis of liver disease, the measured parameters are continuous, yet clinical decisions require cut-offs. Moga and colleagues evaluated time-harmonic elastography against transient elastography in 165 patients and proposed two cut-offs for cirrhosis: a rule-out cut-off below 1.61 m/s (7.77 kPa), with a sensitivity of 90.7% and a specificity of 66.6%, and a rule-in cut-off above 1.83 m/s (10 kPa), with a sensitivity of 65.1% and a specificity of 96.7%; the area under the ROC curve was 0.90 (95% CI 0.82-0.93) [3]. Two cut-offs instead of one are not a needless complication: they acknowledge that the distributions of values in patients with and without cirrhosis overlap. Below the first cut-off lie mostly patients without cirrhosis, above the second mostly patients with cirrhosis, and between them remains a "gray zone" where the decision requires further investigations. Nica and colleagues took a similar approach for the ultrasound-derived fat fraction (USFF) in 169 patients, with cut-offs above 5% for at least mild steatosis (AUROC 0.90), above 7.5% for moderate steatosis (AUROC 0.92) and above 10% for severe steatosis (AUROC 0.91), using CAP as the reference, with cut-offs of 248, 268 and 280 dB/m [65]. What the abstracts do not tell us: the shape of the distributions in each group and the proportion of patients in the gray zone, two pieces of information that are useful for judging whether the cut-offs can be applied in another center. We will return to them in Chapter 12, with ROC curves.

3.6 Sampling Distributions: t, Chi-Square and F

So far we have talked about the distribution of individual values: how patients' blood pressures are spread out. Inferential statistics, however, uses another type of distribution that is just as important: the distribution of a statistic computed from the sample, for example the mean, if we

were to repeat the study a very large number of times. This is called a sampling distribution. Three sampling distributions come up constantly in biostatistics, and we present them now only in terms of their shape, so that you can recognize them in the following chapters (Figure 3.4).

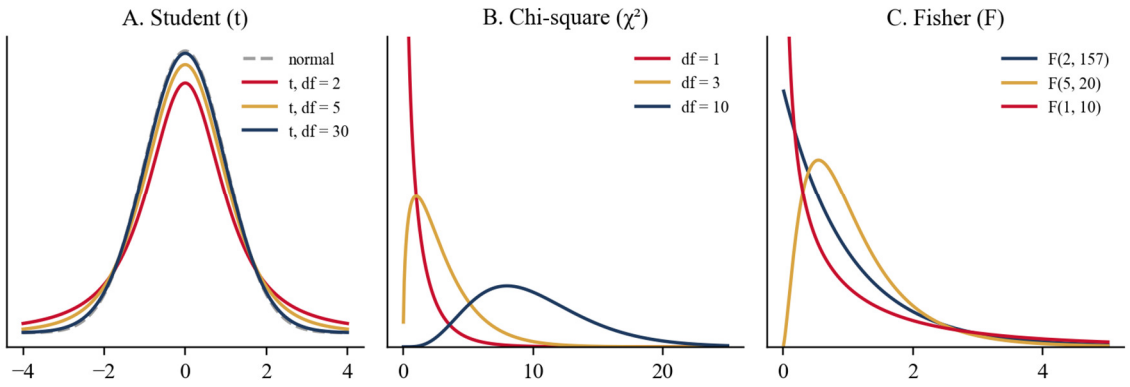


Figure 3.4. Three sampling distributions. A: Student's (t) distribution for 2, 5 and 30 degrees of freedom, compared with the standard normal (dashed line); the fewer the df, the heavier the tails. B: the chi-square distribution for 1, 3 and 10 degrees of freedom, skewed to the right. C: the Fisher (F) distribution for a few pairs of degrees of freedom. Theoretical curves.

Student's (t) distribution was published by W. S. Gosset under the pseudonym "Student". It arises when we standardize a mean using the standard deviation estimated from the sample, s, instead of the true standard deviation σ, which we do not know. It is bell-shaped and symmetric about 0, but its tails are heavier than those of the normal: estimating σ from few data adds uncertainty. Its shape depends on a single parameter, the number of degrees of freedom (df), which for a mean is n - 1. As the df increase, the t distribution approaches the standard normal. The two-sided critical value for 95% is 2.776 at 4 df, 2.262 at 9 df, 2.045 at 29 df, 1.975 at 159 df and 1.96 at infinity (Appendix E, Table E.3). These values will replace 1.96 in confidence intervals for small samples (Chapter 4) and in t-tests (Chapter 6).

Degrees of freedom deserve an intuitive explanation. If we know the mean of 10 values and we know 9 of them, the tenth is determined: it is no longer "free". When we calculate the standard deviation around the sample mean, we are already using a piece of information extracted from the data (the mean), so n - 1 free values remain. This is where the denominator n - 1 in the standard deviation formula comes from, which we will discuss in Chapter 4.

The chi-square (χ^2) distribution arises when we add up the squares of standard normal variables. It takes only positive values, it is skewed to the right, and the skewness decreases as the df increase. We will use it in the chi-square test for contingency tables (Chapter 7), where df = (r - 1) x (c - 1), and in the confidence interval for a variance. The critical value for $\alpha = 0.05$ is 3.84 at 1 df, 5.99 at 2 df and 7.81 at 3 df.

The Fisher (F) distribution is the ratio of two chi-square variables, each divided by its degrees of freedom. It has two numbers of degrees of freedom, one for the numerator and one for the denominator, takes only positive values and is skewed to the right. We will use it in analysis of variance (ANOVA, Chapter 6), in Levene's test for equality of variances and in regression. For example, for an ANOVA with three arms and 160 patients, the df are 2 and 157, and the critical value for $\alpha = 0.05$ is 3.05.

You do not need to calculate these distributions by hand. You need to know what each one is, in which test it appears, and that the critical values decrease as the degrees of freedom increase.

3.7 The Central Limit Theorem

3.7.1 The Statement, Without Formulas

The central limit theorem is probably the most important result in applied statistics. In simple words: if we draw samples of size n from a population with mean μ and standard deviation σ, whatever the shape of the population distribution, the means of these samples have a distribution that:

- is centered on μ (the sample means are not systematically too large or too small);
- has a standard deviation $\frac{\sigma}{\sqrt{n}}$, called the standard error of the mean;
- approaches a normal distribution as n increases.

The last point is the surprising part. The population can be skewed, bimodal, anything at all, and the sample means will still tend toward a bell-shaped curve. The more skewed the population, the larger the n needed. For moderately skewed distributions, 30-50 observations are usually enough; for highly skewed distributions (for example hospitalization costs, where a few patients cost tens of times more than the rest), hundreds may be needed.

3.7.2 A Simulation on Triglycerides

To see the theorem at work, we treated the 160 triglyceride values in mg2026 as a "population" and used the computer to draw from it 10,000 random samples (with replacement) of each size: 2, 5, 10, 30 and 50. For each sample we calculated the mean, and then we looked at the distribution of the 10,000 means (Figure 3.5 and Table 3.5).

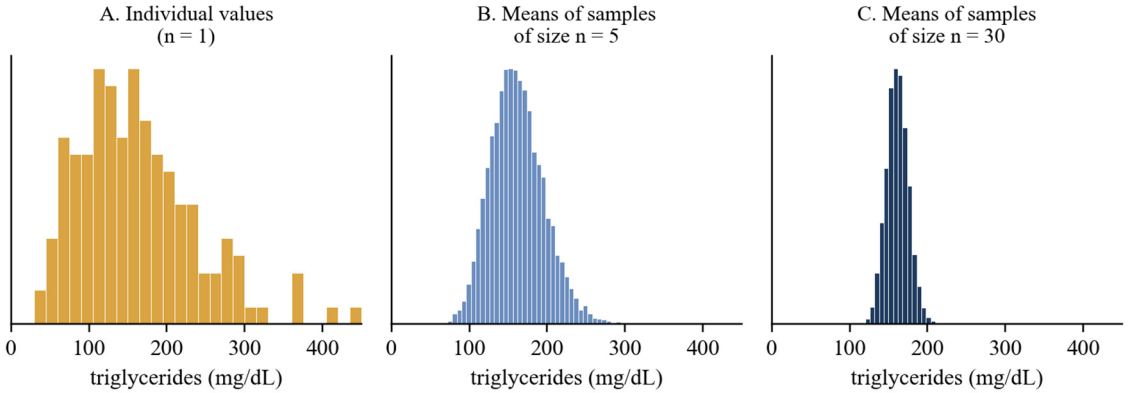


Figure 3.5. The central limit theorem on triglycerides. A: distribution of the individual values, skewed to the right (skewness 0.95). B: distribution of the means of 10,000 samples of 5 patients each, more symmetric. C: distribution of the means of samples of 30 patients each, almost normal and much narrower. Simulation on the mg2026 dataset.

Table 3.5. Simulation of the central limit theorem: 10,000 samples from the triglyceride values (mg2026).

Sample size n	Mean of the 10,000 means (mg/dL)	Standard deviation of the means (mg/dL)	$\frac{\sigma}{\sqrt{n}}$ theoretical (mg/dL)	Skewness of the means
1 (individual values)	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

Three things are immediately apparent. The mean of the means stays around 161 mg/dL, that is, the mean of the "population", whatever n is. The standard deviation of the means decreases

exactly as predicted by the formula $\frac{\sigma}{\sqrt{n}}$: at n = 30 it is 14.1 mg/dL, compared with 76.3 mg/dL for the individual values. The skewness drops from 0.95 to about 0.2. The means of samples of 30 patients are already almost normal, even though individual triglyceride values are not.

3.7.3 Why It Matters to the Physician

The central limit theorem explains why the t-test and the confidence interval for the mean work well even when the data are not perfectly normal, provided the sample is large enough. It also explains why a mean calculated on 160 patients is much more stable than an individual value. In Chapter 4 we will turn these observations into confidence intervals. Keep in mind, however, its limitation: the theorem is about means, not individual values. A reference interval for patients (Section 3.5) requires the individual values to be normal; the central limit theorem does not help us there.

You can repeat the simulation yourselves in Google Sheets, without any special software: in Practical Session 3, samples are drawn with the RAND function and their means are calculated, and in Practical Session 4 you follow how the width of confidence intervals changes.

3.8 Checking Normality

3.8.1 Why We Check and What Exactly We Check

Many methods in the following chapters (the confidence interval based on the t distribution, the t-test, ANOVA, Pearson correlation, linear regression) assume that the data or the residuals come from a normal distribution. If this assumption is seriously violated, especially in small samples, p values and confidence intervals can be misleading. That is why, before choosing a test, we look at the distribution. The check has three components, which are used together: graphical methods, numerical indicators and formal tests [49,53].

A clarification that spares a lot of confusion: what needs to be normal depends on the test. For the independent-samples t-test, we are interested in the distribution of the variable in each group, not in the whole sample. For the paired t-test, we are interested in the distribution of the differences, not of the baseline and final values. For regression, we are interested in the distribution of the residuals. Testing normality on "the whole column" is sometimes irrelevant to the test we are about to apply.

3.8.2 Graphical Methods

The histogram is the first thing we look at. For a normal variable, it shows a single peak, roughly in the middle, and two "shoulders" that slope down symmetrically. The SBP histogram in Figure 3.6A fits this description, whereas the triglyceride histogram in Figure 3.6B has its peak shifted to the left and a long tail to the right. The histogram has one drawback, however: its appearance depends on the number and width of the bins, and in small samples (under 30-40 values) it can look irregular even if the data come from a perfectly normal population.

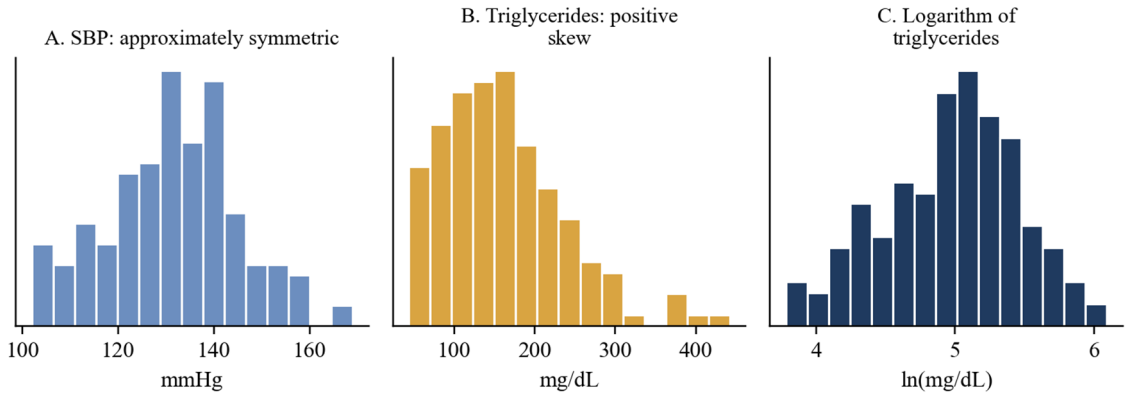


Figure 3.6. Shape of the distribution in the mg2026 dataset. A: baseline systolic blood pressure, approximately symmetric (skewness 0.01). B: triglycerides, with positive skewness (0.95) and a long right tail. C: natural logarithm of triglycerides, approximately symmetric (skewness -0.28). Synthetic data (mg2026 dataset).

The Q-Q plot (quantile-quantile plot) compares the observed quantiles with the quantiles we would expect if the data were normal. Each point is an observation: on the horizontal axis, the value it should have if the distribution were normal; on the vertical axis, its actual value. If the data are normal, the points lie on a straight line. Departures follow easily recognizable patterns:

- the points curve upward at the right end and lie above the line at the left end ("banana" or "U" shape): positive skewness, as with triglycerides;
- the reverse curvature ("n" shape): negative skewness;
- the points fall below the line on the left and rise above the line on the right, like a sideways "S": tails heavier than normal (positive excess kurtosis);
- a single point far from the line, at one end: an extreme value.

In Figure 3.7, the points for SBP follow the line almost perfectly, with small deviations at the ends, which are normal for 160 observations. The points for triglycerides clearly form a curve: the large values are larger than the normal model would predict, and the small values are less small than they should be.

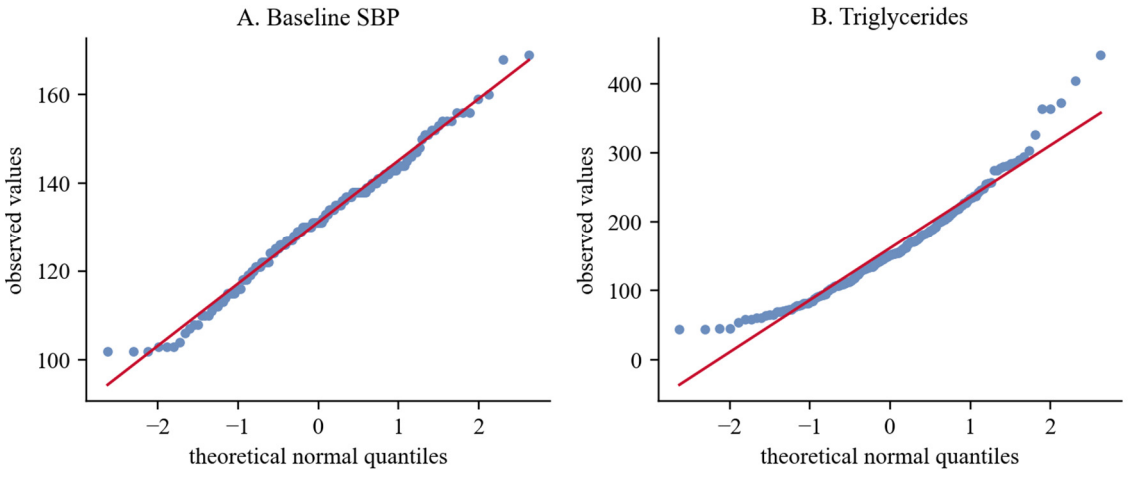


Figure 3.7. Q-Q plots for baseline systolic blood pressure (A), where the points follow the line, and for triglycerides (B), where the points form a curve characteristic of positive skewness. Synthetic data (mg2026 dataset).

The box plot, described in Chapter 2, shows skewness through the position of the median in the box (closer to Q1 when skewness is positive), through the unequal length of the "whiskers" and through the individual points outside them. It is excellent for comparing several groups and for spotting extreme values, but it says little about the exact shape of the distribution.

3.8.3 Numerical Indicators: Skewness and Kurtosis

Skewness measures how unbalanced the distribution is around the mean. For a symmetric distribution, skewness is 0. Positive values indicate a longer tail to the right (large values that are rare but very far out), negative values a tail to the left. Excess kurtosis, often called simply "kurtosis" in software, measures how heavy the tails are compared with a normal distribution: the normal has an excess kurtosis of 0, a distribution with heavy tails and a sharp peak has positive excess kurtosis, and a "flattened" one with thin tails has negative excess kurtosis. Google Sheets (the SKEW and KURT functions), JASP and MedCalc report the sample-corrected versions of these indicators, so they give the same values on the same data.

How large does skewness have to be before it should worry us? The answer depends on n, because in small samples skewness varies a lot from one sample to another even if the population is normal. That is why skewness and kurtosis are judged relative to their standard errors. For large n, approximately:

$$\text{SE(skewness)} \approx \sqrt{\frac{6}{n}} \quad \text{SE(kurtosis)} \approx \sqrt{\frac{24}{n}}$$

The exact formulas, used by software, give for n = 160 a standard error of 0.192 for skewness and 0.381 for kurtosis (the approximations give 0.194 and 0.387). The rule of thumb says that a ratio |skewness / SE| or |kurtosis / SE| greater than 2 (some authors use 1.96 or, for large samples, 3) signals a departure from normality that deserves attention.

Two other simple indicators complete the picture. The first is comparing the mean with the median: in a symmetric distribution they are close, while in one with positive skewness the mean is larger than the median (pulled by the right tail). The second is the "mean minus two standard deviations" test: for a variable that cannot be negative, if m - 2s is below zero, the distribution cannot be normal, because the normal model would predict at least 2.3% negative values, which are impossible [53].

3.8.4 Formal Normality Tests

Normality tests state the hypotheses as follows:

H0: the data come from a population with a normal distribution; H1: the data come from a population whose distribution is not normal.

A small p value (below 0.05) leads us to reject H0 and to say that the distribution departs significantly from normality. A large p value means that we do not have sufficient evidence against normality. Notice the wording: not "the data are normal", but "we found no evidence that they are not". This difference will become central in Chapter 5.

The Shapiro-Wilk test [66] calculates a statistic W, between 0 and 1, which measures how well the ordered values line up with the expected normal quantiles; it is essentially a correlation on the Q-Q plot. Values of W close to 1 indicate normality. It is the test with the best power against most departures from normality, and it is the default test in JASP. It is recommended especially for small and medium-sized samples.

The Kolmogorov-Smirnov test compares the observed cumulative distribution function (the cumulative frequencies of the data) with the cumulative distribution function of the normal, and uses as its statistic D the largest vertical distance between the two curves. The classic version assumes that the population mean and standard deviation are known in advance. When we estimate them from the same data, as is almost always the case, the classic test becomes too "lenient": it gives p values that are too large. The Lilliefors correction solves the problem. In mg2026, the classic Kolmogorov-Smirnov test applied to BMI gives p = 0.082, which would

wrongly suggest normality, whereas the version with the Lilliefors correction gives p = 0.002, and Shapiro-Wilk p < 0.001. Some programs display the classic test without saying so explicitly; it is worth checking the documentation.

The Anderson-Darling test is a variant of the Kolmogorov-Smirnov idea that gives more weight to the tails of the distribution. The statistic A^2 is compared with a critical value (for n = 160 and α = 0.05, approximately 0.75). You will come across it in MedCalc and in laboratory quality control software.

Table 3.6 brings all this information together for the main variables in mg2026.

Table 3.6. Checking normality for several variables in the mg2026 dataset (n = 160; SE of skewness 0.192; SE of kurtosis 0.381).

Variable	Mean	Median	Skewness (ratio to SE)	Kurtosis (ratio to SE)	Shapiro-Wilk W; p	Lilliefors p	Anderson-Darling A^2
Age (years)	56,29	56	0,01 (0,06)	-0,09 (-0,24)	0,993; 0,687	0,835	0,26
SBP0 (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
$\text{BMI}_0\left(\frac{\text{kg}}{m^2}\right)$	29,04	28,25	1,42 (7,42)	3,68 (9,64)	0,904; < 0,001	0,002	2,80
Triglycerides (mg/dL)	161,18	151,5	0,95 (4,97)	1,12 (2,94)	0,945; < 0,001	0,016	1,78
ln(triglycerides)	4,97	5,02	-0,28 (-1,45)	-0,28 (-0,73)	0,988; 0,200	0,364	0,50

Read row by row, the table gives clear conclusions for four of the variables. Age and SBP are compatible with normality on all criteria: skewness and kurtosis practically zero, mean equal to the median, all tests nonsignificant (Shapiro-Wilk p = 0.687 for age and p = 0.303 for SBP; classic Kolmogorov-Smirnov for SBP p = 0.823). BMI is clearly skewed: its skewness of 1.42 is more than 7 times its standard error, the kurtosis of 3.68 indicates heavy tails (a few patients with severe obesity), the mean exceeds the median by 0.8 kg/m², and all tests reject normality. Triglycerides are also skewed, but their logarithm behaves like a normal variable: we say that triglycerides have a log-normal distribution.

3.8.5 The Borderline Case: Fasting Plasma Glucose

Baseline glucose deserves a separate discussion, because here the criteria no longer agree. Shapiro-Wilk gives p = 0.037, so, strictly formally, we reject normality at the 0.05 level. By contrast, the Lilliefors test (p = 0.127) and the Anderson-Darling test ($A^2 = 0.63$, below the critical value of 0.75) do not reject it. The skewness is 0.37, with a ratio to its standard error of 1.92, just below the threshold of 2. The kurtosis is negligible. The mean (105.44 mg/dL) and the median (104.5 mg/dL) differ by less than 1 mg/dL. The glucose histogram in Figure 2.3C (Chapter 2) shows a slight stretching to the right.

What do we do? First we understand where the departure comes from. The sample contains 29 patients with undiagnosed diabetes, whose glucose is higher; they form a "bulge" in the right tail, on top of an otherwise symmetric distribution. This is a plausible biological explanation, not an artifact. Then we ask which analysis comes next. For a confidence interval for the mean or a t-test with 160 patients, a skewness of 0.37 is of no practical consequence: the central limit theorem guarantees that the mean is almost normally distributed (Section 3.7). For a reference interval for individual values, the departure would matter more, because it affects precisely the tails. Finally, we report transparently: "glucose showed a slight positive skewness (0.37; Shapiro-Wilk p =

0.037); we used parametric tests, and the results were confirmed with nonparametric tests". In Chapter 6 you will see that, for comparing glucose between the sexes, the t-test and the Mann-Whitney test lead to the same conclusion.

The general lesson is that the decision "normal or not" is not made mechanically, on the basis of a single p value, but by weighing all the evidence, taking into account the sample size and the analysis that follows [18,42].

3.8.6 Pitfalls in Interpreting Normality Tests

Normality tests have a fundamental problem: their power depends strongly on sample size, in exactly the opposite direction from what we need.

In large samples (hundreds or thousands of observations), the tests become extremely sensitive and reject normality for tiny departures of no practical importance. In a simulation with 2000 values per sample drawn from a distribution with a skewness of only 0.2, the Shapiro-Wilk test rejected normality in over 90% of the repetitions, although a t-test would have worked perfectly on such data. It is precisely in large samples, where the central limit theorem protects us, that the test alarms us most often.

In small samples (under 20-30 observations), the tests have low power and fail to detect even important departures. We drew from mg2026 three random samples of 12 triglyceride values each, a variable we know to be skewed. Shapiro-Wilk gave p = 0.084, p = 0.313 and p = 0.259: no sample "failed" the test, although one of them had a skewness of 1.40. It is precisely in small samples, where normality matters most, that the test falsely reassures us.

Worse still, a p value > 0.05 does not prove normality. It only says that, with the available data, we cannot reject the hypothesis. Statements such as "the Shapiro-Wilk test confirmed normality" are wrong, and they appear frequently in undergraduate theses and even in published articles. The correct wording is "the Shapiro-Wilk test did not indicate a significant departure from normality (p = 0.303)".

The practical conclusion: use the tests as a clue, not as a verdict. Always look at the plots. For large n, judge mainly by the plot, the skewness and the mean-median difference; for small n, be cautious even when the test is not significant, and take into account what you know about the variable from other studies (for example, that lengths of hospital stay are almost always skewed).

> **Case study: the standard deviation that "gives away" skewness**
>
> Cioca and colleagues compared the course of COVID-19 in 486 patients: 200 without diabetes, 62 with type 1 diabetes and 224 with type 2 diabetes [8]. Length of hospital stay was reported as mean ± SD: 11.6 ± 7.0 days for type 2 diabetes and 9.1 ± 5.8 days for type 1. Apply the "m - 2s" test: 11.6 - 2 x 7.0 = -2.4 days and 9.1 - 2 x 5.8 = -2.5 days. If the lengths of stay were normal, about 5% of patients would have stayed in hospital a negative number of days, which is impossible. Therefore the distribution is skewed to the right, as is almost always the case with lengths of hospital stay: most patients stay for a typical period, and a few, with complications, stay for a very long time. The median and the interquartile range would have described the data more faithfully, and the comparison should have been made with a nonparametric test or on a logarithmic scale. The same pattern appears in the study by Popa and colleagues on biomarkers in stroke, in 85 patients [61]: GFAP was 142.91 ± 102.19 pg/mL in stroke patients, and UCH $-L_1$ 1307.68 ± 967.54 pg/mL, standard deviations of 70-75% of the mean, with m - 2s well below zero. For serum biomarkers, a standard deviation close to the mean is almost always the sign of a log-normal distribution. This is not a calculation error but a reporting choice worth discussing: mean ± SD describes well only symmetric data.

3.9 Data Transformation

3.9.1 The Idea of Transformation

When a variable is skewed, we have three options: use methods that do not assume normality (nonparametric methods, Chapter 7), transform the variable so that it becomes approximately normal and then apply parametric methods, or use more advanced models that directly accommodate other distributions. Transformation means applying the same mathematical function to each value: logarithm, square root, reciprocal. The order of the values is preserved (whoever had higher triglycerides still has higher ones), but the distances between them change: large values are "compressed" more than small ones, and the long tail gets shorter [42,53].

3.9.2 The Logarithmic Transformation

The logarithmic transformation is the most widely used in medicine, because many biological processes are multiplicative: a concentration rises "twofold", not "by 50 units". One can use the natural logarithm (ln, base e) or the common logarithm (log10); the shape of the resulting distribution is the same, only the scale differs. In Google Sheets: =LN(R2) or =LOG10(R2). The logarithm requires strictly positive values; for variables that can be zero (for example the number of days of sick leave), ln(x + 1) is sometimes used, with caution in interpretation.

For the triglycerides in mg2026, log transformation reduced the skewness from 0.95 to -0.28 and the kurtosis from 1.12 to -0.28; Shapiro-Wilk goes from $p < 0.001$ to $p = 0.200$ (Figure 3.6C). On the logarithmic scale there are no extreme values either: no patient has a z score above 2.3 in absolute value.

3.9.3 Other Transformations and the "Ladder of Powers"

The usual transformations can be seen as rungs on a ladder, from the mildest to the strongest: square root $\left(\sqrt{x}\right)$, logarithm (ln x), reciprocal square root $\left(\frac{1}{\sqrt{x}}\right)$, reciprocal (1/x). The greater the positive skewness, the higher we climb on the ladder. For negative skewness (rare in medicine, for example the scores on an easy test, where most people obtain almost the maximum score), the "reflected" variable (maximum value + 1 - x) is transformed first. The square root transformation is traditionally used for counts (for example the number of colonies, the number of lesions), and the arcsine transformation for proportions, although nowadays these are usually analyzed with dedicated models.

Table 3.7. Effect of transformations on triglycerides (mg2026, n = 160).

Transformation	Skewness	Kurtosis	Shapiro-Wilk p
no transformation	0,95	1,12	< 0,001
square root $\sqrt{x}$	0,32	-0,10	0,181
logarithm ln x	-0,28	-0,28	0,200
Box-Cox, $\lambda = 0.22$	-0,01	-	0,710
reciprocal 1/x	1,48	2,30	< 0,001

The square root and the logarithm are both acceptable, with skewness values of opposite signs and similar magnitude. The reciprocal is too strong: it turns a moderate positive skewness into an even larger one (the reciprocal reverses the order, and the small values, of 44-60 mg/dL, now become the long tail). The Box-Cox transformation generalizes the ladder: it automatically searches for the exponent λ that makes the data as close to normal as possible ($\lambda = 1$ means no transformation, $\lambda = 0.5$ square root, $\lambda = 0$ logarithm, $\lambda = -1$ reciprocal). For triglycerides, the optimal λ is 0.22, between the square root and the logarithm. In practice, we choose a transformation with a simple interpretation (usually the logarithm) rather than an "exotic"

exponent such as 0.22, even if the latter gives a somewhat larger p; a result expressed in "triglycerides to the power of 0.22" cannot be communicated to anyone.

Transformation does not always solve the problem. BMI in mg2026 remains skewed after log transformation (skewness 0.89, Shapiro-Wilk $p < 0.001$), and Box-Cox calls for an exponent of -2.5, that is, a very strong transformation with no clinical interpretation, after which Shapiro-Wilk gives $p = 0.076$. The skewness of BMI comes largely from a few patients with severe obesity, not from a multiplicative process. Here the better choice is a nonparametric method or reporting the median and IQR.

3.9.4 Back-Transformation: The Geometric Mean

After analyzing the data on the logarithmic scale, the results have to be brought back to the original scale so that a clinician can understand them. The mean of the logarithms of the triglycerides is 4.970. The inverse function of the natural logarithm is the exponential:

$$e^{4.970} = 144.0 \text{ mg/dL}$$

This value is not the arithmetic mean of the triglycerides (161.18 mg/dL). It is the geometric mean, that is, the nth root of the product of the n values. For a log-normal distribution, the geometric mean is close to the median (151.5 mg/dL) and describes the "typical" value better than the arithmetic mean, which is pulled upward by the right tail.

The standard deviation on the logarithmic scale (0.488) is not back-transformed into a standard deviation. Its exponential, $e^{0.488} = 1.63$, is a multiplicative factor called the geometric standard deviation: about two thirds of patients have triglycerides between 144.0 / 1.63 and 144.0 x 1.63, that is, between 88 and 235 mg/dL. Notice the asymmetry of the interval around 144: 56 mg/dL below, 91 mg/dL above.

The confidence interval for the geometric mean is calculated on the logarithmic scale and then back-transformed. The standard error of the mean of the logarithms is 0.488 / √160 = 0.0386, and the critical t value for 159 df is 1.975, so:

$4.9699 \pm 1.975 \times 0.03859 = 4.9699 \pm 0.0762$, that is, 4.8937 to 5.0461

Back-transforming the limits, $e^{4.8937} = 133.44$ and $e^{5.0461} = 155.42$ mg/dL, so the 95% CI for the geometric mean is 133.44-155.42 mg/dL. We deliberately kept four decimal places on the logarithmic scale: the exponential amplifies small errors, and rounding to two decimal places before back-transformation would shift each limit by almost 0.5 mg/dL. The rule of thumb is that on the logarithmic scale you round only at the end, after back-transformation. The back-transformed interval is not symmetric around 144.0, and this is natural. We will return to the calculation in Chapter 4.

One last clarification: a difference on the logarithmic scale becomes, after back-transformation, a ratio. If the mean of the logarithms of triglycerides were 0.10 higher in one group than in another, back-transformation gives $e^{0.10} = 1.105$: the geometric mean is about 10.5% higher, not "0.10 mg/dL higher". This is why studies that analyze biomarkers on a logarithmic scale report ratios of geometric means.

3.9.5 When to Choose Nonparametric Methods Instead of Transformation

Transformation is useful when there is a natural, easily interpreted transformation (the logarithm for concentrations), when we want to estimate parameters with confidence intervals, and when multivariable analyses (regression) will follow. Nonparametric methods, based on ranks, are preferable when no simple transformation normalizes the data (BMI in mg2026), when the variable is ordinal (Likert scores, stages), when the sample is small and we cannot judge the shape of the distribution, when there are extreme values that we cannot exclude, or when the data have a

"ceiling" or a "floor" (many values equal to the minimum or maximum of the scale). The price of nonparametric methods is somewhat lower power when the data are in fact normal, and the fact that they compare distributions or medians, not means. The complete algorithm for choosing a test is presented in Chapter 8.

Case study: skewed data reported as median (IQR)

Gherbon and colleagues followed 200 patients with type 2 diabetes treated with SGLT-2 inhibitors for one year [59]. Glucose decreased from 180.00 mg% (IQR 154.50-207.00) to 130.00 mg% (IQR 117.50-150.00), HbA1c from 8.40% (IQR 7.98-9.15%) to 7.30% (IQR 6.90-7.95%), and SBP from 140.00 mmHg (130.00-160.00) to 120.00 mmHg (115.50-130.00), all with $p < 0.001$. Notice that the authors chose the median and IQR for the skewed variables, but mean ± SD for waist circumference (107.05 ± 16.39 cm, then 102.50 ± 15.11 cm; $p = 0.042$), which was probably distributed almost symmetrically. This is the right choice: the summary measure matches the distribution of each variable. Look at the quartiles of baseline glucose: 180 - 154.5 = 25.5 mg% below the median, 207 - 180 = 27 mg% above it, a slight asymmetry. For SBP, Q1 = 130 lies 10 mmHg below the median of 140, and Q3 = 160 lies 20 mmHg above it: the distribution is stretched to the right. One detail for critical reading: the abstract calls these values "means", although they are medians; this kind of terminological slip is common, and you should learn to recognize it. Frandeş and colleagues did the same in a cross-sectional study of 304 adults with diabetes, reporting digital engagement scores as median [Q1-Q3]: 1.71 [1.71-2.39] versus 3.86 [3.86-4.43] [7]. The fact that the median coincides with Q1 in both groups shows that at least a quarter of the participants have exactly the minimum value of their group, a distribution with a "floor" for which the mean ±SD ar fi would have been misleading; the authors consistently used beta regression, which is suited to bounded scores. Medians were also used to compare values of the TyG index across three endocrine-metabolic phenotypes in 387 patients evaluated for thyroid dysfunction: 8.24, 8.74 and 8.87, $p < 0.001$ [67].

3.10 Extreme Values

3.10.1 How to Recognize Them

An extreme value (outlier) is an observation that lies unusually far from the rest of the data. It may be an error (a typing error, a wrong unit of measurement, an instrument error), a real but rare value (a patient with severe obesity in a prevention study), or a sign that the distribution actually has a long tail. The first step is always to identify it; the second, to understand where it comes from. In mg2026_brut.csv, the SBP value of 1540 and the height of 1.72 are obvious errors (an extra zero, meters instead of centimeters), corrected in Chapter 2. But what do we do with values that are plausible yet far out?

Three simple criteria are in common use:

- The z score: values with $|z| > 3$ are suspect. The criterion assumes an approximately normal distribution; in a normal distribution, only 0.3% of values lie more than 3 standard deviations away. It has one flaw: the extreme value itself inflates the mean and the standard deviation, so it partly "hides" its own extremeness.
- The 1.5 x IQR rule (Tukey): the "fences" Q1 - 1.5 x IQR and Q3 + 1.5 x IQR are calculated; values outside them are flagged as extreme (the individual points on the box plot). For very extreme values, 3 x IQR is used. The advantage is that it relies on quartiles, which are not influenced by extreme values.
- Grubbs' test: a formal test for a single extreme value in a sample assumed to be normal, with the statistic $G = \frac{|x-m|}{s}$ compared with a critical value that depends on n (for n = 160 and α = 0.05,

approximately 3.54). There are also variants for several extreme values. The test is applied to nearly normal data; on a skewed variable it will find "extreme values" that are in fact the natural tail of the distribution.

Table 3.8. Extreme values in the mg2026 dataset according to the z score and IQR criteria.

Variable	Q1-Q3	1.5 x IQR fences	Values outside the fences	Values with |z| > 3	Maximum value and its z score
SBP0 (mmHg)	122-140	95-167	2 (168, 169)	0	169; 2,73
Age (years)	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
$\text{BMI}_0\left(\frac{\text{kg}}{m^2}\right)$	26,2-31,4	18,4-39,2	3 (40,0; 43,4; 48,3)	2	48,3; 4,92
Triglycerides (mg/dL)	107-203,25	-37,4-347,6	5 (364; 364; 373; 405; 442)	2	442; 3,68
ln(triglycerides)	4,67-5,31	3,71-6,28	0	0	6,09; 2,30

Notice that the IQR rule flags one or two values even for normal variables such as SBP and age. This is to be expected: even in a perfectly normal distribution, about 0.7% of values fall outside the Tukey fences, that is, one or two patients out of 160. For BMI, the Grubbs statistic for the value 48.3 is 4.92, well above the critical value, and for triglycerides the value 442 has G = 3.68, just above the threshold. After log transformation, however, the triglycerides no longer have any extreme value: the "extreme values" were the normal tail of a log-normal distribution.

3.10.2 What to Do with Them

The golden rule: extreme values are never deleted silently. The order of steps is as follows. First we check whether the value is an error, by comparing it with the source document (the patient's record, the laboratory report). If it is an error and the correct value can be recovered, we correct it and document the correction. If it is an error and cannot be recovered, we treat it as a missing value and report it. If the value is biologically plausible, we keep it. A BMI of 48.3 kg/m² is rare, but it exists; a patient with grade III obesity is exactly the kind of patient who belongs to the population under study.

For plausible values, we choose methods that tolerate them: median and IQR instead of mean and SD, nonparametric tests, logarithmic transformation, robust methods. We then perform a sensitivity analysis: we repeat the analysis with and without the extreme values and report whether the conclusion changes. If we exclude the patient with a BMI of 48.3, the mean BMI drops from 29.04 to 28.92 kg/m², the standard deviation from 3.91 to 3.61, and the skewness from 1.42 to 0.95; Shapiro-Wilk nevertheless remains significant ($p < 0.001$). Excluding a single value does not "repair" the distribution, and continuing to exclude values until the test becomes nonsignificant is a form of data manipulation. In Chapter 9 you will see how much a single extreme value can change a correlation coefficient.

Any exclusion must be specified in advance in the protocol (for example, "glucose values above 600 mg/dL will be checked against the laboratory report") and reported in the article, with the number of excluded observations and the reason.

Case study: times, scores and rank-based tests

Time variables in emergency medicine are almost always skewed: they cannot be negative, most patients fall within a typical range, and a few are delayed for a very long time. Popa and colleagues analyzed 258 eligible patients with acute ischemic stroke (out of 355 assessed), of whom only 46 received intravenous thrombolysis, a rate of 17.83% [62]. The times were described with medians: a median of 100 minutes from onset to arrival in the emergency department ($p < 0.001$) and a door-to-CT time of 15.5 minutes ($p = 0.009$).

Choosing the median is correct here: a mean would have been pulled upward by the few patients who arrived after many hours. In infectious diseases, Marin and colleagues compared immunological markers across three types of response to antiretroviral therapy in 462 adults with HIV, using the Kruskal-Wallis test, with H = 153.62 ($p < 0.001$, $\varepsilon^2 = 0.33$) and H = 115.10 ($p < 0.001$, $\varepsilon^2 = 0.25$) [68]. The CD4/CD8 ratio, a ratio of two cell counts, has by its nature a skewed distribution, and the rank-based test, accompanied by an effect size (epsilon squared), is the natural choice. In the same spirit, Gherbon and colleagues used the Kruskal-Wallis test ($p = 0.003$) for diabetes distress scores in 453 adults with type 2 diabetes, of whom 19 (4.2%) had severe distress [69]: questionnaire scores are ordinal and often skewed, and a test comparing means would have been hard to justify. What is worth noticing in all three studies: the choice of method follows the shape of the distribution and the type of variable, not habit.

3.11 Overlapping Distributions: From the z Score to Clinical Decisions

So far we have looked at a single distribution. In diagnosis, we usually have two: the distribution of a marker in patients with the disease and its distribution in those without it. How well the marker separates the two groups depends on the distance between the means, measured in standard deviations, which is exactly the idea behind the z score.

Lupușoru and colleagues evaluated contrast-enhanced ultrasound with parametric imaging (CEUS-PAT) for the diagnosis of cirrhosis in 64 people: 37 patients with cirrhosis and 27 healthy volunteers [2]. The measured index was 0.83 ± 0.09 in patients with cirrhosis and 0.49 ± 0.11 in volunteers ($p < 0.0001$). Reporting as mean ± SD is appropriate here, because the intervals $m \pm 2s$ remain within the plausible range of the index. The difference between the means, 0.34, represents more than three standard deviations: the two distributions hardly overlap. Not surprisingly, the cut-off > 0.7 gave an area under the ROC curve of 0.98, a sensitivity of 89.19% and a specificity of 100%. In Worked Problem 3.6 we will check that the normal model, applied to the two means and standard deviations, predicts very similar figures. What the study does not tell us is how well the index would perform in patients with intermediate fibrosis: comparing clear-cut patients with healthy volunteers maximizes the distance between the distributions, a phenomenon called spectrum bias, which we will discuss in Chapter 11.

Bende and colleagues measured the UGAP attenuation parameter in 179 patients, compared with CAP [58]. The means increased from one grade of steatosis to the next: 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($p < 0.001$). The difference between S0 and S1 is only 18.6 dB/m, less than one standard deviation: the distributions of the two grades overlap considerably. The difference between S0 and S3 is more than 70 dB/m, that is, around 2.5 standard deviations. The reported ROC areas follow exactly this logic: 0.83 for the S1 cut-off (192.5 dB/m), 0.90 for S2 (231 dB/m) and 0.91 for S3 (248 dB/m). The farther apart the distributions, the better a cut-off separates them.

Case study: the distribution of a difference

Muntean and colleagues followed 921 first-year university students and measured the change in BMI between the beginning and the end of the year [5]. The mean difference was -0.45 ± 1.07 kg/m² (95% CI -0.52 to -0.38; $p < 0.001$). Look at the two figures: the standard deviation (1.07) is more than twice the mean in absolute value (0.45). This means that the individual changes varied widely and were spread on both sides of zero, even though the mean clearly decreases. The abstract confirms this: 48.2% of students had a decrease in BMI, 38.7% remained stable, and 13.1% gained weight, with a clinically significant increase in 7.2%. A mean of differences describes the group, not the individual; a study reporting only that "BMI

> decreased by 0.45 kg/m² on average" would hide the fact that one in eight students gained weight. It is the distribution of the differences, not of the baseline or final values, that matters for the paired t-test (Chapter 6). And the fact that the authors analyzed the same variable both as a continuous difference (linear regression) and dichotomized as a "clinically significant increase" (logistic regression) shows that a continuous distribution can be viewed from two angles, each with its own question.

3.12 Worked Problems

Worked Problem 3.1. Percentages under the Normal Curve for SBP

Statement. In the mg2026 dataset, baseline SBP has a mean of 131.13 mmHg and a standard deviation of 13.88 mmHg, and the distribution is compatible with normality (Shapiro-Wilk p = 0.303). (a) What proportion of patients should, in theory, have an SBP between 110 and 150 mmHg? (b) Compare it with the observed proportion, 83.8%. (c) What SBP value separates off the 10% of patients with the lowest values?

Solution. First we check the condition: the normal model can be used, because the skewness (0.01) and the kurtosis (-0.10) are practically zero, and the normality tests are not significant.

(a) We calculate the z scores of the two limits:

$$z_1 = \frac{110 - 131.13}{13.88} = -\frac{21.13}{13.88} = -1.52 z_2 = \frac{150 - 131.13}{13.88} = \frac{18.87}{13.88} = 1.36$$

From Table E.1: Φ(1.36) = 0.9131 and Φ(1.52) = 0.9357, so Φ(-1.52) = 1 - 0.9357 = 0.0643. The proportion between the two limits is:

$$P(110 \leq \text{SBP} \leq 150) = \Phi(1.36) - \Phi(-1.52) = 0.9131 - 0.0643 = 0.8488$$

In theory, about 84.9% of patients have an SBP between 110 and 150 mmHg. In Google Sheets: =NORM.DIST(150, 131.13, 13.88, TRUE) - NORM.DIST(110, 131.13, 13.88, TRUE).

(b) The observed proportion is 83.8% (134 patients out of 160). The difference of 1.1 percentage points is of the size of the sampling fluctuations expected for n = 160. The normal model describes the data well.

(c) We look for the z for which Φ(z) = 0.10. By symmetry, it is the negative of the value for which Φ(z) = 0.90, that is, z = -1.28. Then x = 131.13 - 1.28 x 13.88 = 131.13 - 17.77 = 113.4 mmHg. In Sheets: =NORM.INV(0.1, 131.13, 13.88) gives 113.3 mmHg (the difference comes from rounding z). The observed 10th percentile in the sample is 112.0 mmHg.

Clinical interpretation. Almost five out of six patients at cardiometabolic risk in this sample have an SBP between 110 and 150 mmHg, and one in ten has an SBP below 113 mmHg. The figures describe the sample and do not replace the clinical thresholds in the guidelines.

Worked Problem 3.2. The Binomial Distribution in Diabetes Screening

Statement. The prevalence of undiagnosed diabetes in a population of adults at cardiometabolic risk is 18%. A family physician refers 20 randomly chosen patients for an OGTT. (a) What are the expected number of cases and the standard deviation? (b) What is the probability of finding no cases? (c) And the probability of finding 5 or more? (d) The physician finds 8 cases. Is this a surprising result?

Solution. The conditions for the binomial distribution are met: n = 20 independent trials, two possible outcomes, constant p = 0.18.

(a) Mean = n x p = 20 x 0.18 = 3.6 cases. Variance = 20 x 0.18 x 0.82 = 2.952, and standard deviation = √2.952 = 1.72 cases.

(b) $P(X = 0) = 0.82^{20}$. We calculate step by step: 0.82² = 0.6724; 0.82^4 = 0.4521; 0.82^8 = 0.2044; 0.82^16 = 0.0418; $0.82^{20} = 0.0418 \times 0.4521 = 0.0189$. The probability of finding no cases is 1.9%.

(c) P(X ≥ 5) = 1 - P(X ≤ 4). Calculating the five terms by hand is laborious; in Google Sheets we use =1 - BINOM.DIST(4, 20, 0.18, TRUE), which gives 0.2849. In about 28% of groups of 20 patients we will find at least 5 cases.

(d) P(X ≥ 8) = 1 - BINOM.DIST(7, 20, 0.18, TRUE) = 0.0177. Such a result occurs in fewer than 2% of groups if the true prevalence is 18%. It is therefore unusual and suggests either that this physician's patients are at higher risk than the reference population (for example, the physician selected older or more obese patients) or a rare chance occurrence. Notice that we have carried out a statistical test without calling it by name: we calculated the probability of a result at least as extreme under a given hypothesis. This is exactly the logic of the p value in Chapter 5.

Worked Problem 3.3. The Poisson Distribution and the Alarm Threshold

Statement. A vaccination center reports on average 1.5 minor adverse reactions per month. (a) What is the probability of a month with no reactions at all? (b) The team wants an alarm threshold such that false alarms occur in at most 5% of months. What is the smallest number of reactions at which the alarm is triggered?

Solution. Rare events, counted over a time interval, with a known mean: the Poisson distribution with λ = 1.5.

(a) $P(X = 0) = e^{-1.5} = 0.2231$. Roughly one month in four or five has no reported reaction.

(b) We calculate the successive probabilities with the formula $P(X = k) = \frac{e^{-\lambda} \times \lambda^k}{k!}$:

$$P(1) = 0.2231 \times 1.5 = 0.3347 P(2) = \frac{0.2231 \times 1.5^2}{2} = 0.2510 P(3) = \frac{0.2231 \times 1.5^3}{6}$$
$$= 0.1255 P(4) = \frac{0.2231 \times 1.5^4}{24} = 0.0471$$

P(X ≤ 3) = 0.2231 + 0.3347 + 0.2510 + 0.1255 = 0.9343, so P(X ≥ 4) = 0.0657 (6.6%), above the 5% limit. P(X ≤ 4) = 0.9814, so P(X ≥ 5) = 0.0186 (1.9%). The smallest threshold with false alarms below 5% is 5 reactions per month. In Sheets: =1 - POISSON.DIST(4, 1.5, TRUE) = 0.0186.

Interpretation. A threshold of 4 reactions would trigger a false alarm in about one month out of fifteen; a threshold of 5, in about one month out of fifty. Choosing the threshold is a trade-off: a lower threshold detects a real problem sooner, but produces more false alarms. We will meet the same dilemma with diagnostic tests (sensitivity versus specificity).

Worked Problem 3.4. A Reference Interval and Multiple Tests

Statement. In a sample of 400 healthy women, hemoglobin has a mean of 13.8 g/dL and a standard deviation of 1.0 g/dL, with an approximately normal distribution (fictitious data). (a) Construct the 95% reference interval. (b) A patient has 11.6 g/dL. Calculate the z score and the percentile. (c) A healthy woman has 12 independent tests, each with a 95% reference interval. What is the probability that all of them are "within normal limits"?

Solution. $(a)\, m \pm 1.96s = 13.8 \pm 1.96 \times 1.0 = 13.8 \pm 1.96$, that is, 11.8-15.8 g/dL (rounded to one decimal place).

(b) $z = \frac{11.6-13.8}{1.0} = -2.20$. From the table, Φ(2.20) = 0.9861, so Φ(-2.20) = 1 - 0.9861 = 0.0139. The patient is at the 1.4th percentile: only 1.4% of healthy women in the reference population have a value this low or lower. The value is unusual and justifies investigating anemia, but statistics alone does not make the diagnosis.

(c) $P(\text{ all } 12 \text{ normal }) = 0.95^{12} = 0.540$. Almost half (46%) of healthy women who have 12 independent tests will have at least one result outside the reference interval. Hence the practical advice: an isolated result slightly outside the interval, in a person without symptoms, should be repeated and interpreted in context before any conclusion is drawn.

Worked Problem 3.5. Choosing a Transformation for Triglycerides

Statement. The triglycerides in mg2026 have a mean of 161.18 mg/dL, a median of 151.5 mg/dL, a standard deviation of 76.30 mg/dL, a skewness of 0.95, a kurtosis of 1.12 and Shapiro-Wilk p < 0.001. (a) Argue why the distribution is not normal. (b) Choose a transformation based on Table 3.7. (c) Calculate the geometric mean and an interval containing about 95% of patients.

Solution. (a) Four pieces of evidence converge. The skewness divided by its standard error is 0.95 / 0.192 = 4.95, well above 2. The kurtosis relative to its standard error is 1.12 / 0.381 = 2.94, also above 2. The mean exceeds the median by almost 10 mg/dL. The Shapiro-Wilk test rejects normality. The Q-Q plot (Figure 3.7B) has the curved shape typical of positive skewness.

(b) From Table 3.7, the square root (skewness 0.32; p = 0.181) and the logarithm (skewness -0.28; p = 0.200) both normalize sufficiently. The reciprocal is too strong. Box-Cox gives the best numerical result (λ = 0.22), but with no clinical interpretation. We choose the logarithm: it is as effective as the square root, makes biological sense (concentrations vary multiplicatively) and allows back-transformation into a geometric mean and into ratios.

(c) The mean of the logarithms is 4.970, their standard deviation 0.488. The geometric mean is $e^{4.970} = 144.0$ mg/dL. The 95% interval on the logarithmic scale is $4.970 \pm 1.96 \times 0.488 = 4.970 \pm 0.957$, that is, 4.013-5.927. Back-transformed: $e^{4.013} = 55.3$ and $e^{5.927} = 374.9$ mg/dL. About 95% of patients have triglycerides between 55 and 375 mg/dL; in the sample, 95.6% of the values lie within this interval, and the observed 2.5th and 97.5th percentiles are 53.8 and 364.0 mg/dL.

Interpretation. On the original scale, the interval is not symmetric around the geometric mean (89 mg/dL below, 231 mg/dL above), exactly as the histogram shows. The naively calculated interval, 11.6-310.7 mg/dL, would have been wrong at both ends.

Worked Problem 3.6. The Normal Model Applied to Published Data

Statement. In a diagnostic study of cirrhosis [2], the CEUS-PAT index was 0.83 ± 0.09 in 37 patients with cirrhosis and 0.49 ± 0.11 in 27 healthy volunteers. At the cut-off > 0.7, the authors reported a sensitivity of 89.19% and a specificity of 100%. Assuming that the index is normally distributed in each group, what sensitivity and specificity would the model predict?

Solution. Sensitivity is the proportion of diseased individuals with a value above the cut-off. In patients with cirrhosis:

$z = \frac{0.7-0.83}{0.09} = -1.44$; Φ(-1,44) = 1 - 0,9251 = 0,0749

The proportion above the cut-off is 1 - 0.0749 = 0.925, so the predicted sensitivity is 92.5%.

Specificity is the proportion of those without the disease with a value below the cut-off. In the volunteers:

$z = \frac{0.7-0.49}{0.11} = 1.91$; Φ(1,91) = 0,9719

The predicted specificity is 97.2%.

Interpretation. The normal model predicts 92.5% and 97.2%; the study observed 89.19% and 100%. The agreement is good, given the small samples: 100% of 27 volunteers means that none exceeded the cut-off, and the model says we would expect 0.8 volunteers, that is, zero or one. The exercise shows that if we know the mean and standard deviation in two groups and the distributions are normal, we can anticipate the performance of a cut-off. It also shows the limitation: the result depends on the normality assumption and on the composition of the groups (here, definite cases versus healthy volunteers, which exaggerates the separation).

Worked Problem 3.7. The Normality Decision for HbA1c

Statement. For baseline HbA1c in mg2026: n = 160, mean 5.58%, standard deviation 0.45%, median 5.6%, skewness 0.14, kurtosis 0.16, Shapiro-Wilk W = 0.990, p = 0.304. (a) Is the normal model reasonable? (b) How many patients should, in theory, have HbA1c ≥ 6.5%? There are 4 in the sample.

Solution. (a) The ratio skewness / SE $= \frac{0.14}{0.192} = 0.73$; the ratio kurtosis / SE $= \frac{0.16}{0.381} = 0.42$; both well below 2. The mean and the median are practically equal. The Shapiro-Wilk test does not indicate a significant departure from normality. With 160 observations, the test has good power, so the nonsignificant result is informative. The normal model is reasonable.

(b) $z = \frac{6.5-5.58}{0.45} = 2.04$; $\Phi(2.04) = 0.9793$; the theoretical proportion above the threshold is 2.1% (with the exact z, 1.9%), that is, about 3 patients out of 160. We observe 4 (2.5%). Very good agreement.

Interpretation. The fact that only 4 patients have HbA1c ≥ 6.5%, even though 29 have diabetes confirmed by OGTT, is not a statistical contradiction but a clinical one: HbA1c and fasting plasma glucose identify partly different people, and undiagnosed diabetes is often at an early stage. The correlation between glucose and HbA1c will be studied in Chapter 9.

3.13 Proposed Exercises

1. SBP in mg2026 has a mean of 131.13 mmHg and an SD of 13.88 mmHg. Calculate the z scores for the values 125 and 150 mmHg and the corresponding percentiles, using Table E.1 and then the NORM.S.DIST function.
2. Total cholesterol has a mean of 209.39 mg/dL and an SD of 36.91 mg/dL (Shapiro-Wilk p = 0.927). Calculate the theoretical proportion of patients with cholesterol ≥ 240 mg/dL and with cholesterol < 190 mg/dL. In the sample, the observed proportions are 23.8% and 30.0%. Comment.
3. HDL has a mean of 49.52 mg/dL and an SD of 11.78 mg/dL. Construct the interval $m \pm 1.96s$. What proportion of patients should it contain? Calculate the theoretical proportion of those with HDL < 40 mg/dL (observed: 19.4%).
4. In the population from which mg2026 comes, the proportion of current smokers is approximately 0.33. For a group of 6 randomly chosen patients, calculate: the mean and variance of the number of smokers, the probability that there is no smoker, and the probability that there are at least 2.
5. A hospital ward records on average 0.5 serious adverse events per month. Calculate the probability that at least one event occurs in a month and the probability that 3 or more occur.
6. In a healthy reference population, serum potassium has a mean of 4.3 mmol/L and an SD of 0.35 mmol/L, with a normal distribution (fictitious data). Construct the 95% reference interval. A patient has 5.1 mmol/L: calculate the z score and the percentile.

7. In an article, the length of hospital stay in patients with type 1 diabetes is reported as 9.1 ± 5.8 days. Without any other information, what can you say about the shape of the distribution? Which summary measures would have been more appropriate?
8. For the FINDRISC score in mg2026 (n = 160): mean 10.37, median 10, skewness 0.39, kurtosis 0.74, Shapiro-Wilk p = 0.034. Calculate the ratios skewness / SE and kurtosis / SE (SE 0.192 and 0.381) and argue which descriptive statistics and which type of test you would choose, bearing in mind that the score takes only integer values between 0 and 26.
9. The mean of the natural logarithms of a variable is 4.97, and their standard deviation is 0.49. Calculate the geometric mean, the geometric standard deviation (as a factor) and the interval containing about 95% of the values on the original scale.
10. Baseline glucose in mg2026 has Q1 = 93 and $Q_3 = 116$ mg/dL, a mean of 105.44 and an SD of 15.76 mg/dL; the maximum value is 154 mg/dL. Is 154 an extreme value according to the 1.5 x IQR rule? And according to the |z| > 3 criterion? What would you do with this value?
11. Describe what the Q-Q plot would look like for: (a) a normal variable; (b) length of hospital stay; (c) a variable with tails heavier than normal; (d) a normal variable with a single typing error (an extra zero).

Common Pitfalls

- Believing that a nonsignificant normality test "proves" normality. It only means that we found no evidence against it, and in small samples the test has low power.
- Automatically rejecting parametric methods in a large sample just because Shapiro-Wilk gives p < 0.05. Look at the plot, the size of the skewness and the analysis that follows.
- Testing normality on the whole sample when the test requires normality in each group or normality of the differences.
- Confusing the cumulative distribution function with the density: in Google Sheets, NORM.DIST with the argument FALSE does not give a probability.
- Believing that standardization (z scores) turns a skewed distribution into a normal one.
- Reporting mean ± SD for variables for which m - 2s is negative, although the variable cannot be negative.
- Back-transforming the standard deviation calculated on a logarithmic scale as if it were a standard deviation on the original scale, or rounding the interval limits on the logarithmic scale before back-transformation.
- Deleting extreme values without checking and without reporting, or excluding values until the normality test becomes nonsignificant.
- Using the classic Kolmogorov-Smirnov test, without the Lilliefors correction, when the mean and standard deviation are estimated from the same data.
- Interpreting a laboratory result slightly outside the reference interval as evidence of disease.

Key Points

- The normal distribution is completely determined by μ and σ; it is symmetric, the mean, median and mode coincide, and 68.3%, 95.4% and 99.7% of the values lie within 1, 2 and 3 standard deviations of the mean (95% within 1.96 standard deviations).
- The z score = (x - m) / s expresses the distance from the mean in standard deviations; Table E.1 or the NORM.S.DIST function converts it into a percentile. In mg2026, an SBP of 160 mmHg has z = 2.08 (98th percentile).
- The binomial distribution (n trials, probability p, mean n x p) and the Poisson distribution (rare events, mean = variance = λ) describe counts.
- The reference interval $m \pm 1.96s$ leaves, by construction, 5% of healthy people outside it; for skewed variables, percentiles or the logarithmic scale are used.

- The t, χ^2 and F distributions are sampling distributions, with shapes that depend on the degrees of freedom; the central limit theorem makes means almost normal in sufficiently large samples.
- Normality is judged by plots, indicators (skewness and kurtosis relative to their standard errors, mean versus median) and tests (Shapiro-Wilk, Lilliefors, Anderson-Darling), all together.
- The logarithmic transformation often normalizes serum concentrations; the back-transformed mean is the geometric mean (triglycerides: 144.0 mg/dL), and differences become ratios.
- Extreme values are identified ($|z| > 3$, 1.5 x IQR, Grubbs), checked at the source and reported; they are not deleted silently.

Self-Assessment Questions

1. (SC) What proportion of the values of a normal distribution lies between $\mu - 2\sigma$ and $\mu + 2\sigma$?
    a) 68,3%
    b) 90,0%
    c) 95,4%
    d) 99,7%
    e) 50,0%
2. (SC) In the mg2026 dataset, SBP has a mean of 131.13 mmHg and a standard deviation of 13.88 mmHg. The z score of a value of 160 mmHg is:
    a) 1,64
    b) 2,08
    c) 2,58
    d) 28,87
    e) -2,08
3. (SC) Which of the following tests is recommended first and foremost for checking normality in small and medium-sized samples?
    a) the chi-square test of association
    b) the Shapiro-Wilk test
    c) the independent-samples t-test
    d) Levene's test
    e) the Mann-Whitney test
4. (SC) An article reports length of hospital stay as 11.6 ± 7.0 days. The most likely shape of the distribution is:
    a) normal, because the mean is reported
    b) skewed to the right (positive skewness)
    c) skewed to the left (negative skewness)
    d) uniform
    e) with negative excess kurtosis and no skewness
5. (SC) If we calculate the mean of the natural logarithms of a variable and then apply the exponential function, we obtain:
    a) the arithmetic mean
    b) the exact median
    c) the geometric mean
    d) the harmonic mean
    e) the mode

6. (MC) Which of the following statements about the normal distribution are true?
    a) it is symmetric about the mean
    b) the mean, median and mode coincide
    c) it is completely determined by the mean and the standard deviation
    d) the tails of the curve touch the horizontal axis at $\mu \pm 3\sigma$
    e) the total area under the curve is equal to 1
7. (MC) In a sample of 12 patients, the Shapiro-Wilk test gives p = 0.30. Which statements are correct?
    a) the test proves that the data are normal
    b) the hypothesis of normality is not rejected
    c) the test may have low power in such a small sample
    d) a parametric test must be used
    e) the plots (histogram, Q-Q) should also be examined
8. (MC) Which of the following indicate positive skewness?
    a) the mean is larger than the median
    b) the ratio skewness / standard error is greater than 2
    c) on the Q-Q plot, the points curve upward at the right end
    d) for a variable that cannot be negative, m - 2s is negative
    e) the excess kurtosis is equal to zero
9. (MC) Which statements about extreme values are correct?
    a) they are deleted automatically if their z score exceeds 3 in absolute value
    b) they are checked against the source document before any decision
    c) a sensitivity analysis, with and without them, is recommended
    d) the 1.5 x IQR rule is based on quartiles
    e) Grubbs' test assumes an approximately normal distribution
10. (MC) Regarding a reference interval constructed as $m \pm 1.96s$ in a healthy population, which statements are correct?

a) it contains about 95% of healthy people

b) about 5% of healthy people have values outside it

c) a value outside the interval proves disease

d) for skewed variables, the 2.5th and 97.5th percentiles or the logarithmic scale can be used

e) with 20 independent tests, about 64% of healthy people will have at least one result outside the interval

Chapter 4. Parameter Estimation and Confidence Intervals; Principles of Sample Size Calculation

Objectives

By the end of this chapter, the student will be able to:

- distinguish point estimation from interval estimation and explain the properties of a good estimator (unbiasedness, consistency, efficiency);
- calculate the standard error of the mean and not confuse it with the standard deviation;
- construct, on the same data, the 68.3%, 95.4% and 99.7% intervals for individual values and for the mean, and explain why they differ so much;
- calculate by hand a confidence interval for the mean using the t distribution and check it in JASP, MedCalc and Google Sheets;
- interpret a confidence interval correctly and recognize incorrect interpretations;
- calculate the confidence interval for a proportion using the Wald and Wilson methods and apply the rule of three for zero events;
- construct confidence intervals for the difference between two means and between two proportions, and read intervals for OR, RR, AUC and for skewed data;
- calculate the sample size for estimating a mean or a proportion and for comparing two means, taking into account power, effect size and losses during the study;
- critically read the confidence intervals reported in published articles.

4.1 Point Estimation and Interval Estimation

4.1.1 Parameters, Estimators, Estimates

In Chapter 1 we distinguished the population from the sample. The population has parameters, which are fixed but unknown values: the mean systolic blood pressure μ of all adults at cardiometabolic risk in the west of the country, the proportion π of those with undiagnosed diabetes. The sample gives us statistics, calculated from the data: the mean m, the proportion p. When we use a statistic to approximate a parameter, we call it an estimator, and the actual value obtained in our sample is called an estimate.

Point estimation gives a single number: "the mean SBP in the population is estimated at 131.13 mmHg". It is simple, but incomplete. If we repeated the study on another 160 patients, we would almost certainly obtain a different mean, perhaps 129.8 or 132.5 mmHg. A single number says nothing about how far we can rely on it. Interval estimation completes the information: "the mean SBP in the population lies, with 95% confidence, between 128.96 and 133.30 mmHg". The interval expresses the precision of the estimate, and its width tells us how much we actually know [42,53].

Current reporting guidelines for studies require the main results to be presented as estimates with confidence intervals, not just as p values. The reason will become clear by the end of the chapter: the confidence interval contains everything the p value contains and, in addition, the size of the effect and its precision [70].

4.1.2 Properties of a Good Estimator

An estimator is unbiased if, on average, over all possible samples, it hits the parameter. The sample mean is an unbiased estimator of the population mean: some samples overestimate it, others underestimate it, but without any systematic tendency. The simulation in Chapter 3 (Table 3.5) showed exactly this: the mean of the 10,000 means stayed around 161 mg/dL for every n.

This is where we find the explanation for a detail that puzzles students from the very first seminar: why do we divide by n - 1, and not by n, in the formula for the standard deviation?

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

The answer has to do with unbiasedness. The deviations are calculated from the sample mean m, not from the true mean μ, which we do not know. The sample mean is, by construction, the point about which the sum of squared deviations is smallest, so deviations from m are systematically a little smaller than those from μ. Dividing by n would underestimate the population variance. We checked this by simulation: from 20,000 samples of 5 values each drawn from a normal population with σ = 14 mmHg (hence σ² = 196), the variance calculated with n - 1 averaged 196.2, whereas the one calculated with n averaged only 157.0, that is, exactly 196 x 4/5. The n - 1 correction (Bessel's correction) removes this bias. For large n the difference becomes negligible: for SBP in mg2026, the standard deviation with n - 1 is 13.88 mmHg, and with n it would be 13.84 mmHg. In Google Sheets, STDEV (or STDEV.S) divides by n - 1, and STDEVP (STDEV.P) by n; for samples, always use the former.

An estimator is consistent if, as n increases, it gets closer and closer to the parameter. The sample mean is consistent: its standard error $\frac{\sigma}{\sqrt{n}}$ decreases toward zero as n increases. A larger sample does not guarantee a correct estimate if the sample is biased (for example, only patients from a university hospital), but it does guarantee an increasingly precise estimate of whatever it measures.

An estimator is efficient if, among unbiased estimators, it has the smallest variability. For normal data, both the mean and the median estimate the center of the distribution, but the mean is more efficient: the standard error of the median is approximately $\frac{1.25\times\sigma}{\sqrt{n}}$. For SBP in mg2026, this would be 1.25 x 1.097 = 1.37 mmHg, compared with 1.10 mmHg for the mean. To obtain the same precision with the median, we would need about 57% more patients. For skewed data or data with extreme values, the situation can be reversed, and the median becomes the more stable estimator. The choice of estimator therefore depends on the shape of the distribution, that is, on everything we discussed in Chapter 3.

4.2 The Sampling Distribution of the Mean and the Standard Error

4.2.1 The Standard Error of the Mean

The central limit theorem (Section 3.7) told us that the means of samples of size n have an approximately normal distribution, centered on μ, with standard deviation $\frac{\sigma}{\sqrt{n}}$. This standard deviation of the means is called the standard error of the mean, denoted SE or SEM. Since σ is unknown, we estimate it with s:

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

For SBP in mg2026:

$$\text{SE} = \frac{13.88}{\sqrt{160}} = \frac{13.88}{12.649} = 1.097 \text{ mmHg}$$

The standard error is almost 13 times smaller than the standard deviation, because $\sqrt{160} \approx 12.65$. It does not describe the spread of the patients, but the spread of the means we would obtain if we repeated the study many times with 160 patients each time. In Google Sheets it is calculated with =STDEV(M2:M161)/SQRT(COUNT(M2:M161)); JASP displays it under Descriptives, option "Std. error of mean", and MedCalc reports it automatically under Summary statistics.

4.2.2 Standard Deviation or Standard Error?

This is probably the most common confusion in the medical literature, and it comes up in the exam year after year. The two quantities answer different questions.

The standard deviation (SD) describes the variability of the data: how different the patients are from one another. It does not decrease as the sample grows; with more patients we simply estimate it more precisely. It is reported in the descriptive part (Table 1 of an article), in the form "SBP 131.13 ± 13.88 mmHg".

The standard error (SE) describes the precision of the mean: how much the mean would vary from one sample to another. It decreases as n increases. It is not usually reported on its own, but serves to construct the confidence interval, which is much easier to interpret.

Reporting "mean ± SE" instead of "mean ± SD" makes the data look less variable than they are, and the error bars on a graph look shorter. When you read an article, always check what the ± sign stands for. If it is not stated and n is known, you can work it out: an "SD" of 1.1 mmHg for systolic blood pressure in a sample of 160 patients is almost certainly a standard error. The recommended reporting rule is simple: mean ± SD to describe the data, mean with 95% CI to express the precision of the estimate.

4.3 The Interval for Individual Values and the Interval for the Mean

In the tradition of our discipline, the first estimation exercises are worded as follows: "estimate, with 95.4% confidence, the height values of the students" and "estimate, with 95.4% confidence, the mean height in the student population". The wordings are similar, but they ask for different things, and the difference is exactly that between SD and SE.

The interval for individual values (also called the prediction interval or, in the laboratory, the reference interval) uses the standard deviation:

$m \pm k \times s$, with k = 1 (68.3%), 2 (95.4%) or 3 (99.7%)

It answers the question: where do most patients lie? It assumes that the individual values are approximately normal (Chapter 3).

The interval for the mean uses the standard error:

$m \pm k \times \text{SE}$, with k = 1 (68.3%), 2 (95.4%) or 3 (99.7%)

It answers the question: where does the population mean probably lie? Thanks to the central limit theorem, it also works for moderately skewed data, if n is large enough.

Table 4.1. Intervals for individual values and for mean SBP (mg2026: m = 131.13 mmHg; s = 13.88 mmHg; SE = 1.097 mmHg; n = 160).

Confidence level	For individual values: $m \pm k \times s$	Width	For the mean: $m \pm k \times \text{SE}$	Width
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	4.4 mmHg

Confidence level	For individual values: $m \pm k \times s$	Width	For the mean: $m \pm k \times \text{SE}$	Width
			mmHg	
99.7% ($k = 3$)	89.5-172.8 mmHg	83.3 mmHg	127.84-134.42 mmHg	6.6 mmHg

The difference between the two columns is striking (Figure 4.1). The 95.4% interval for the values is more than 55 mmHg wide and covers patients from low-normal blood pressure to grade II hypertension. The 95.4% interval for the mean is less than 4.5 mmHg wide: we know fairly precisely where the population mean lies, even though individual patients differ greatly from one another. The ratio between the widths is $\sqrt{160} \approx 12.6$.

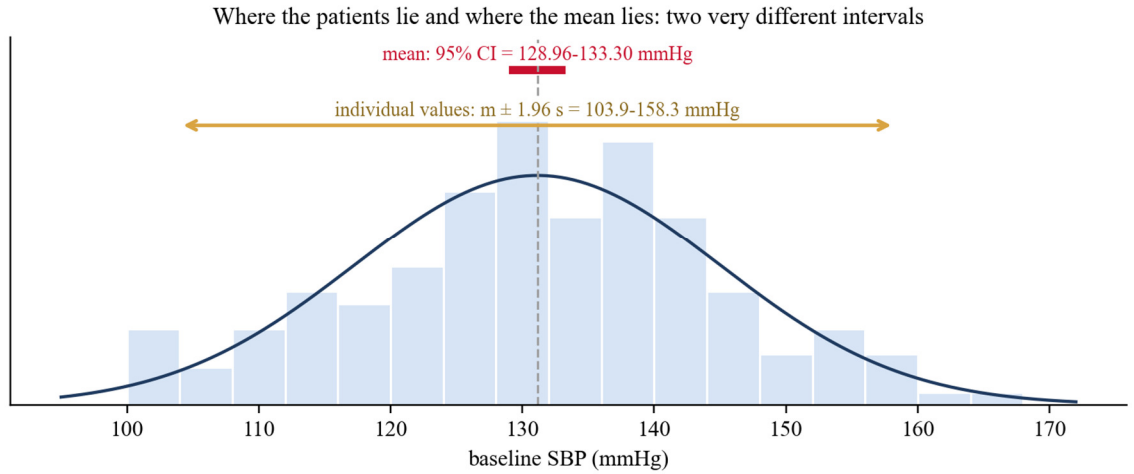


Figure 4.1. Histogram of baseline SBP in the mg2026 dataset, with the fitted normal curve. The yellow arrow shows the interval m ± 1.96 s, which contains about 95% of patients (103.9-158.3 mmHg). The short red bar shows the 95% confidence interval for the population mean (128.96-133.30 mmHg). Synthetic data (mg2026 dataset).

For the exact 95% interval, 1.96 is used instead of 2. For individual values: 103.9-158.3 mmHg (Chapter 3); for the mean, with z: $131.13 \pm 1.96 \times 1.097 = 128.98 - 133.28$ mmHg. Software, however, uses the t distribution for the mean, not the normal distribution, and the difference shows up in the second decimal place, as we will see shortly.

4.4 The Confidence Interval for the Mean Using the t Distribution

4.4.1 Why t and Not z

The formula $m \pm 1.96 \times \text{SE}$ assumes that we know σ. In reality, we estimate it with s, and this estimate also carries uncertainty, which is greater in small samples. The t distribution (Section 3.6), with its heavier tails, takes this additional uncertainty into account. The confidence interval for the mean becomes:

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

where t(df; 1 - α/2) is the critical value from Table E.3 in Appendix E. For a 95% CI, α = 0.05, and we look up the column 0.05 two-sided (or 0.975 cumulative).

Table 4.2. Critical t values for 95% and 99% confidence intervals.

n	$\text{df} = n - 1$	t for 95% CI	t for 99% CI
5	4	2,776	4,604
10	9	2,262	3,250
20	19	2,093	2,861
30	29	2,045	2,756

n	df = $n - 1$	t for 95% CI	t for 99% CI
40	39	2,023	2,708
60	59	2,001	2,662
100	99	1,984	2,626
160	159	1,975	2,607
very large	infinity	1,960	2,576

For n = 160, the difference between t (1.975) and z (1.96) is small. For n = 10, the t value of 2.262 widens the interval by 15% compared with the one calculated with z. For n = 5, by 42%. The old rule "for n ≥ 30 we use z, for n < 30 we use t" dates from the era of printed tables; today software always uses t, because it is correct for any n.

4.4.2 The Complete Calculation for SBP in mg2026

Step 1. Descriptive statistics: n = 160, m = 131.13 mmHg, s = 13.88 mmHg.

Step 2. The standard error: SE $= \frac{13.88}{\sqrt{160}} = 1.097$ mmHg.

Step 3. The critical value: df = 159; t(159; 0.975) = 1.975.

Step 4. The margin of error: E = 1.975 x 1.097 = 2.167 mmHg.

Step 5. The interval: 131.13 - 2.167 = 128.96 mmHg and 131.13 + 2.167 = 133.30 mmHg.

95% CI for mean SBP: 128.96-133.30 mmHg.

For a 99% CI, t(159; 0.995) = 2.607, the margin is 2.607 x 1.097 = 2.861 mmHg, and the interval is 128.27-133.99 mmHg. For a 90% CI, t(159; 0.95) = 1.654, and the interval is 129.32-132.95 mmHg. Notice the rule: the more confidence we demand, the wider the interval. There is no interval that is at the same time very reliable and very narrow, unless we enlarge the sample.

In Google Sheets, the margin of error is obtained directly with =CONFIDENCE.T(0.05, STDEV(M2:M161), COUNT(M2:M161)), which gives 2.167. The CONFIDENCE.NORM function (or the old CONFIDENCE) uses z instead of t and gives 2.151. The limits are then calculated as =AVERAGE(M2:M161) - CONFIDENCE.T(...) and =AVERAGE(M2:M161) + CONFIDENCE.T(...). In JASP, the interval appears under T-Tests > One Sample T-Test, with the "Location parameter" and "Confidence interval" options checked [48,49]; in MedCalc, under Statistics > Summary statistics, in the row "95% CI for the mean" [50].

For baseline glucose, with $m = 105.44$ mg/dL, $s = 15.76$ mg/dL and n = 160, SE = 1.246 mg/dL, and the 95% CI is $105.44 \pm 1.975 \times 1.246$, that is, 102.98-107.90 mg/dL. The slight skewness of glucose (Chapter 3) has practically no effect on this interval, because at n = 160 the distribution of the mean is almost normal.

4.4.3 A Small Sample: 10 Patients

Suppose we had only 10 patients, drawn at random from mg2026. Their SBP values, in order, are: 102, 103, 110, 130, 131, 136, 137, 147, 150, 154 mmHg.

The mean: the sum is 1300, so m = 130.0 mmHg.

The standard deviation: the deviations from the mean are -28, -27, -20, 0, 1, 6, 7, 17, 20, 24; their squares are 784, 729, 400, 0, 1, 36, 49, 289, 400, 576, with a sum of 3264. The variance $s^2 = \frac{3264}{9} = 362.67$, and s = 19.04 mmHg.

The standard error: SE $= \frac{19.04}{\sqrt{10}} = \frac{19.04}{3.162} = 6.022$ mmHg.

The critical value: df = 9, t = 2.262.

The margin: 2.262 x 6.022 = 13.62 mmHg.

95% CI: 130.0 ± 13.62, that is, 116.38-143.62 mmHg, with a width of 27.2 mmHg.

With z instead of t, we would have obtained 118.20-141.80 mmHg, an interval that is too narrow and would have exaggerated the precision. Compare this with the interval based on 160 patients, width 4.3 mmHg. With 10 patients, the population mean could lie anywhere between "normal" and "grade I hypertension"; with 160, we know it to a precision of ±2 mmHg.

4.5 Correct Interpretation of a Confidence Interval

4.5.1 What "95%" Means

The correct interpretation of a 95% CI is a frequentist one and refers to the procedure, not to any particular interval: if we repeated the study a very large number of times, each time with a new sample and calculating the interval in the same way, about 95% of the intervals obtained would contain the true population mean. The remaining 5% would miss it, because of sampling fluctuations.

Figure 4.2 makes this idea visible. We simulated 50 studies, each on 25 patients drawn from a population with a known mean SBP (131 mmHg) and a standard deviation of 14 mmHg, and calculated a 95% CI for each one. The intervals have different positions and widths, because each sample has a different mean and a different standard deviation. 48 of them contain the true mean; 2 (marked in red) do not. The proportion of 96% is very close to 95%, as expected. In real life we carry out a single study and do not know whether our interval is one of the 95% "good" ones or one of the 5% that "missed".

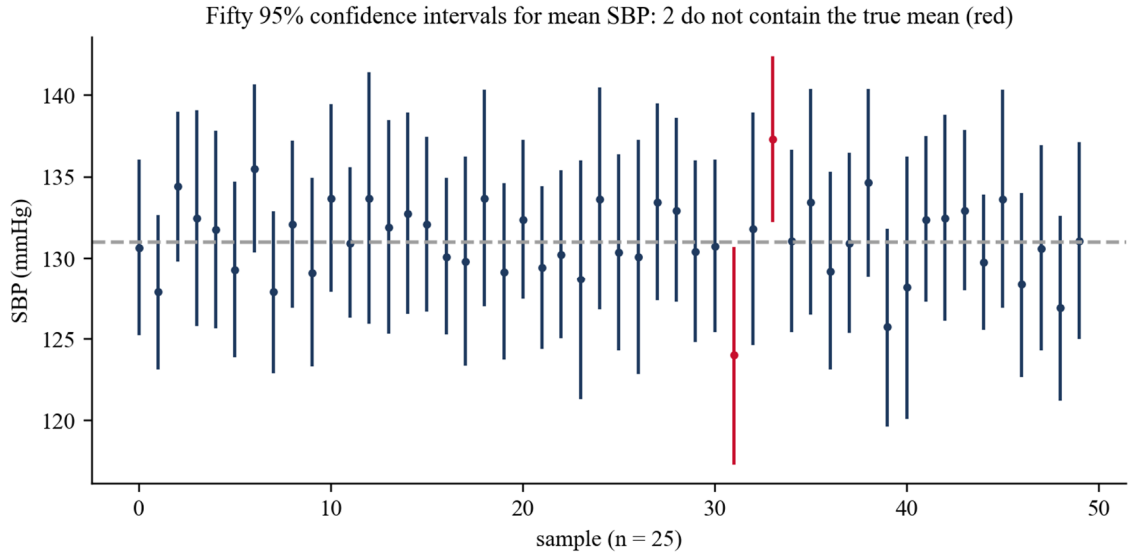


Figure 4.2. Fifty 95% confidence intervals for mean SBP, calculated on 50 simulated samples of 25 patients each from a population with a mean of 131 mmHg (dashed line). Two intervals, marked in red, do not contain the true mean. Simulation.

In everyday language, we can say: "we are 95% confident that the population mean lies between 128.96 and 133.30 mmHg", with the understanding that "confidence" refers to the method of construction, which succeeds in 95% of cases. The Bayesian approach, which is beyond the scope of this course, allows direct probability statements about the parameter, but it requires additional information (a prior distribution).

4.5.2 Incorrect Interpretations

The following statements appear frequently in papers and in exams, and they are wrong:

- "There is a 95% probability that the population mean lies between 128.96 and 133.30 mmHg." In the frequentist approach, the population mean is a constant: either it is in the interval or it is not. The 95% probability refers to the procedure.
- "95% of patients have an SBP between 128.96 and 133.30 mmHg." This is the confusion between the interval for the mean and the interval for the values. In reality, only 25 of the 160 patients (15.6%) have an SBP within this narrow interval.
- "If we repeat the study, the mean of the new sample will fall in this interval with 95% probability." No: the new sample will have its own variability, and the probability that its mean falls within our interval is, on average, around 83%, not 95%.
- "The values in the center of the interval are certain, those at the ends impossible." Values close to the point estimate are more compatible with the data than those at the ends, but there is no sharp boundary at the limits of the interval [70].
- "A wider interval means a larger effect." The width expresses precision (which depends on n and on variability), not the size of the effect.

4.6 Width of the Confidence Interval and Sample Size

The width of a CI for the mean is $\frac{2 \times t \times s}{\sqrt{n}}$. It depends on three things: the confidence level (through t), the variability of the data (through s) and the sample size (through $\sqrt{n}$). We choose the first ourselves, the second depends on biology and on the quality of measurement, and the third depends on the design of the study. The most important message is that the width decreases with the square root of n, not with n: to halve the width of the interval, we need to make the sample four times larger (Figure 4.3).

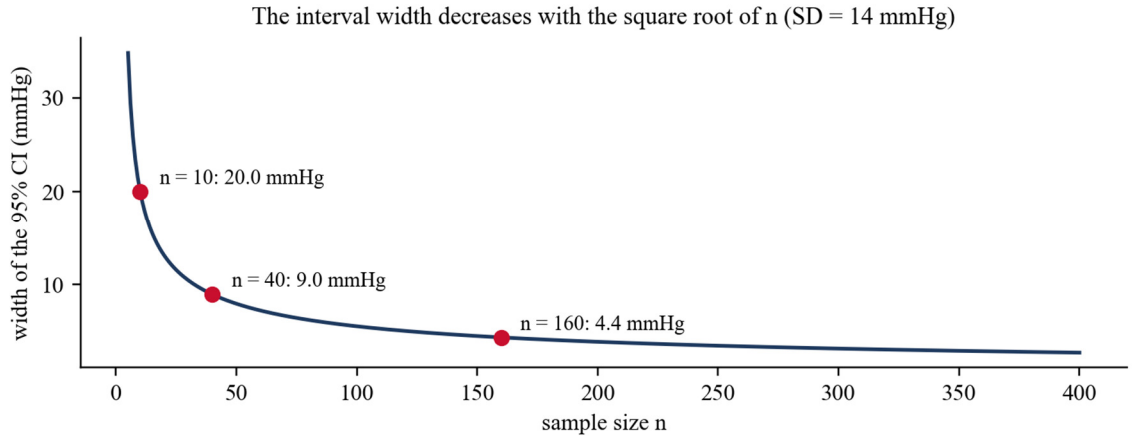


Figure 4.3. Width of the 95% confidence interval for mean SBP as a function of sample size, for a standard deviation of 14 mmHg. The curve falls quickly at first and then more and more slowly: doubling the precision requires four times as many patients. Theoretical curve.

Using the real data in mg2026, we randomly drew a subsample of 10 patients and one of 40 patients and calculated the 95% CI for mean SBP (Table 4.3).

Table 4.3. The 95% confidence interval for mean SBP according to sample size (mg2026 dataset).

n	Mean (mmHg)	95% CI (mmHg)	Width (mmHg)	Theoretical width for 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

The observed widths differ from the theoretical ones, because each subsample has its own standard deviation: the subsample of 10 patients happened to have s = 19.04 mmHg, larger than that of the full sample, and the subsample of 40 had s = 16.54 mmHg. The trend, however, is clear: from 10 to 40 patients (four times as many), the width decreases from 27.2 to 10.6 mmHg; from 40 to 160 (again four times as many), from 10.6 to 4.3 mmHg. Notice also that all three intervals contain the mean of the full sample, and the intervals from the smaller samples "catch" it with a wider margin. In Practical Session 4 you will repeat this simulation dozens of times in Google Sheets and see how the intervals vary from one sample to another.

The practical consequence for research is important: the gain in precision shows diminishing returns. Going from 40 to 160 patients gains 6.3 mmHg in width, while going from 160 to 640 would gain only 2.2 mmHg. This is why sample size calculation (Section 4.13) starts from the question "what precision do we need?", not from "how many patients can we gather?".

> **Case study: the confidence interval of a mean change**
>
> Muntean and colleagues followed 921 first-year university students and reported a mean change in BMI of -0.45 ± 1.07 kg/m², with a 95% CI from -0.52 to -0.38 kg/m² ($p < 0.001$) [5]. We can check the interval by hand. The standard error of the mean difference is 1.07 / √921 = 1.07 / 30.35 = 0.0353 kg/m². The margin of error is $1.96 \times 0.0353 = \frac{0.069\ \text{kg}}{m^2}$ (for 920 degrees of freedom, t = 1.963, practically 1.96). The interval is -0.45 ± 0.069, that is, from -0.519 to -0.381, which, once rounded, coincides with the published one. The lesson of the figures: the standard deviation of the changes (1.07) is large relative to the mean, but with 921 participants the standard error is 30 times smaller, and the interval is very narrow. We know very precisely that mean BMI decreased, by about half a unit; at the same time, 13.1% of the students gained weight. The precision of a mean says nothing about the heterogeneity of individuals. A second detail is worth noticing: the interval excludes zero, and $p < 0.001$; we will see in Section 4.12 that these two pieces of information are linked.

4.7 The Confidence Interval for a Proportion

4.7.1 The Wald Method

A proportion calculated on a sample is also an estimate, with its own standard error. If k out of n subjects in the sample have the characteristic studied, the observed proportion is $p_i = \frac{k}{n}$, and its standard error is:

$$\text{SE}(p_i) = \sqrt{\frac{p_i \times (1 - p_i)}{n}}$$

The simplest method for calculating the interval, called the Wald method, is based on the normal approximation to the binomial distribution (Section 3.2.1):

$$\text{CI}\,95\% = p_i \pm 1.96 \times \sqrt{\frac{p_i \times (1 - p_i)}{n}}$$

In mg2026, 29 of the 160 patients have diabetes confirmed by OGTT: $p_i = \frac{29}{160} = 0.181 (18.1\%)$. The standard error is $\sqrt{\frac{0.181 \times 0.819}{160}} = \sqrt{0.000927} = 0.0305$. The margin of error: 1.96 x 0.0305 = 0.0597. The Wald interval: 0.181 ± 0.060, that is, 12.2%-24.1%.

4.7.2 The Wilson Method and Why It Is Preferred

The Wald method is easy to calculate by hand, but it has well-known flaws. The interval is always symmetric, even though the distribution of a small proportion is not symmetric (it cannot go below 0, but it can rise a long way). For proportions close to 0 or 1 and for small samples, the Wald interval can go below 0% or above 100% and, more seriously, it covers the true parameter less often than it promises: a Wald "95% CI" may in reality have a coverage of 85-90%. That is why modern guidelines and software recommend the Wilson method (Wilson score) or the exact Clopper-Pearson method [42,53].

The Wilson formula looks intimidating, but it can be calculated step by step:

center = $(p_i + z^2 / (2n)) / (1 + z^2 / n)$ margin = $\left(\frac{z}{1+\frac{z^2}{n}}\right) \times \sqrt{\frac{p_i \times (1-p_i)}{n} + \frac{z^2}{4n^2}}$

For 29/160, with $z = 1.96 (z^2 = 3.8416)$:

$\frac{z^2}{n} = \frac{3.8416}{160} = 0.0240$; $\frac{z^2}{2n} = 0.0120$ center = (0.1813 + 0.0120) / 1.0240 = 0.1887 under the square root: $\frac{0.1813 \times 0.8187}{160} + \frac{3.8416}{102400} = 0.000927 + 0.0000375 = 0.000965$; square root = 0.0311 margin = $\left(\frac{1.96}{1.0240}\right) \times 0.0311 = 1.914 \times 0.0311 = 0.0595$; CI 95% Wilson: 0.1887 ± 0.0595, that is, 12.9%-24.8%

Notice two things. The center of the Wilson interval (18.9%) is slightly shifted toward 50% relative to the observed proportion (18.1%), and the interval is no longer symmetric around 18.1%: 5.2 percentage points below and 6.7 above. For n = 160 and p_i = 18%, the difference from Wald is small. The exact Clopper-Pearson method gives 12.5%-25.0%, somewhat wider, because it is conservative.

The differences become important in small samples. If 2 out of 20 patients have a complication (10%), the Wald interval is 10% ± 13.1%, that is, from -3.1% to 23.1%, an impossible lower limit that software "truncates" at 0. The Wilson interval is 2.8%-30.1%, and the exact interval 1.2%-31.7%. Both are asymmetric and realistic. The practical conclusion: for proportions, report the Wilson (or exact) interval; the Wald method is useful only for understanding the idea and for quick calculations by hand, with large n and p_i far from 0 and 1. In MedCalc, the exact interval appears under Statistics > Summary statistics for dichotomous variables or in the calculator for a single proportion; in JASP, under Frequencies > Binomial Test, which displays the exact Clopper-Pearson interval. Google Sheets has no dedicated function, but the Wilson formula can be written in a cell, as we will do in Practical Session 4.

4.7.3 Zero Events: The Rule of Three

A special situation, very common in medicine: a study of 50 patients treated with a new drug observes no serious adverse reaction. The observed proportion is 0/50 = 0%, and the Wald interval is 0% ± 0, that is, "exactly zero", which is obviously absurd. No one would argue that the true risk is zero simply because we did not see it in 50 patients.

The rule of three offers a simple solution: when no event is observed in a sample of n subjects, the upper limit of the 95% CI for the true proportion is approximately 3 / n. For 50 patients: 3/50 = 6%. The true risk could therefore be as high as about 6%, that is, up to 1 patient in 17. The exact method gives 7.1%; the rule works best for larger n (for 100 patients, 3% compared with an exact 3.6%). For 30 patients without events, the limit is 3/30 = 10%. The rule explains why rare adverse reactions (1 in 10,000) cannot be detected in phase III clinical trials with a few thousand patients, and why post-marketing pharmacovigilance is indispensable.

Case study: Wilson intervals for prevalences

Moleriu and colleagues studied 136 patients from a gastroenterology clinic, 52 with a disorder of gut-brain interaction phenotype and 84 asymptomatic controls [71]. Only 4.4% of them (95% CI 2.0-9.3%) had a formal diagnosis of irritable bowel syndrome, whereas 47.8% (95% CI 39.6-56.1%) met the criteria for a compatible symptom cluster, giving an underdiagnosis ratio of 10.8. Look at the first interval: 4.4% lies 2.4 points from the lower limit and 4.9 points from the upper one. This is the typical asymmetry of a Wilson or exact interval for a small proportion. We can check: 4.4% of 136 means 6 patients, and the Wilson interval for 6/136 is exactly 2.0-9.3%; for 65/136 (47.8%), the Wilson interval is 39.6-56.1%, again identical to the published one. A Wald interval for 6/136 would have given 1.0-7.9%, symmetric and too narrow at the top. The second interval, for a proportion close to 50%, is almost symmetric, because there the methods give similar results. Mischie and colleagues explicitly reported prevalences with Wilson intervals in an online survey of 753 young people aged 18-30 years, with 11.6% underweight and 29.3% overweight or obese [20]. As an exercise, the Wilson interval we calculated for 29.3% of 753 is approximately 26.2-32.7%. The precision is good thanks to the large sample, but the interval does not correct another shortcoming, which the authors acknowledge in the title itself: the sample is a convenience sample, with 93.5% women. The confidence interval measures only random sampling error, not systematic selection error.

4.8 The Confidence Interval for the Difference between Two Means and between Two Proportions

4.8.1 The Difference between Two Means

When we compare two independent groups, the parameter of interest is the difference between the population means, μ1 - μ2. The point estimate is m1 - m2, and its standard error combines the standard errors of the two means:

$$\mathrm{SE}(m_1 - m_2) = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

The confidence interval is $(m_1 - m_2) \pm t \times \mathrm{SE}(m_1 - m_2)$, with a number of degrees of freedom calculated by the Welch formula (software calculates it automatically) or, if the variances are equal, n1 + n2 - 2.

In mg2026, SBP is 130.85 ± 13.25 mmHg in the 74 men and 131.37 ± 14.47 mmHg in the 86 women. The difference men minus women is -0.52 mmHg. The standard error:

$$\mathrm{SE} = \sqrt{\frac{13.25^2}{74} + \frac{14.47^2}{86}} = \sqrt{2.372 + 2.435} = \sqrt{4.807} = 2.193 \text{ mmHg}$$

With 157.4 degrees of freedom (Welch), t = 1.975, and the margin is 1.975 x 2.193 = 4.33 mmHg. 95% CI for the difference: -0.52 ± 4.33, that is, from -4.9 to 3.8 mmHg.

The interval contains zero: the data are compatible both with a slightly higher SBP in men (by up to 3.8 mmHg) and with a higher one in women (by up to 4.9 mmHg), and of course with no difference at all. But the interval also tells us something more than "the difference is not significant" (p = 0.81, Chapter 6): a true difference of 10 mmHg between the sexes, in this population, is very unlikely, because it lies far outside the interval. A nonsignificant result with a narrow interval is informative; one with a very wide interval tells us almost nothing.

4.8.2 The Difference between Two Proportions

For two independent proportions $p_1 = \frac{k_1}{n_1}$ and $p_2 = \frac{k_2}{n_2}$:

$$\mathrm{SE}(p_1 - p_2) = \sqrt{\frac{p_1 \times (1 - p_1)}{n_1} + \frac{p_2 \times (1 - p_2)}{n_2}}$$

In the simulated mg2026 study, a weight loss of at least 5% at 6 months was achieved by 16 of the 54 patients in arm C (29.6%) and by 6 of the 53 in arm A (11.3%). The absolute difference is 18.3 percentage points. The standard error:

$$\mathrm{SE} = \sqrt{\frac{0.296 \times 0.704}{54} + \frac{0.113 \times 0.887}{53}} = \sqrt{0.00386 + 0.00189} = \sqrt{0.00576} = 0.0759$$

95% CI: $0.183 \pm 1.96 \times 0.0759 = 0.183 \pm 0.149$, that is, from 3.4 to 33.2 percentage points. The interval excludes zero, so the difference is statistically significant. It is, however, very wide: the true benefit of teleconsultation could be modest (3 more patients out of 100) or very large (33 out of 100). From this difference we calculate the number needed to treat, NNT = $\frac{1}{0.183} = 5.5$, with an interval from 3.0 to 29.1 patients, which we will discuss in Chapter 10.

4.9 Confidence Intervals for Skewed Data

4.9.1 The Logarithmic Scale and the Geometric Mean

For a log-normal variable, such as the triglycerides in mg2026, the confidence interval is calculated on the logarithmic scale and then back-transformed (Section 3.9.4). The mean of the logarithms is 4.970, their standard deviation 0.488, the standard error 0.488 / √160 = 0.0386, and the 95% CI on the logarithmic scale is $4.9699 \pm 1.975 \times 0.0386$, that is, 4.8937 to 5.0461. Back-transformed, we obtain $e^{4.8937} = 133.44$ and $e^{5.0461} = 155.42$ mg/dL, around the geometric mean of 144.0 mg/dL (the limits on the logarithmic scale are kept with all their decimal places until back-transformation). The interval for the arithmetic mean, calculated directly with the t formula, would be 149.26-173.09 mg/dL. The two intervals do not contradict each other: they estimate different parameters. The first estimates the "typical" center of a skewed distribution, the second the arithmetic mean, pulled upward by the right tail. When you report, state explicitly what you have estimated.

4.9.2 The Confidence Interval for the Median

For variables for which there is no suitable transformation, the confidence interval of the median can be estimated directly. The classic method, based on the binomial distribution, assumes no particular shape of the distribution. The values are sorted in ascending order and the ranks are calculated:

lower rank $\approx \frac{n}{2} - \frac{1.96 \times \sqrt{n}}{2}$; upper rank $\approx \frac{n}{2} + \frac{1.96 \times \sqrt{n}}{2}$ (rounded outward)

For n = 160: $80 \pm \frac{1.96 \times 12.65}{2} = 80 \pm 12.4$, so ranks 67 and 93. The values in these positions give the interval. For triglycerides (median 151.5 mg/dL), the 67th and 93rd ordered values are 134 and 162 mg/dL: 95% CI for the median 134-162 mg/dL. For SBP (median 131 mmHg), the interval is 129-134 mmHg, slightly wider than that of the mean (128.96-133.30 mmHg, width 4.3 versus 5), illustrating the lower efficiency of the median for normal data. MedCalc automatically calculates the confidence interval of the median in Summary statistics.

4.10 Confidence Intervals for Ratios (OR, RR) and for the ROC Area

4.10.1 Why Intervals for OR and RR Are Asymmetric

The odds ratio (OR) and the relative risk (RR), which we will study in Chapter 10, are ratios: they take values between 0 and infinity, with the value 1 meaning "no association". Their sampling distribution is skewed, but their logarithm has an approximately normal distribution. The interval is therefore calculated on the logarithmic scale and back-transformed, just as for the geometric mean.

For the obesity-diabetes association in mg2026 (2 x 2 table with a = 18, b = 36, c = 11, d = 95), $\mathrm{OR} = \frac{18\times95}{36\times11} = 4.32$. On the logarithmic scale:

$$\ln(\mathrm{OR}) = \ln(4.32) = 1.463;\ \mathrm{SE}(\ln \mathrm{OR}) = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{18}+\frac{1}{36}+\frac{1}{11}+\frac{1}{95}} =$$

$\sqrt{0.1848} = 0.430 \ln(\mathrm{OR}) \pm 1.96 \times 0.430 = 1.463 \pm 0.843$, that is, 0.620 to 2.306 95% CI for the OR: $e^{0.620} = 1.86$ to $e^{2.306} = 10.03$

The interval is strongly asymmetric around 4.32: 2.46 units below, 5.71 above. The asymmetry is not a flaw but a consequence of the multiplicative scale. A useful check: the point estimate is the geometric mean of the limits, $\sqrt{1.86 \times 10.03} = 4.32$. For the RR of weight loss in arm C versus arm A (16/54 versus 6/53), RR = 2.62, and the 95% CI, calculated in the same way, is 1.11-6.17.

4.10.2 The Interval for the Area under the ROC Curve

The area under the ROC curve (AUC), which we will study in Chapter 12, takes values between 0.5 (useless test) and 1 (perfect test). Its confidence interval is also often asymmetric when the AUC is close to 1, because the upper limit cannot exceed 1. In mg2026, the FINDRISC score has an AUC of 0.852 (95% CI 0.762-0.943) for diabetes confirmed by OGTT. Gherbon and colleagues report for age, as a predictor of severe diabetes distress, an AUC of 0.726 (95% CI 0.602-0.827), with only 19 cases, which explains the large width of the interval [69]. The reading rule is the same as for any estimate: we look first at the lower limit. An AUC of 0.726 with a lower limit of 0.602 describes a predictor ranging from "almost useless" to "acceptable".

Moga and colleagues report for time-harmonic elastography, in the diagnosis of cirrhosis, an AUROC of 0.90 (95% CI 0.82-0.93) [3]: the estimate is much closer to the upper limit than to the lower one (0.08 below, 0.03 above), a natural asymmetry for a quantity bounded at 1. The same study also reports, with confidence intervals, the reproducibility of the measurements, in a subgroup of 27 patients: an overall inter-observer agreement of 0.94 (95% CI 0.89-0.97) and intraclass correlation coefficients of 0.92 (0.82-0.96), 0.94 (0.87-0.97) and 0.97 (0.95-0.99). The interval for 0.92 spans 14 hundredths, which reflects the small size of the subgroup; an "excellent" reproducibility estimated on 27 patients remains compatible with one that is merely "good".

Rațiu and colleagues compared three methods for diagnosing malignant biliary obstruction in 160 patients, with an AUROC of 0.88 (95% CI 0.75-0.90) for MR cholangiography, 0.94 (0.85-0.99) for ERCP and 0.80 (0.70-0.82) for histology [57]. Two observations for critical reading. The intervals partly overlap, but the overlap of two intervals does not automatically mean that the difference is not significant; comparing two AUCs measured on the same patients requires a dedicated test that takes into account the correlation between them (Chapter 12). Second, for MR cholangiography and histology, the point estimate lies very close to the upper limit (0.88 in an interval that rises only to 0.90; 0.80 in one that rises only to 0.82). This pattern is unusual and may stem from the calculation method or from an editorial slip; it is worth checking in the full text before the figures are used any further.

Case study: odds ratios with confidence intervals, read correctly

Three studies by colleagues from our discipline, on students and patients in Timișoara, illustrate how to read intervals for ratios. Moleriu and colleagues analyzed 921 students, 24.1% of whom were overweight or obese, and reported adjusted odds ratios of 7.20 (95% CI 4.26-12.17) and 1.98 (95% CI 1.26-3.10), plus an OR of 1.43 per point of a dietary risk score (95% CI 1.23-1.66) [4]. Check the geometric mean property: $\sqrt{4.26 \times 12.17} = 7.20$, $\sqrt{1.26 \times 3.10} = 1.98$, $\sqrt{1.23 \times 1.66} = 1.43$. All three intervals exclude 1, so the associations are significant, but the first one is wide: the data are compatible with odds that are 4 times or 12 times higher. Piroș and colleagues found, in 693 students, a prediabetes prevalence of 21.1% and an OR of 1.78 per dietary pattern category (95% CI 1.38-2.30; $p < 0.001$) [6], a narrower interval that comfortably excludes the value 1. At the opposite end, Popa and colleagues, comparing 100 women and 113 men treated in the emergency department for suspected bacterial infections, reported an OR of 0.71 (95% CI 0.34-1.49) for a clinical outcome recorded in 14 women (14.0%) and 21 men (18.6%), and an OR of 0.87 (95% CI 0.31-2.43) for in-hospital mortality, with 7 and 9 deaths [9]. These intervals include 1 and are very wide: for mortality, the data are compatible both with odds three times lower and with odds more than twice as high in women. The correct conclusion is not "there are no differences between the sexes" but "the study does not have enough events to answer the question". The same study found, for a urinary source of infection, an OR of 2.14 (95% CI 1.20-3.81). By contrast, linear regression coefficients have symmetric intervals, because they lie on an additive scale: Cîrstea and colleagues report B = 0.076 (95% CI 0.017-0.135; $p = 0.011$) for the association between satisfaction with pharmaceutical services and treatment adherence in 220 dermatology patients [72]; the estimate lies exactly in the middle, 0.076 ± 0.059.

4.11 The Bootstrap, Explained Simply

For many statistics (the median, the ratio of two medians, the Spearman correlation coefficient, the AUC) there are no simple formulas for the standard error, or the formulas require assumptions that the data do not meet. The bootstrap method, proposed in the 1970s, gets around the problem with an almost naive idea: if our sample is representative of the population, we can simulate "repeating the study" by drawing new samples from the sample itself.

The procedure is as follows. From the 160 triglyceride values, we randomly draw 160 values with replacement: some values will appear two or three times, others not at all. We calculate the median of this "bootstrap sample". We repeat this 10,000 times. We obtain 10,000 medians, whose distribution approximates the sampling distribution of the median. The 2.5th and 97.5th percentiles of this distribution form the 95% bootstrap confidence interval (percentile method).

For the median of triglycerides in mg2026, the result is 137-162 mg/dL (Figure 4.4), very close to the interval obtained by the rank method (134-162 mg/dL). The distribution of the bootstrap medians is irregular, with "jumps", because the median can only take values from the dataset or averages of two neighboring values. For mean SBP, the bootstrap gives 128.99-133.26 mmHg, almost identical to the classic t interval (128.96-133.30 mmHg), and the standard deviation of the 10,000 bootstrap means is 1.09 mmHg, practically equal to the standard error calculated with the formula (1.097 mmHg). When the assumptions of the classic method are met, the two methods agree; when they are not, the bootstrap remains valid.

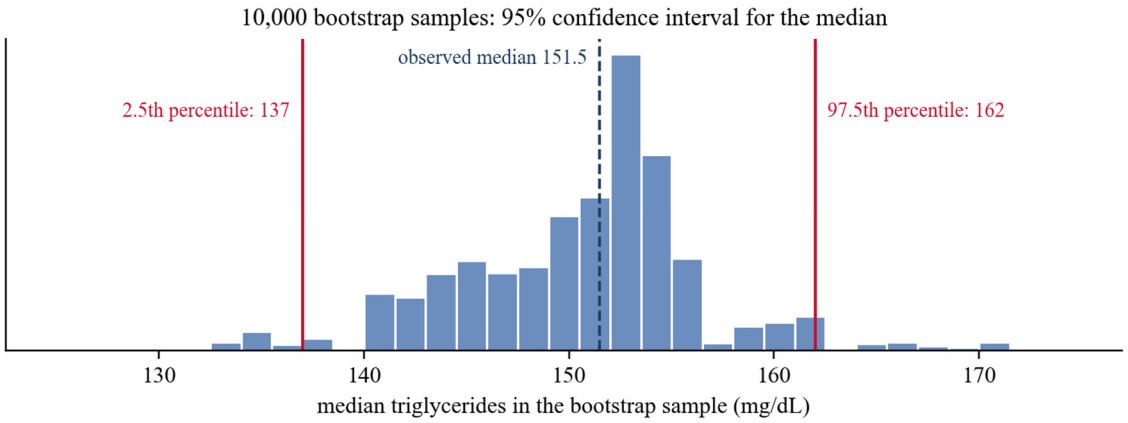


Figure 4.4. Distribution of the medians of triglycerides in 10,000 bootstrap samples drawn from the mg2026 dataset. The red lines mark the 2.5th percentile (137 mg/dL) and the 97.5th percentile (162 mg/dL), which form the 95% bootstrap confidence interval for the median; the dashed line is the observed median (151.5 mg/dL). Simulation on the mg2026 dataset.

The bootstrap is not a magic solution: it cannot correct a sample that is unrepresentative or too small (with 10 values, the bootstrap samples endlessly repeat the same 10 numbers). It is, however, used more and more often in medical articles, for example for AUC intervals or for mediation models. Dicu and colleagues used 5000 bootstrap resamples in validating a consumer assessment instrument in 504 participants [73]. JASP offers a "Bootstrap" option in many analyses, and MedCalc uses it for some intervals.

4.12 The Confidence Interval and the p Value

Chapter 5 is devoted to hypothesis testing, but the link with confidence intervals deserves to be announced now, because it greatly simplifies reading articles. For a comparison (difference of means, difference of proportions, OR, RR), there is a "null value" that means "no effect": 0 for differences, 1 for ratios. The rule is:

- if the 95% CI does not contain the null value, the corresponding two-sided test gives $p < 0.05$;
- if the 95% CI contains the null value, $p \geq 0.05$;
- if the null value lies exactly at one limit of the interval, $p \approx 0.05$.

In mg2026: the SBP difference men-women, 95% CI from -4.9 to 3.8 mmHg, contains 0, and $p = 0.81$. The RR of weight loss C versus A, 95% CI 1.11-6.17, does not contain 1, and $p < 0.05$. The obesity-diabetes OR, 95% CI 1.86-10.03, does not contain 1, and $p < 0.001$; the fact that the lower limit is far from 1 corresponds to a very small p value. The rule works exactly for methods that use the same standard error for the interval and for the test; for other combinations (for example, a Wilson interval and an exact test), small discrepancies may appear at the boundary.

The interval, however, contains more information than the p value. The p value only tells us whether the data are compatible with "no effect". The interval tells us which effect sizes are compatible with the data: if the interval is narrow and close to zero, the effect is small, even if significant; if it is wide, the result is imprecise, whatever p is [70,74].

Case study: what is missing when an odds ratio has no confidence interval

Cioca and colleagues compared the course of COVID-19 in 486 patients (200 without diabetes, 62 with type 1 diabetes and 224 with type 2 diabetes) and reported, for type 2 diabetes, an OR of 2.24 for admission to intensive care and an OR of 2.46 for mechanical ventilation [8]. In the abstract, these values are not accompanied by confidence intervals or p values. What can we not know without them? Whether the OR of 2.24 comes from an interval

of 1.9-2.6 or of 0.8-6.3; in the first case, the association would be certain and precisely estimated, in the second, it would also be compatible with no association at all. With 62 patients in the type 1 diabetes group and a probably small number of intensive care admissions, the interval is very likely wide. The critical reader does not reject the result, but treats it as an estimate of unknown precision and looks for the intervals in the full text. The opposite situation, a suspiciously narrow interval, appears in the study by Albai and colleagues on the ISARIC-4C score in 159 patients with type 2 diabetes and COVID-19, with a mortality of 25.15% [75]. The sensitivity of 87.80% is reported with a 95% CI of 87.66-87.93%, that is, a width of less than 0.3 percentage points. For a proportion estimated on about 40 deceased patients (25.15% of 159), the standard error would be around $\sqrt{\frac{0.878\times0.122}{40}} = 0.052$, hence a Wilson interval of about ±10 percentage points (around 74-95%). Intervals this narrow probably stem from a calculation or reporting error, and they illustrate why it is worth checking, at least roughly, whether the width of an interval matches the number of subjects.

4.13 Principles of Sample Size Calculation

4.13.1 Why It Is Calculated in Advance

"How many patients do I need?" is probably the question a statistician hears most often from clinicians. The answer cannot be "as many as possible", nor "as many as we manage to collect before the thesis defense". A study that is too small produces intervals so wide that it does not answer the question, and it exposes patients to procedures without scientific benefit. An unnecessarily large study consumes resources and sometimes exposes more patients than necessary to an inferior treatment. The sample size calculation is done before data collection, written into the protocol, and required by ethics committees, funders and journals [18,53].

All the formulas below follow the same logic: we require a certain precision (an acceptable margin of error) or the ability to detect a certain difference, and we start from an estimate of variability, taken from previous studies or from a pilot study. If the variability is underestimated, the calculated sample will be too small.

4.13.2 For Estimating a Mean

If we want the margin of error of the 95% CI (half its width) to be at most E, we start from $E = \frac{z\times\sigma}{\sqrt{n}}$ and solve it for n:

$$n = \left(\frac{z \times \sigma}{E}\right)^2$$

Example. We want to estimate mean SBP in a population of patients at cardiometabolic risk with a margin of ± 3 mmHg(CI 95%). From previous studies, the standard deviation is around 14 mmHg.

$n = \left(\frac{1.96\times14}{3}\right)^2 = (9.147)^2 = 83.7$, rounded up to 84 patients.

Rounding is always done upward; otherwise we no longer obtain the required precision. Notice how sensitive the formula is: for a margin of ±2 mmHg, $n = \left(\frac{1.96\times14}{2}\right)^2 = 188.2$, so 189 patients; for ±5 mmHg, only 31. Halving the margin quadruples the sample. For fasting plasma glucose, with $\sigma \approx 15.76$ mg/dL and a margin of ±3 mg/dL, 107 patients would be needed.

4.13.3 For Estimating a Proportion

The analogous formula for a proportion is:

$$n = \frac{z^2 \times p_i \times (1 - p_i)}{E^2}$$

where p_i is the anticipated proportion and E is the desired margin (as a fraction, not a percentage).

Example. A physician wants to estimate the prevalence of undiagnosed diabetes among the at-risk patients on his or her list. The physician anticipates a prevalence of around 20% and wants a margin of ±5 percentage points.

$n = \frac{1.96^2 \times 0.20 \times 0.80}{0.05^2} = \frac{3.8416 \times 0.16}{0.0025} = 245.9$, so 246 patients.

If we have no idea of the prevalence, we use $p_i = 0.5$, which gives the maximum value of the product p_i x (1 - p_i) and therefore the largest, "safe" sample: $n = \frac{3.8416 \times 0.25}{0.0025} = 384.2$, so 385. For $p_i = 0.18$ and a margin of ±3 percentage points, n = 631. The formula is based on the Wald approximation; for very small proportions (below 5%), more exact variants, available in software, are used.

4.13.4 For Comparing Two Means: Power and Effect Size

When the aim is to compare two groups (for example, a new treatment versus the standard one), the calculation requires four ingredients:

- the significance level α, usually 0.05 two-sided (z = 1.96);
- the desired power, 1 - β, usually 80% (z = 0.84) or 90% (z = 1.28): the probability that the study will detect the difference if it truly exists;
- the minimum clinically relevant difference Δ, which we do not want to miss;
- the standard deviation σ of the variable.

The last two are combined into the standardized effect size, Cohen's d = Δ / σ [76]. By convention, d = 0.2 is a small effect, 0.5 a medium one and 0.8 a large one. The approximate formula for the number of patients in each group is:

n per group = 2 x $(z(1-\alpha/2) + z(1-\beta))^2 / d^2$

For d = 0.5, α = 0.05 and 80% power: $n = \frac{2 \times (1.96 + 0.84)^2}{0.25} = \frac{2 \times 7.85}{0.25} = 62.8$. Software, which uses the t distribution instead of the normal approximation, gives 64 patients per group. For d = 0.3, the same calculation gives 174.4, and software gives 176 per group; for d = 0.8, 26 per group; for d = 0.2, 394 per group. Small effects require large samples: to detect an effect half as large, we need four times as many patients. If we require 90% power instead of 80%, for d = 0.5 the number rises from 64 to 86 per group.

A clinical example. We want to know whether an intervention lowers SBP by at least 5 mmHg compared with the control group, with σ = 14 mmHg. $d = \frac{5}{14} = 0.36$, and software indicates 125 patients per group for 80% power. Let us now go back to the comparison of SBP between the sexes in mg2026, with 74 men and 86 women: the power of this sample to detect a true difference of 5 mmHg is only about 61%. The fact that the observed difference was not significant (p = 0.81) therefore does not rule out a difference of 5 mmHg; the confidence interval (-4.9 to 3.8 mmHg) says the same thing, more directly.

For two proportions, the formula is longer, but the logic remains the same. To detect an increase in the proportion of patients with weight loss ≥ 5% from 11.3% (arm A) to 29.6% (arm C), with α = 0.05 and 80% power, about 76 patients per arm would be needed. The actual arms in mg2026 have 53-54 patients each, and their power for this difference is about 67%. The study

nevertheless found a significant difference (95% CI for the difference 3.4-33.2 percentage points), but with a wide interval, exactly what we expect from a study with moderate power.

4.13.5 Practical Adjustments: Losses during the Study and Small Populations

Losses during the study (patients who drop out, who can no longer be contacted, who have missing data) reduce the analyzed sample. If we anticipate a loss rate r, we recruit n / (1 - r) patients. For 64 patients per group and losses estimated at 20%: 64 / 0.8 = 80 per group, that is, 160 in total. The common mistake is to multiply by 1.2 (64 x 1.2 = 76.8, so 77), which underestimates the requirement: of 77 recruited, with 20% losses only 61.6 remain.

The finite population correction is applied when the sample represents a substantial fraction (over 5-10%) of a small, known population, for example all at-risk patients on a family physician's list:

$$n \text{ corrected } = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$$

For n0 = 246 and a list of N = 1000 at-risk patients, $n = \frac{246}{1+\frac{245}{1000}} = \frac{246}{1.245} = 197.6$, so 198. For N = 500, 166 patients. For large populations (tens of thousands), the correction is negligible.

4.13.6 Software for Sample Size Calculation

In practice, nobody does these calculations by hand, but it is important that you understand what you enter into the software. G*Power is a free, widely used program that covers t-tests, ANOVA, chi-square, correlation and regression [77]. For two means, you choose the "t tests" family, the test "Means: Difference between two independent means (two groups)", the "A priori" type of analysis, and you enter the effect size d, α, the power and the allocation ratio between groups; for d = 0.5 the result is 64 per group. MedCalc has a dedicated menu (Sampling) for a mean, a proportion, the comparison of two means or two proportions, ROC curves and more [50]. Epi Info offers, in the StatCalc module, the calculators "Population survey" (for prevalences, with the finite population correction built in), "Cohort or cross-sectional" and "Unmatched case-control" [51]. The calculators are used in Practical Session 4.

4.13.7 What to Write in the Protocol

A correct paragraph on sample size allows anyone to redo the calculation exactly. It must contain: the primary variable on which the calculation was based; the minimum clinically relevant difference or the desired precision margin, with justification; the estimate of variability (the standard deviation or the anticipated proportion) and its source; α and whether the test is two-sided; the power; the anticipated loss rate; the final number of patients to be recruited; the software used. For example: "To detect a difference of 5 mmHg in SBP between groups (estimated standard deviation 14 mmHg, based on study X), with a two-sided t-test, α = 0.05 and 80% power, 125 patients per group are required (calculation in G*Power). Anticipating 15% losses, we will recruit 148 patients per group."

Calculating power after the fact ("post hoc" power, computed from the observed effect) is of limited value: it is a direct function of the p value and brings no new information. For a study that has already been completed, the confidence interval shows more clearly which effects are or are not compatible with the data.

4.13.8 Consequences of Samples That Are Too Small

A study with low power has three problems. The first is obvious: it may miss a real effect (type II error, Chapter 5). The second is less obvious: if it does find a significant effect, the estimate tends

to be exaggerated, because only large estimates, arising by chance, cross the significance threshold. The third: the confidence intervals are so wide that the result, significant or not, is hard to use clinically.

> **Case study: a study with 29.3% power**
>
> Popovici and colleagues compared, in a prospective cohort of 100 patients undergoing ERCP, sedation guided by entropy monitoring with conventional sedation [78]. Hypertensive episodes were less frequent in the entropy-guided group (6% versus 26%, p = 0.007), and Aldrete recovery scores were higher at 5 minutes (p = 0.003) and at 15 minutes (p < 0.001). The authors, however, honestly point out a limited statistical power of 29.3%. A power of 29.3% means that, if the difference being sought really existed, a study of this size would have only about 3 chances in 10 of detecting it. A nonsignificant result obtained under these conditions cannot be read as "there is no difference"; it is simply inconclusive. The observational, nonrandomized design adds a second limitation: the groups may also differ in characteristics other than the monitoring method. The correct conclusion, which the title also expresses ("is associated with"), is that the results generate hypotheses that must be confirmed in a randomized trial, with a sample size calculated in advance for the primary variable.

4.14 Worked Problems

Worked Problem 4.1. 95% CI and 99% CI for Mean Glucose

Statement. In mg2026, baseline fasting plasma glucose has n = 160, $m = 105.44$ mg/dL and $s = 15.76$ mg/dL. Calculate by hand the 95% CI and the 99% CI for the population mean and interpret them.

Solution. The glucose distribution has a slight skewness (0.37), but for n = 160 the mean is almost normally distributed (central limit theorem), so the t interval is appropriate.

$\text{SE} = \frac{15.76}{\sqrt{160}} = \frac{15.76}{12.649} = 1.246$ mg/dL.

95% CI: t(159; 0.975) = 1.975; margin = 1.975 x 1.246 = 2.461; the interval 105.44 ± 2.46, that is, 102.98-107.90 mg/dL.

99% CI: t(159; 0.995) = 2.607; margin = 2.607 x 1.246 = 3.248; the interval 105.44 ± 3.25, that is, 102.19-108.69 mg/dL.

Check in Google Sheets: =CONFIDENCE.T(0.05, STDEV(S2:S161), 160) gives 2.461, and =CONFIDENCE.T(0.01, STDEV(S2:S161), 160) gives 3.248.

Statistical interpretation. With 95% confidence, the mean fasting plasma glucose in the population of at-risk adults from which the sample comes lies between 102.98 and 107.90 mg/dL; the 99% interval is wider, because we demand greater certainty. Clinical interpretation. The entire interval lies within the impaired fasting glucose range (100-125 mg/dL), which confirms that the population studied has, on average, impaired glucose metabolism.

Worked Problem 4.2. Values or Mean? The Patients' Age

Statement. The age of the patients in mg2026 has m = 56.29 years, s = 8.77 years, n = 160, and a normal distribution (Shapiro-Wilk p = 0.687). (a) Estimate, with 95.4% confidence, the interval in which the patients' ages lie. (b) Estimate, with 95.4% confidence, the mean age in the population. (c) Comment on the difference.

Solution. (a) For individual values we use the standard deviation: $m \pm 2s = 56.29 \pm 17.54$, that is, 38.75-73.83 years. In the sample, 94.4% of patients have an age within this interval.

(b) For the mean we use the standard error: SE $= \frac{8.77}{\sqrt{160}} = 0.693$ years; $m \pm 2\text{SE} = 56.29 \pm 1.39$, that is, 54.90-57.68 years.

(c) The first interval is 35 years wide and describes the diversity of the patients: from young adults to older people. The second is less than 3 years wide and describes how precisely we know the mean age of the population. The ratio between the widths is √160 ≈ 12.6. In the exam, the key words in the statement are "the values" (SD) versus "the mean" (SE).

Worked Problem 4.3. A Small Sample

Statement. In 8 patients with type 2 diabetes, HbA1c after three months of treatment has the following values (fictitious data): 6.1, 5.8, 6.4, 7.0, 6.2, 5.9, 6.6, 6.0%. Calculate the 95% CI for mean HbA1c.

Solution. With only 8 values we cannot seriously check normality; HbA1c usually has an approximately symmetric distribution in homogeneous groups, and the values show no obvious extreme value, so we use the t interval, with caution.

The mean: sum = 50.0, so $m = \frac{50.0}{8} = 6.25\%$.

The deviations from the mean: -0.15; -0.45; 0.15; 0.75; -0.05; -0.35; 0.35; -0.25. The squares: 0.0225; 0.2025; 0.0225; 0.5625; 0.0025; 0.1225; 0.1225; 0.0625; sum = 1.12.

$s^2 = \frac{1.12}{7} = 0.16$; s = 0.40%.

$\text{SE} = \frac{0.40}{\sqrt{8}} = \frac{0.40}{2.828} = 0.141\%$.

df = 7; t(7; 0.975) = 2.365; margin = 2.365 x 0.141 = 0.334.

95% CI: 6.25 ± 0.33, that is, 5.92-6.58%.

Interpretation. The mean of the population of similarly treated patients lies, with 95% confidence, between 5.92% and 6.58%. Had we wrongly used z = 1.96, the margin would have been 0.28, and the interval 5.97-6.53%, about 17% narrower than the correct one. With 8 patients, the difference between t and z is not a subtlety.

Worked Problem 4.4. CI for a Proportion: Wald, Wilson and the Rule of Three

Statement. (a) In an audit, 12 of 80 hypertensive patients have uncontrolled BP on treatment. Calculate the 95% CI for the proportion by the Wald method and by the Wilson method. (b) In another practice, no case of angioedema was observed in 40 patients treated with a new antihypertensive drug. What can you say about the true risk?

Solution. $(a)\ p = \frac{12}{80} = 0.15$. SE $= \sqrt{\frac{0.15 \times 0.85}{80}} = \sqrt{0.001594} = 0.0399$.

Wald: $0.15 \pm 1.96 \times 0.0399 = 0.15 \pm 0.078$, that is, 7.2%-22.8%.

Wilson: $\frac{z^2}{n} = \frac{3.8416}{80} = 0.0480$; $\frac{z^2}{2n} = 0.0240$. Center = (0.15 + 0.024) / 1.048 = 0.1660. Under the square root: $0.001594 + \frac{3.8416}{4 \times 6400} = 0.001594 + 0.000150 = 0.001744$; square root = 0.0418. Margin = $\left(\frac{1.96}{1.048}\right) \times 0.0418 = 1.870 \times 0.0418 = 0.0781$. Wilson 95% CI: 0.166 ± 0.078, that is, 8.8%-24.4%.

The Wilson interval is shifted toward 50% and asymmetric around 15% (6.2 points below, 9.4 above), as it should be for a small proportion in a moderate-sized sample.

(b) Zero events out of 40: the rule of three gives an upper limit of the 95% CI of approximately 3 / 40 = 7.5%. The data are compatible with a true risk of angioedema of up to about 1 case in 13

patients. The absence of cases in 40 patients does not prove the drug is safe with respect to a rare reaction.

Worked Problem 4.5. The Difference between Two Proportions and the Prevalence Ratio

Statement. In mg2026, at enrollment, diabetes was present in 18 of 54 obese patients (33.3%) and in 11 of 106 non-obese patients (10.4%). Calculate the 95% CI for the difference in prevalences and for the prevalence ratio.

Solution. The difference: 0.333 - 0.104 = 0.230 (23.0 percentage points).

$$\text{SE} = \sqrt{\frac{0.333\times0.667}{54} + \frac{0.104\times0.896}{106}} = \sqrt{0.004115 + 0.000878} = \sqrt{0.004993} = 0.0707.$$

95% CI: $0.230 \pm 1.96 \times 0.0707 = 0.230 \pm 0.139$, that is, 9.1-36.8 percentage points.

The prevalence ratio: $\text{PR} = \frac{0.333}{0.104} = 3.21$. On the logarithmic scale: ln(3.21) = 1.167; $\text{SE}(\ln \text{PR}) = \sqrt{\frac{1}{18} - \frac{1}{54} + \frac{1}{11} - \frac{1}{106}} = \sqrt{0.0556 - 0.0185 + 0.0909 - 0.0094} = \sqrt{0.1185} = 0.344$. $\ln(\text{PR}) \pm 1.96 \times 0.344 = 1.167 \pm 0.675$, that is, 0.492 to 1.842; back-transformed: $e^{0.492} = 1.64$ and $e^{1.842} = 6.31$.

95% CI for the PR: 1.64-6.31.

Interpretation. The prevalence of diabetes is about 3 times higher in obese patients, with an interval compatible with a ratio from 1.6 to 6.3; in absolute terms, the excess is 23 percentage points (9-37). Both intervals exclude the null value (0 and 1, respectively), so the association is statistically significant. Being a cross-sectional analysis, it does not establish causality (Chapter 10).

Worked Problem 4.6. Sample Size for a Study in a Family Practice

Statement. (a) A family physician wants to estimate the prevalence of uncontrolled hypertension among the 1500 hypertensive patients on the list, anticipating 30%, with a margin of ±5 percentage points. The physician estimates that 10% of the invited patients will not respond. (b) A second physician wants to compare the reduction in glucose between a new intervention and standard counseling: the minimum relevant difference is 8 mg/dL, and the standard deviation of the reduction, from mg2026, is around 11 mg/dL; α = 0.05, 80% power, estimated losses 15%.

Solution. (a) Step 1: $n_0 = \frac{1.96^2\times0.30\times0.70}{0.05^2} = \frac{3.8416\times0.21}{0.0025} = 322.7$, so 323. Step 2, the finite population correction: $n = \frac{323}{1+\frac{322}{1500}} = \frac{323}{1.215} = 265.8$, so 266. Step 3, nonresponse: 266 / 0.90 = 295.6, so 296 patients invited.

(b) $d = \frac{8}{11} = 0.73$. n per group = $\frac{2\times(1.96+0.84)^2}{0.73^2} = \frac{15.7}{0.529} = 29.7$, so 30 (G*Power, using the t distribution, gives 31). With 15% losses: 31 / 0.85 = 36.5, so 37 patients per group, 74 in total.

Interpretation. The finite population correction reduced the sample by almost 60 patients, because the physician's list is small. For the second study, the anticipated effect is large (d = 0.73), so the sample is modest; if the relevant difference were only 4 mg/dL($d = 0.36$), about 120 patients per group would be needed, four times as many.

4.15 Proposed Exercises

1. Total cholesterol in mg2026 has n = 160, $m = 209.39$ mg/dL, $s = 36.91$ mg/dL. Calculate the standard error and the 95% CI for the population mean (t = 1.975).

2. HDL in mg2026 has $m = 49.52$ mg/dL, $s = 11.78$ mg/dL, n = 160. Calculate the 68.3%, 95.4% and 99.7% intervals for individual values and for the mean, and present them in a table like Table 4.1.
3. In a sample of 25 patients with diabetes, HbA1c has a mean of 7.2% and a standard deviation of 1.5%. Calculate the 95% CI for the mean. Which critical value did you use, and why?
4. In mg2026, 41 of 160 patients have SBP ≥ 140 mmHg. Calculate the 95% CI for the proportion by the Wald method and by the Wilson method.
5. A new vaccine was given to 120 volunteers, without any anaphylactic reaction. What is the approximate upper limit of the 95% CI for the risk of anaphylaxis?
6. Two independent groups: n1 = 40, m1 = 142 mmHg, s1 = 15 mmHg; n2 = 45, m2 = 134 mmHg, s2 = 14 mmHg. Calculate the difference between the means, its standard error and the approximate 95% CI (with z = 1.96). Is the difference significant?
7. In a case-control study, 20 of 50 cases and 10 of 50 controls were exposed. Calculate the OR and the 95% CI on the logarithmic scale.
8. How many patients are needed to estimate: (a) a mean with σ = 10 and a margin of ±2 units (95% CI); (b) a completely unknown proportion with a margin of ±4 percentage points?
9. For comparing two means with d = 0.4, α = 0.05 two-sided and 80% power, calculate the number of patients per group with the approximate formula. How many should be recruited if 20% losses are anticipated?
10. An article reports "SBP 131 ± 1.1 mmHg$(n = 160)$". What does the value 1.1 most likely represent? Calculate the corresponding standard deviation. Another article reports OR = 1.75 (95% CI 0.95-3.10): is the association significant at the 0.05 level?
11. A 95% CI for a mean, calculated on 40 patients, has a width of 10 mmHg. Approximately how many patients would be needed for a width of 5 mmHg, assuming the same standard deviation?

Common Pitfalls

- Confusing the standard deviation with the standard error: the SD describes the patients, the SE describes the precision of the mean. "Mean ± SE" makes the data look less variable.
- Using the interval for the mean when the question is about individual values (and vice versa), especially in problems of the type "estimate the values/the mean with 95.4% confidence".
- Interpreting the 95% CI as "95% of patients lie within the interval" or as "a 95% probability that the mean lies in this particular interval".
- Using z = 1.96 instead of the t value for small samples.
- Reporting the Wald interval for small proportions or small samples, with negative limits or insufficient coverage.
- Concluding "there is no effect" on the basis of a wide interval that includes the null value; the correct statement is "the data do not allow a conclusion".
- Calculating the interval for an OR or RR directly on the original scale, symmetrically, instead of on the logarithmic scale.
- Rounding the sample size down, or adjusting for losses by multiplying by (1 + r) instead of dividing by (1 - r).
- Calculating "post hoc" power from the observed effect in order to "explain" a nonsignificant result.

- Ignoring the fact that the confidence interval measures only random error; a biased sample gives narrow intervals around a wrong value.

Key Points

- The standard error of the mean is SE = $\frac{s}{\sqrt{n}}$; for SBP in mg2026, 1.097 mmHg, compared with an SD of 13.88 mmHg.
- The interval for individual values is $m \pm k \times s$, the interval for the mean is m ± k x SE (k = 1, 2, 3 for 68.3%, 95.4%, 99.7%).
- CI for the mean: $m \pm \frac{t \times s}{\sqrt{n}}$, with df = n - 1; for SBP, 95% CI 128.96-133.30 mmHg and 99% CI 128.27-133.99 mmHg.
- A 95% CI means that the calculation procedure captures the true parameter in 95% of repeated studies; its width decreases with $\sqrt{n}$.
- For proportions, the Wilson interval is preferred (diabetes in mg2026: 12.9-24.8%, compared with Wald 12.2-24.1%); for zero events out of n, the upper limit is approximately 3 / n.
- Intervals for the OR, RR and geometric mean are calculated on the logarithmic scale and are asymmetric; the point estimate is the geometric mean of the limits.
- If the 95% CI of a difference excludes 0 (or that of a ratio excludes 1), p < 0.05; the interval additionally shows the size and precision of the effect.
- Sample size: $n = \left(\frac{z \times \sigma}{E}\right)^2$ for a mean (84 for ±3 mmHg, σ 14), $n = \frac{z^2 \times p_i \times (1-p_i)}{E^2}$ for a proportion (246 for 20% ± 5%), approximately 2 x (1.96 + 0.84)² / d² per group for two means (64 for d = 0.5; 176 for d = 0.3).

Self-Assessment Questions

1. (SC) In mg2026, SBP has s = 13.88 mmHg and n = 160. The standard error of the mean is approximately:
    a) 0.087 mmHg
    b) 1.097 mmHg
    c) 3.72 mmHg
    d) 13.88 mmHg
    e) 27.2 mmHg
2. (SC) To halve the width of a confidence interval for the mean, keeping the same variability and the same confidence level, the sample must be:
    a) doubled
    b) tripled
    c) increased fourfold
    d) increased eightfold
    e) halved
3. (SC) A study reports OR = 1.40 (95% CI 0.90-2.18). The correct statement is:
    a) the association is statistically significant at the 0.05 level
    b) the interval includes the value 1, so p ≥ 0.05
    c) the interval proves the absence of any association
    d) the interval should be symmetric around 1.40
    e) the p value is certainly less than 0.01
4. (SC) In a study of 60 patients, no serious adverse event was observed. The approximate upper limit of the 95% CI for the true risk is:

a) 0%
b) 1,7%
c) 3%
d) 5%
e) 10%

5. (SC) How many patients are needed to estimate a mean with a margin of ± 3 mmHg(CI 95%), if σ = 14 mmHg?
   a) 19
   b) 47
   c) 84
   d) 176
   e) 246
6. (MC) Regarding the interpretation of a 95% CI for mean SBP of 128.96-133.30 mmHg, which statements are correct?
   a) if the study were repeated many times, about 95% of the intervals calculated in this way would contain the true mean
   b) 95% of patients have an SBP between 128.96 and 133.30 mmHg
   c) the interval expresses the precision of the estimate of the mean
   d) a 99% CI calculated on the same data would be wider
   e) the interval would become narrower if the sample were larger
7. (MC) Which statements about the standard deviation (SD) and the standard error (SE) are correct?
   a) the SD describes the variability of individual values
   b) the SE decreases as the sample size increases
   c) $\mathrm{SE} = \frac{\mathrm{SD}}{\sqrt{n}}$
   d) the reference interval for individual values is constructed with the SE
   e) reporting "mean ± SE" instead of "mean ± SD" makes the data look less variable
8. (MC) Which statements about the confidence interval for a proportion are correct?
   a) the Wald method can give limits below 0% for small proportions
   b) the Wilson method gives asymmetric intervals for proportions close to 0
   c) the Wald method has insufficient coverage in small samples
   d) the Wilson interval for diabetes in mg2026 is 12.9-24.8%
   e) the Wald interval is always wider than the Wilson interval
9. (MC) Which of the following factors increase the required sample size for comparing two means?
   a) a smaller minimum relevant difference
   b) a larger standard deviation
   c) a higher desired power (90% instead of 80%)
   d) a higher α level (0.10 instead of 0.05)
   e) a higher anticipated loss rate
10. (MC) Which statements about confidence intervals for ratios (OR, RR) are correct?

a) they are calculated on the logarithmic scale and back-transformed
b) they are usually asymmetric around the point estimate
c) the point estimate is approximately the geometric mean of the limits

d) an interval that excludes 1 corresponds to a result that is significant at the 5% level

e) a very wide interval indicates a very large effect

Chapter 5. Foundations of Statistical Testing: Hypotheses, Significance Level, Type I and II Errors, Decision and Interpretation

Objectives

By the end of this chapter, the student will be able to:

- turn a clinical question into a research hypothesis and then into a pair of statistical hypotheses, H0 and H1, written in words and mathematically;
- choose between a two-sided and a one-sided test and justify the choice before seeing the data;
- define the p value correctly and recognize its most common misinterpretations;
- apply the decision rule for a given α level and state the statistical conclusion without the phrase "H0 is accepted";
- describe type I and II errors, explain the relationship between α, β, statistical power 1 - β and the confidence level 1 - α, and list the factors that influence power;
- calculate and interpret an effect size (Cohen's d, dz, r, eta², OR) and distinguish statistical significance from clinical importance;
- solve by hand a z-test and a t-test for a single mean, using three equivalent routes: the critical value, the p value and the confidence interval;
- explain the inflation of the type I error with multiple comparisons and apply the Bonferroni, Holm and Benjamini-Hochberg corrections;
- report a statistical test according to current standards (the exact p value, the effect size, the confidence interval) and recognize practices such as p-hacking and HARKing.

5.1 From the Clinical Question to a Testable Hypothesis

The previous chapters described the data (descriptive statistics), showed how a normal variable behaves and how a confidence interval for a mean or a proportion is constructed. Here begins the second part of the course, inferential statistics in the strict sense: we make decisions about a population on the basis of a single sample. All the tests in Chapters 6-9 (the t-test, ANOVA, Mann-Whitney, chi-square, the test for correlation) use the same logic, which we now build once, carefully. Once the logic is understood, each new test means only a new formula for the "test statistic"; the rest of the reasoning stays the same [19,63].

5.1.1 The Research Question and the Research Hypothesis

Every study starts from a clinical question. The family physician who followed the patients in the mg2026 study for six months wonders whether the mobile app, added to counseling, lowers fasting plasma glucose more than counseling alone. The question is a good one, but still vague: what does "more" mean, in whom, after how long, measured how? The researcher's first task is to specify the population, the intervention, the comparator and the outcome (the PICO framework, discussed in the medical informatics volume [1]). The result is a research hypothesis, a statement about the real world that the study tries to support: "in adults at cardiometabolic risk in family practices, adding the self-monitoring app to counseling produces, over six months, a greater reduction in fasting plasma glucose than counseling alone".

The research hypothesis is formulated in the clinician's language and usually expresses what the researcher expects to find: a difference, an association, an effect. It cannot be tested directly,

because we do not have access to all patients at cardiometabolic risk in the country. We have 160 patients. To move from these 160 to the population, the research hypothesis must be translated into a form that a calculation can check: a pair of statistical hypotheses about population parameters (means μ, proportions π, correlation coefficients ρ).

5.1.2 Statistical Hypotheses: H0 and H1

The null hypothesis (H0, the hypothesis of zero) states that in the population there is no difference, association or effect of the kind being sought. The difference observed in the sample would then be merely the result of sampling variability, that is, of the chance that decides which patients end up in the study. The alternative hypothesis (H1, sometimes denoted Ha) negates the null hypothesis: in the population, the difference exists. The two hypotheses are mutually exclusive and together cover all possibilities.

Why do we start from H0, that is, from the hypothesis that, more often than not, we hope to refute? Because H0 is the only one that fixes a precise value of the parameter (for example, a difference between means equal to zero). Only by starting from a precise value can we calculate how probable our data would be. The alternative hypothesis "the difference is not zero" covers an infinity of possible values and does not allow a single calculation. The reasoning resembles a proof by contradiction: we assume that H0 is true, calculate what we should see in that case, and, if the real data are very hard to reconcile with the assumption, we abandon it.

For the app example, the hypotheses are written as follows:

H0: the mean reduction in glucose at 6 months is the same in the population treated with counseling + app and in the population treated with counseling alone; mathematically, μB = μA.

H1: the mean reduction in glucose differs between the two populations; mathematically, μB ≠ μA.

Notice that the hypotheses refer to the population (Greek letters μ, π, ρ), not to the sample. We already know the sample means (3.0 mg/dL in arm A and 11.1 mg/dL in arm B); there is no point in making hypotheses about them. In the exam, writing the hypotheses with m1 and m2 (the sample means) instead of μ1 and μ2 is a common mistake and loses points.

Table 5.1. Examples of translating clinical questions into statistical hypotheses (mg2026 dataset and clinical examples).

Clinical question	H0 (in words)	H0 (mathematically)	H1 (mathematically)
Does the patients' mean SBP differ from 125 mmHg?	the population mean is 125 mmHg	$\mu = 125$	μ ≠ 125
Does SBP differ between women and men?	the means of the two sexes are equal	$\mu_F = \mu_M$	μF ≠ μM
Does glucose change between enrollment and 6 months (same patients)?	the mean of the individual differences is zero	$\mu_D = 0$	μD ≠ 0
Does the reduction in glucose differ among the three arms?	the three means are equal	$\mu_A = \mu_B = \mu_C$	at least one mean differs
Does the proportion of patients with weight loss ≥5% differ between arms C and A?	the proportions are equal	$\pi_C = \pi_A$	πC ≠ πA
Are BMI and SBP correlated?	the population correlation coefficient is zero	$\rho = 0$	ρ ≠ 0

5.1.3 Two-Sided and One-Sided Tests

The alternative hypothesis in the examples above is two-sided: it allows a difference in either direction, μB > μA or μB < μA. A one-sided test looks at a single direction: H1: μB > μA. For the same dataset, the one-sided p value is half the two-sided p value, if the observed difference goes in the stated direction. This is precisely where the temptation lies: a result with a two-sided p of 0.08 becomes "significant" (p = 0.04) if we decide, after seeing the data, that the test "should have been" one-sided.

A one-sided test is rarely justified, and only if three conditions are met at the same time. First, the direction is fixed before data collection, in the protocol. Second, a difference in the opposite direction would be biologically impossible or, if it did occur, would have the same practical consequence as no difference. Third, the reader is explicitly informed. The classic example is the non-inferiority trial, in which the only question is whether the new treatment is not worse by more than an accepted margin (Section 5.12). In ordinary studies in general medicine, a new drug can also worsen the patient's course; that is why the default test, required by most journals and used in this course, is the two-sided one [42,53]. Software (JASP, MedCalc, Google Sheets) offers the two-sided version by default; if you choose otherwise, you must be able to say why.

5.2 The Logic of the Test: The Test Statistic and Its Distribution under H0

Suppose we want to know whether the mean SBP in the population from which the 160 patients come differs from 125 mmHg, a reference value chosen for the example. The observed mean is 131.13 mmHg. The difference of 6.13 mmHg could be real, but it could also be the product of chance: if the population mean were indeed 125, a sample of 160 patients would not have a mean of exactly 125, but a nearby value, higher or lower. The right question becomes: how unusual is a deviation of 6.13 mmHg if H0 were true?

To answer, we compare the observed difference (the "signal") with the variability we would expect from chance alone (the "noise"). The measure of noise for a mean is the standard error of the mean, $\mathrm{SE} = \frac{s}{\sqrt{n}}$, which you met in Chapter 4. The ratio of signal to noise is called the test statistic:

$$t = \frac{m - \mu_0}{\mathrm{SE}} = \frac{m - \mu_0}{\frac{s}{\sqrt{n}}}$$

For SBP: $\mathrm{SE} = \frac{13.88}{\sqrt{160}} = \frac{13.88}{12.649} = 1.097$ mmHg, so $t = \frac{131.13 - 125}{1.097} = \frac{6.13}{1.097} = 5.59$. The observed mean lies almost 5.6 standard errors away from the value stated by H0.

The test statistic has one property that is decisive for the reasoning: we know how it is distributed when H0 is true. If we repeated the study a very large number of times, with samples of 160 patients from a population with a mean of exactly 125, the t values obtained would follow a Student distribution with n - 1 = 159 degrees of freedom, very similar to the standard normal distribution: centered at zero, symmetric, with about 95% of the values between -1.975 and +1.975. A value of t = 5.59 lies far outside this zone. Under H0, it would be extraordinarily rare.

Each test has its own statistic (z, t, F, χ^2, U, H), but the idea is the same: a number that grows as the data move away from H0 and whose distribution under H0 is known, from theory or from tables. The values of the test statistic so extreme that, under H0, they occur with a probability lower than a chosen threshold (usually 5%) form the critical region, or rejection region. Its boundary is called the critical value.

5.3 The p Value

5.3.1 The Correct Definition

The p value is the probability of obtaining, by the mere play of chance, a test statistic at least as extreme as the one observed, assuming that the null hypothesis is true and that all the conditions for applying the test are met. Every word in the definition matters. "At least as extreme" means that we add up the probability of the observed value and of all values farther from H0; in a two-sided test, we add up both tails of the distribution. "Assuming that H0 is true" means that p is calculated in a hypothetical world in which H0 is true; the p value tells us nothing about the probability that this world is the real one.

For SBP, the two-sided p value associated with t = 5.59 and 159 degrees of freedom is approximately 0.0000001 (in Google Sheets, =T.DIST.2T(5.59, 159)). In other words, if the population mean were 125 mmHg, only about one in ten million samples of 160 patients would have a mean at least as far from 125 as ours. Such a value is reported as "p < 0.001", since more decimal places no longer add useful information.

5.3.2 What the p Value Is Not

The p value is probably the most frequently misinterpreted number in the medical literature. In 2016, the American Statistical Association published an official statement, with six principles, precisely to put a stop to these confusions [74]. A landmark article by Greenland and colleagues lists more than twenty misinterpretations commonly found in published articles [70]. We focus on those that also come up in the final degree exam and in the residency examination.

The p value is not the probability that H0 is true. A p of 0.03 does not mean "a 3% chance that the treatment has no effect". The probability of the hypothesis, given the evidence, requires a different kind of calculation (Bayesian reasoning, which you will meet with diagnostic tests in Chapter 11), which also takes into account how plausible the hypothesis was before the study.

The p value is not the probability that the result is "due to chance". The wording found in many older textbooks ("p is the probability that the observed differences arose by chance") is a convenient shorthand, but a misleading one. The correct version is: p is the probability of data this extreme if chance were the only mechanism at work.

The p value does not measure the size of the effect or its clinical importance. A p of 0.0001 can accompany a difference of 0.04 diopters, of no relevance whatsoever to the patient, if the sample is large; a p of 0.20 can accompany a one-third reduction in mortality, if the sample is small. We will return to this with examples in Section 5.6.

A p > 0.05 does not prove that H0 is true. The lack of a statistically significant difference does not mean that the difference is absent; it means that the data were not sufficient to demonstrate it. Section 5.4.2 is devoted to this idea.

The 0.05 threshold is not a natural boundary. The difference between p = 0.049 and p = 0.051 is negligible as evidence, although convention places them on either side of the decision. A result does not become "true" when p drops below 0.05 and "false" when it rises above it.

The p value says nothing about the quality of the design. A study with biased patient selection, with flawed measurements or with uncontrolled confounders can produce very small p values for effects that do not exist. The statistical test only checks whether the data, as they are, can be reconciled with H0; it does not check whether the data are any good.

5.3.3 The Exact p Value and How to Report It

In the past, when p values were looked up in printed tables, authors wrote only "p < 0.05" or "NS" (nonsignificant). Today software calculates the exact p, and reporting guidelines require the exact value, with two or three decimal places: p = 0.034, p = 0.81, p = 0.0008. For very small values, write p < 0.001. Never write p = 0.000, as software sometimes displays when it rounds: the probability is not zero, merely smaller than 0.0005. Reporting the exact value allows readers to judge for themselves how close to the threshold the result was and makes it possible to include the study in a meta-analysis.

5.4 The Significance Level α and the Decision Rule

5.4.1 The α Level and the Decision

The significance level α is the maximum probability of a type I error (rejecting a true H0) that the researcher is willing to accept. It is set before the analysis, in the protocol. The conventional values are 0.05, 0.01 and 0.001; in clinical medicine, α = 0.05 is the standard, and stricter thresholds appear when the consequences of a false positive are serious (for example, approving a drug) or when a very large number of tests are performed (genetic studies use thresholds of the order of 0.00000005).

The decision rule is simple:

- if p < α, we reject H0; we say that the result is statistically significant at level α;
- if p ≥ α, we do not reject H0; the result is not statistically significant at level α.

For SBP versus 125 mmHg, p < 0.001 < 0.05, so we reject H0: the mean SBP in the population of patients at cardiometabolic risk from which the sample comes differs significantly from 125 mmHg (it is higher). If we repeat the test with the reference value of 130 mmHg, we obtain $t = \frac{131.13-130}{1.097} = 1.03$ and p = 0.304. Now p > 0.05: we do not reject H0, and the data are compatible with a population mean of 130 mmHg. Same sample, same mean, two different conclusions, because we asked two different questions.

5.4.2 Why "We Do Not Reject H0" and Not "We Accept H0"

The phrase "we accept H0", still found in some teaching materials, suggests that we have demonstrated the absence of a difference. We have demonstrated nothing of the kind. A statistical test can provide evidence against H0, but it cannot provide evidence in its favor. Altman and Bland summed up the idea in a title that has become proverbial: absence of evidence is not evidence of absence [79].

An example from the mg2026 dataset makes this visible. Let us take just 10 patients at random and test again whether the mean SBP differs from 125 mmHg. The 10 patients have the values 150, 154, 131, 130, 137, 102, 110, 103, 147 and 136 mmHg, with a mean of 130.0 and a standard deviation of 19.04. The standard error is 19.04 / √10 = 6.02, and $t = \frac{130.0-125}{6.02} = 0.83$, with 9 degrees of freedom; p = 0.43. We do not reject H0. It would be absurd to "accept" that the mean is 125 mmHg, since in all 160 patients we found an extremely significant difference. The 95% confidence interval for the mean of the 10 patients, 116.38-143.62 mmHg, shows why: it includes 125, but it includes 140 just as well. The small sample is compatible with almost any plausible value. The correct conclusion is: the data do not allow H0 to be rejected; the difference, if it exists, could not be demonstrated with this sample.

In a clinician's words, a nonsignificant result can mean two very different things: an absent effect or an undetermined effect. We tell them apart by looking at the confidence interval. If the

interval is narrow and includes only clinically unimportant differences, we can say that a relevant effect is unlikely. If the interval is wide and includes both zero and important differences, the result is undetermined, and the study was probably too small.

5.4.3 The Course's Verbal Scale and Its Limitations

In the tradition of our discipline, the decision is accompanied by a verbal scale that describes how small the p value is [14,19]:

Table 5.2. The verbal scale used in the course for $\alpha = 0.05$.

p value	Decision	Wording
$p \geq 0.05$	H0 is not rejected	statistically nonsignificant difference
$0.01 \leq p < 0.05$	H0 is rejected	significant difference
$0.001 \leq p < 0.01$	H0 is rejected	highly significant difference
$p < 0.001$	H0 is rejected	extremely significant difference

The scale is useful for assignments and for the practical exam, because it forces students to read the p value carefully. It does, however, have two limitations that you need to know. First, "extremely significant" describes how strong the evidence against H0 is, not how large or important the effect is. Height differs between the women and men in mg2026 by 13.4 cm ($p < 0.001$), and SBP differs between the 160 patients and the value of 125 mmHg also with $p < 0.001$, even though in the second case the difference is 6 mmHg; the scale puts them on the same step. Second, the scale has no steps below "nonsignificant", so it does not distinguish between an undetermined result and one that rules out a relevant effect. That is why, in scientific papers, the verbal label never replaces the exact p value, the effect size and the confidence interval.

Two studies by colleagues from our discipline show what these steps look like in practice. In a survey of 121 medical students on data science competencies, the correlation between measured and self-assessed knowledge was $R = 0.107$, with $p = 0.246$, hence nonsignificant, whereas the correlation between the ratings of two courses was $R = 0.869$, $p < 0.001$, extremely significant [16]. The nonsignificant result does not prove that students cannot assess their own knowledge; it says that, with 121 respondents, a weak association could not be distinguished from zero. The confidence interval for R, had it been reported, would have shown how wide the zone of uncertainty is. In the multicenter study of hemodialysis patients from 10 Romanian centers, comparing the 2015 cohort (n = 1401) with the 2019 cohort (n = 1698), one-year mortality was 9.9% versus 10.7% ($p = 0.46$, nonsignificant), and the prevalence of hepatitis C virus infection fell from 16.9% to 10.5% ($p < 0.0001$, extremely significant) [39]. In a sample of more than 3000 patients, a nonsignificant difference of 0.8 percentage points in mortality has a much more solid interpretation than the same difference in 100 patients: the confidence interval is narrow, so an important change in mortality is unlikely.

5.5 Type I and II Errors, Statistical Power

5.5.1 The Four Possible Situations

The statistical decision is made on the basis of a sample, so it can be wrong. We do not know the truth about the population (H0 true or false); we do know our decision (we reject or do not reject H0). Combining the two gives four situations.

Table 5.3. The four possible outcomes of a statistical test.

Decision made	In reality, H0 is true (the effect does not exist)	In reality, H0 is false (the effect exists)
We reject H0	type I error (false positive); probability α	correct decision; probability 1 - β (power)
We do not reject H0	correct decision; probability 1 - α (confidence level)	type II error (false negative); probability β

A type I error means declaring an effect that does not exist: an ineffective drug appears effective, a factor unrelated to the disease appears to be a risk factor. Its probability, when H0 is true, is exactly α, the level we chose. A type II error means missing an effect that does exist: a useful treatment is abandoned because the study "showed no differences". Its probability is denoted β. The complement of β, 1 - β, is the statistical power: the probability of detecting the effect when it truly exists and has a given size. The complement of α, 1 - α, is the confidence level: the probability of not rejecting H0 when it is true; the same quantity appears in the name of the 95% confidence interval.

5.5.2 Two Analogies: The Court Verdict and the Diagnostic Test

The oldest analogy is the criminal trial. The defendant is presumed innocent (H0) until the evidence contradicts this beyond reasonable doubt. Convicting an innocent person is the type I error; acquitting a guilty one is the type II error. The legal system tolerates more wrongful acquittals than wrongful convictions, just as medical statistics sets α strictly (0.05) and accepts a larger β (usually 0.20). Notice one more detail: the verdict is "not guilty" or, more precisely, "guilt has not been proven", not "innocence has been proven". Exactly the same nuance separates "we do not reject H0" from "we accept H0".

The second analogy will be familiar to you from the chapter on diagnostic tests. The statistical test behaves like a laboratory test applied to the hypothesis: "positive" means rejecting H0, "disease" means that the effect exists. A type I error is a false positive result, and α corresponds to 1 - specificity; a type II error is a false negative, and power 1 - β corresponds to sensitivity. As with a diagnostic test, the two errors are linked: if we want fewer false positive results (a smaller α), we will have, with the same sample, more false negative results (a larger β). The analogy goes even further. The predictive value of a positive test depends on the prevalence of the disease; likewise, the probability that a "significant" result corresponds to a real effect depends on how plausible the hypotheses tested in a given field are. In a field where many implausible hypotheses are tested, a large share of the significant results are false positives.

5.5.3 Graphical Representation of α, β and Power

Figure 5.1 places the two possible worlds on the same axis. The curve on the left is the distribution of the test statistic when H0 is true; the curve on the right is its distribution when there is an effect of a given size. The vertical dotted line is the critical value (1.96 for a two-sided z-test with $\alpha = 0.05$). The area under the H0 curve to the right of the critical value is α/2 (the other half lies in the left tail, not drawn). The area under the H1 curve to the left of the critical value is β: in these situations the effect exists, but the statistic does not exceed the threshold. The area under the H1 curve to the right of the critical value is the power.

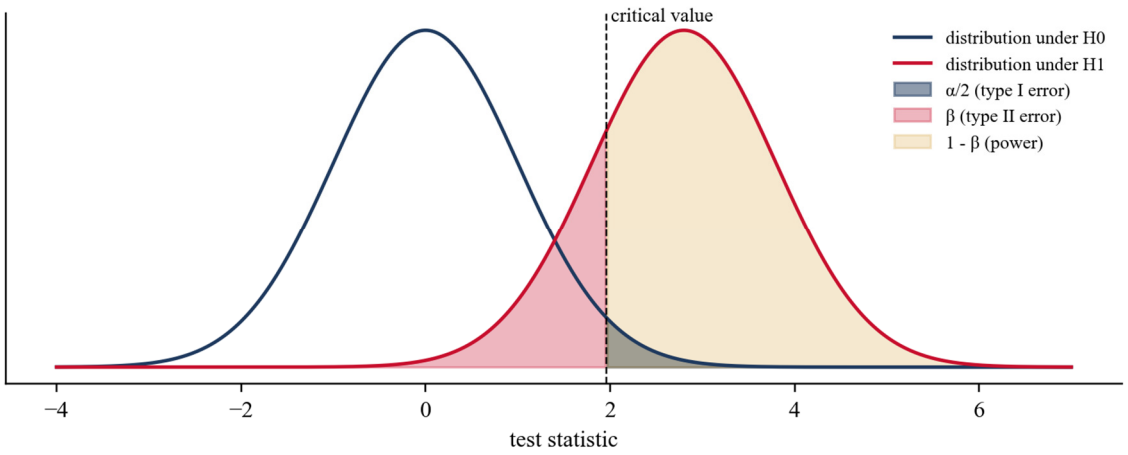


Figure 5.1. Distribution of the test statistic under the null hypothesis (left) and under the alternative hypothesis (right). The critical value (dotted line) divides the axis: the H0 area to its right is α/2 (type I error), the H1 area to its left is β (type II error), and the H1 area to its right is the power 1 - β.

The figure immediately explains why α and β vary in opposite directions. If we move the critical value to the right (a smaller α, for example 0.01), the α area shrinks, but the β area grows. The only ways to reduce both errors at the same time are those that move the curves apart or make them narrower: a larger effect or a smaller standard error, that is, more patients or more precise measurements.

5.5.4 Factors That Influence Power

The power of a test depends on four elements, which must be fixed at the planning stage of the study.

Effect size. A large effect (for example, a difference of 20 mmHg between two antihypertensive treatments) is easily detected, even with few patients. A small effect requires large samples.

Variability. The larger the standard deviation of the variable, the more easily noise drowns out the signal. Standardized measurements, calibrated instruments and inclusion criteria that make the sample more homogeneous increase power. The paired design (the same patients measured twice) increases power precisely because it eliminates between-patient variability (Chapter 6).

Sample size. The standard error decreases with $\sqrt{n}$, so the curves in Figure 5.1 become narrower as n increases. This is the factor the researcher controls most directly.

The α level. A stricter α (0.01 instead of 0.05) reduces power, and a (justified) one-sided test increases it, because the entire α area is placed in a single tail.

Figure 5.2 shows how the power of the t-test for two independent groups increases with the number of patients in each group, for three effect sizes expressed as Cohen's d (Section 5.6). For a medium effect, d = 0.5, 64 patients are needed in each group for 80% power at α = 0.05; for d = 0.3, 176 per group are needed, for d = 0.2 almost 394, and for a large effect, d = 0.8, 26 are enough [76]. If we want 90% power instead of 80% for d = 0.5, the number rises to 86 per group; if we keep 80% power but lower α to 0.01, 96 per group are needed. Calculations of this kind are done with dedicated software, such as G*Power [77], or with the sample size modules in MedCalc and JASP.

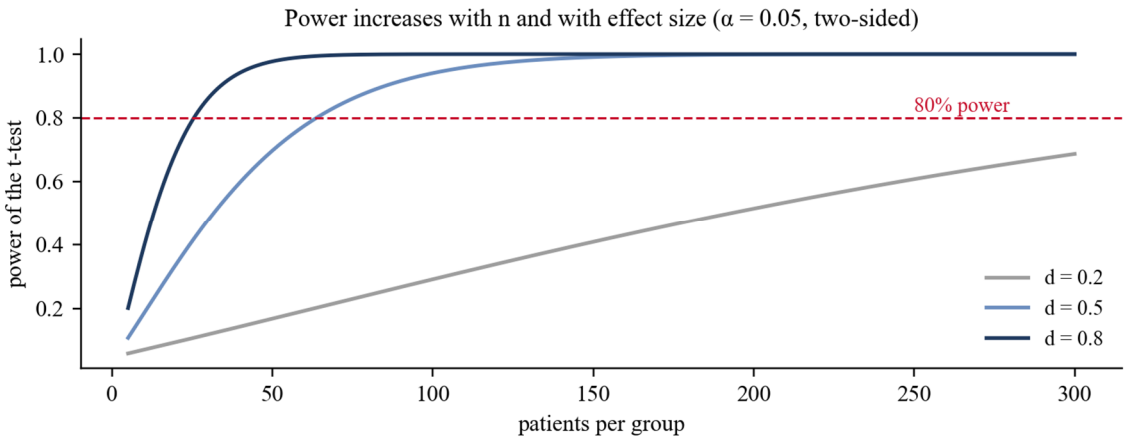


Figure 5.2. Power of the t-test for two independent groups as a function of the number of patients in each group, for three effect sizes (d = 0.2; 0.5; 0.8), at α = 0.05 two-sided. The horizontal line marks 80% power, the minimum value usually accepted when planning a study.

The usual convention, β = 0.20 (80% power), means that we accept missing a real effect in one study out of five. The asymmetry relative to α = 0.05 reflects the view that declaring a nonexistent effect is more serious than missing a real one. In some situations (a serious disease with no other treatment), this ratio could be debated, but the convention remains the one used by ethics committees and journals.

A less intuitive consequence of low power is also worth noting. When an underpowered study nevertheless obtains a significant result, the estimated effect is, on average, exaggerated: only the samples in which chance has "inflated" the difference manage to cross the threshold. That is why spectacular results from small studies must be confirmed in larger studies, and reporting the power (or, better still, the confidence interval) helps readers judge how far they can trust the estimate.

Case study: limited statistical power in an observational cohort

A prospective observational study of 100 patients undergoing endoscopic retrograde cholangiopancreatography (ERCP) compared two strategies for monitoring sedation: entropy-guided sedation (an index derived from the electroencephalogram) versus standard monitoring [78]. Hypertensive episodes occurred in 6% versus 26% of patients (p = 0.007), and Aldrete recovery scores were higher at 5 minutes (p = 0.003) and at 15 minutes (p < 0.001). The authors explicitly report a limited statistical power of 29.3%. What does this mean for the reader? A power of 29.3% means that, if the effect assumed in the calculation truly existed, a study of this size would detect it only in about three cases out of ten. For the analysis to which the calculation refers, β would therefore be about 70%, and a nonsignificant result would say very little about the absence of an effect. For the significant results, low power has a different consequence: the effect estimates may be exaggerated, and confirmation in a larger study is needed. The second point for critical reading is the design. The groups were not randomized, so the differences may also reflect other characteristics of the patients or of the teams. The authors present the results cautiously, as an association ("is associated with"), not as a causal effect, which is exactly the right wording for an observational study with limited power. For your own assignments, remember the order: first the design, then the power, then the p value.

5.6 Effect Size and Clinical Significance

5.6.1 Measures of Effect Size

The p value answers the question "is there an effect?", but not the question that interests the physician: "how large is the effect?". For the second question we use the effect size, a measure that does not depend on sample size. Effect sizes can be expressed in the original units (a difference of 10.19 mg/dL, a risk difference of 18.3 percentage points) or in a standardized form, which allows comparison across different studies and variables.

Cohen's d, the standardized mean difference, divides the difference between two means by the pooled standard deviation:

$$d = \frac{m_1 - m_2}{s_{\text{pooled}}}$$

A d of 0.5 means that the means lie half a standard deviation apart. For paired data, dz = mean of the differences / standard deviation of the differences is used. The correlation coefficient r is itself an effect size, and r² shows the proportion of variability explained. In ANOVA (Chapter 6), η^2 = SS between groups / SS total shows what proportion of the total variability is due to group membership. For binary variables, the odds ratio (OR) and the relative risk (RR) are effect sizes (Chapter 10); there are no universal thresholds for them, and their interpretation depends on the frequency of the disease and on the context.

Table 5.4. Indicative thresholds for interpreting effect size (after Cohen [76]); to be used only in the absence of a clinical criterion.

Measure	Small effect	Medium effect	Large effect
Cohen's d, dz	0,2	0,5	0,8
r (correlation)	0,1	0,3	0,5
η^2 (ANOVA)	0,01	0,06	0,14

Cohen's thresholds are general conventions, originally proposed for the behavioral sciences, and they do not take the place of clinical judgment. A reduction in SBP of 0.2 standard deviations (about 3 mmHg) is "small" according to Cohen, but applied across an entire population it has real effects on the number of strokes.

A few examples from mg2026 give a sense of the orders of magnitude. The difference in height between men and women has d = 2.19, an enormous effect, visible to the naked eye. The reduction in glucose between enrollment and 6 months has dz = 0.88, a large effect. Mean SBP versus 125 mmHg corresponds to $d = \frac{131.13-125}{13.88} = 0.44$, a small-to-medium effect. The difference in SBP between the sexes has d = -0.04, practically zero. In the ANOVA for the reduction in glucose by arm, $\eta^2 = 0.226$, so almost 23% of the variability in the reduction in glucose is explained by the treatment arm, a large effect.

5.6.2 Statistical Significance and Clinical Significance

A result is statistically significant when the data are hard to reconcile with H0. It is clinically significant (clinically relevant) when the effect is large enough to change the physician's decision or the patient's life. The two do not automatically coincide. The threshold of clinical relevance is called the minimal clinically important difference (MCID): the smallest change that patients perceive as beneficial or that justifies changing treatment. It is established on clinical grounds,

from patient studies or from guidelines, not from statistics, and it should be fixed before the data are analyzed.

Combining the two criteria gives four situations that you should learn to recognize quickly:

- statistically significant and clinically relevant: a certain and important effect; the ideal example;
- statistically significant but not clinically relevant: typical of very large samples; the effect exists, but it is too small to matter;
- statistically nonsignificant, with a confidence interval that includes important effects: an undetermined result, probably a study that was too small;
- statistically nonsignificant, with a narrow confidence interval around zero: a relevant effect is unlikely.

The first example from mg2026 is the difference in SBP between the sexes. Women have a mean of 131.37 mmHg (SD 14.47; n = 86), men 130.85 mmHg (SD 13.25; n = 74). The Welch t-test gives t = -0.24, p = 0.81, so the result is nonsignificant. The estimated difference men minus women is -0.5 mmHg, with a 95% CI from -4.9 to 3.8 mmHg, and d = -0.04. Is this an undetermined effect or an absent effect? The confidence interval excludes differences larger than about 5 mmHg in either direction. If we accept that a mean SBP difference of less than 5 mmHg between the sexes has no practical consequences in this group, we can say that a clinically relevant difference is unlikely. The power of the test with 86 and 74 patients was 88% for a medium effect (d = 0.5, that is, about 7 mmHg), but only 24% for a small effect (d = 0.2, about 3 mmHg). Therefore, the study could detect a mean difference of 7 mmHg, but it cannot tell us anything certain about a difference of 2-3 mmHg.

The second example is the reduction in fasting plasma glucose between enrollment and 6 months in the same 160 patients: from 105.4 to 95.2 mg/dL, with a mean difference of 10.19 mg/dL (SD 11.57), 95% CI 8.38-11.99 mg/dL, t = 11.14, 159 degrees of freedom, $p < 0.001$. This is an extremely significant result. Is it also clinically important? Here we have to bring medicine into the discussion. In a patient with fasting plasma glucose in the prediabetes range (100-125 mg/dL, according to the ADA criteria [80]), a reduction of 10 mg/dL can move the value below the 100 mg/dL threshold; as a group mean, the effect is therefore relevant. If, hypothetically, we had decided in advance that only a reduction of at least 15 mg/dL is clinically meaningful, the confidence interval (8.38-11.99) would have shown that the mean effect is certainly below this threshold, even though it is extremely significant statistically. The same figure, two different clinical conclusions, depending on the chosen relevance threshold. It should also be noted that a study without a control group cannot attribute the reduction to the intervention: glucose could also fall through regression to the mean or through the effect of taking part in a study. In mg2026, however, we have three randomized arms, and the correct comparison of efficacy is the one between arms (Chapter 6).

Our colleagues' publications offer examples of both pitfalls. In a four-year analysis of colon cancer care at a Romanian university hospital, in 867 patients, the authors reported an inverse correlation between patient age and the duration of surgery, r = -0.088, p = 0.01 [55]. The result is statistically significant, but $r^2 = 0.0077$, that is, age explains less than 1% of the variability in the duration of surgery. In a sample this large, almost any association that differs from zero becomes significant. In the study of hemodialysis patients, mean age differed between cohorts (59.7 ± 12.92 versus 61 ± 12.93 years, p = 0.005) [39]; a difference of 1.3 years between two cohorts of almost 1500 patients each is highly significant statistically, but it has no clinical importance in itself. An agreement study of three ophthalmic devices, in 170 eyes, found for K1 keratometry values of 43.45 ± 1.64 D with one device and 43.41 ± 1.70 D with another, a significant difference (p = 0.019) [81]; a difference of 0.04 diopters, however, is far below the precision needed to calculate an intraocular lens implant. In the opposite direction, the study on smart

hearing aids reported a huge effect: patients with sensorineural hearing loss who wore hearing aids had self-esteem scores of 35.41 ± 5.32 versus 22.99 ± 4.53 in those without hearing aids ($p < 0.001$, Cohen's d = 2.515), and a three-group ANOVA gave F = 299.00 with $\eta^2 = 0.712$ [82]. Here statistical and practical significance go together, although, since this is a cross-sectional study, the result describes an association and does not show that the hearing aid increased self-esteem.

Case study: a very weak but significant correlation

The retrospective analysis of 867 patients with colon cancer, treated before and during the COVID-19 pandemic in a university hospital in Romania, compared the periods using a series of tests and found a marked decrease in elective surgery ($p < 0.001$), an increase in emergency surgery ($p < 0.001$) and an increase in stage IV cases in the second year of the pandemic ($p = 0.045$) [55]. As a secondary result, the abstract reports an inverse correlation between age and the duration of surgery, r = -0.088, p = 0.01. The critical reading exercise has three steps. First, significance: with n = 867, the standard error of r is approximately $\frac{1}{\sqrt{n}}$, that is, 0.034, and any |r| above about 0.067 becomes significant at $\alpha = 0.05$; the significance threshold is easily exceeded. Next, magnitude: $r^2 = 0.0077$, less than 1% of the variability, a negligible association on any scale (the Colton scale, used in this course, considers any |r| below 0.25 to be "no correlation"). Finally, interpretation: the result does not allow the duration of surgery to be predicted from age and does not justify any clinical decision. The main results of the study (the change in the structure of surgical activity) are differences in proportions with real relevance for the organization of services. What would have added value to the reporting? The confidence interval for r and, for the comparisons of proportions, the absolute differences with their confidence intervals. The lesson for students is that "significant" does not mean "important", especially once n exceeds a few hundred.

5.7 Three Routes to the Same Decision: The Critical Value, the p Value and the Confidence Interval

The decision for a two-sided test at level α can be made along three routes, which always lead to the same conclusion. We show them using the SBP example, with H0: μ = 125 mmHg, n = 160, m = 131.13 mmHg, s = 13.88 mmHg, SE = 1.097 mmHg.

The critical value route. We calculate the test statistic, t = 5.59, and compare it with the critical value of the Student distribution for 159 degrees of freedom and $\alpha = 0.05$ two-sided, critical t = 1.975 (in Google Sheets, =T.INV.2T(0.05, 159)). The rule: we reject H0 if |t| > critical t. Since 5.59 > 1.975, we reject H0. This was the route used when calculations were done with printed tables.

The p value route. We calculate the probability of a value of |t| at least equal to 5.59 under H0: p = T.DIST.2T(5.59, 159), approximately 0.0000001. The rule: we reject H0 if $p < \alpha$. Since $p < 0.05$, we reject H0. This is the route used by all modern software, and it has the advantage of showing how strong the evidence is, not just whether the threshold has been crossed.

The confidence interval route. We construct the 1 - α = 95% confidence interval for the population mean: $m \pm t$ critical × SE $= 131.13 \pm 1.975 \times 1.097 = 131.13 \pm 2.17$, that is, 128.96-133.30 mmHg (the same values as in Chapter 4). The rule: we reject H0 if the value stated by H0 (here 125) lies outside the interval. The value 125 is outside the interval, so we reject H0. This route has the greatest information content: it shows at the same time the decision, the direction and size of the effect, and the precision of the estimate.

The equivalence is not a coincidence. The value 125 falls outside the interval exactly when |m - 125| > critical t × SE, that is, exactly when |t| > critical t, that is, exactly when $p < 0.05$.

Check the same thing for the reference value 130: t = 1.03 < 1.975; p = 0.304 > 0.05; 130 lies within the interval 128.96-133.30. All three routes say: we do not reject H0.

Table 5.5. The three decision routes for mean SBP (n = 160, m = 131.13 mmHg, α = 0.05).

Value under H0	**Calculated t**	**Critical t**	**p value**	**Does the 95% CI contain μ0?**	**Decision**
$\mu_0 = 125$ mmHg	5,59	1,975	< 0,001	no	H0 is rejected
$\mu_0 = 130$ mmHg	1,03	1,975	0,304	yes	H0 is not rejected
$\mu_0 = 140$ mmHg	-8,08	1,975	< 0,001	no	H0 is rejected

The correspondence holds for a 1 - α CI and a two-sided test at level α: a 95% CI with α = 0.05, a 99% CI with α = 0.01. The 99% interval for SBP, 128.27-133.99 mmHg, also excludes 125, so the result remains significant at α = 0.01 as well.

5.8 The Algorithm of a Statistical Test

Whatever the test, the solution follows the same steps. We number them here and will use them in all the following chapters, as well as in the a-g format of the discipline's assignments.

1. State the research hypothesis in clinical language (assignment format: item a).
2. State the statistical hypotheses H0 and H1, in words and mathematically, using the population parameters (item b). Specify whether the test is two-sided or one-sided.
3. Set the significance level α (0.05 by default) before looking at the results.
4. Identify the type of variables and the design (one mean compared with a value, two independent groups, two measurements on the same patients, more than two groups, qualitative variables) and choose the test (item c). The complete selection algorithm appears in Chapter 8.
5. Check the conditions for applying the test (normality, equality of variances, independence of observations, expected frequencies); if they are not met, choose the appropriate alternative, usually a nonparametric test.
6. Calculate the test statistic, by hand or with software (item d), and obtain the p value (and/or the critical value).
7. Calculate the effect size and the confidence interval for the effect (item e, together with the graph).
8. Make the statistical decision: p < α, H0 is rejected; p ≥ α, H0 is not rejected; add the course's verbal label (item f).
9. State the research conclusion in words, with the main figures and the clinical interpretation, including the limitations (item g).

Steps 3 and 4 are done before step 6. Changing the test or the threshold after you have seen the p value turns the analysis into a search for convenient results (Section 5.11).

5.9 The z and t Tests for a Single Mean

The simplest test compares the mean of a sample with a value given in advance: a reference value from the literature, a therapeutic target, the national mean. The discipline's course distinguishes

two variants, the z-test and the t-test, depending on what we know about the variability of the population.

5.9.1 The z-Test

The z-test is used when the population standard deviation, σ, is known (for example, from a large national registry or from the specifications of a measuring instrument) and the sample is large enough (n ≥ 30) for the mean to follow the normal distribution, even if the variable itself is not perfectly normal (the central limit theorem, Chapter 4). The test statistic is:

$$z = \frac{m - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Under H0, z follows the standard normal distribution. The two-sided critical values are 1.96 for α = 0.05, 2.576 for α = 0.01 and 3.291 for α = 0.001; for a one-sided test with α = 0.05, the critical value is 1.645.

Suppose, for the sake of the example, that a large population study has established, for adults of the same age in the region, a mean total cholesterol of 200 mg/dL, with $\sigma = 40$ mg/dL. In mg2026, the mean cholesterol is 209.39 mg/dL in n = 160 patients. The standard error calculated with known σ is 40 / √160 = 40 / 12.649 = 3.162 mg/dL, and $z = \frac{209.39-200}{3.162} = \frac{9.39}{3.162} = 2.97$. Since 2.97 > 1.96 (and even > 2.576), we reject H0; the two-sided p value is 2 × P(Z > 2.97) = 0.003, so the difference is highly significant. The patients at cardiometabolic risk in the study have a higher mean cholesterol than the general reference population.

5.9.2 The t-Test for a Single Mean

In practice, σ is almost always unknown, and we estimate it by the sample standard deviation, s. Replacing σ with s adds uncertainty, especially in small samples, and the statistic

$$t = \frac{m - \mu_0}{\frac{s}{\sqrt{n}}}$$

no longer follows the normal distribution but Student's t distribution, described in 1908 by William Gosset, a chemist at a Dublin brewery, who published under the pseudonym "Student" [83]. The t distribution resembles the normal distribution, but it has heavier tails; its shape depends on the number of degrees of freedom, df = n - 1. The larger n is, the closer the t distribution comes to the normal, and for df above 120 the difference becomes negligible.

Table 5.6. Critical values of the t distribution (two-sided α = 0.05 and 0.01; one-sided α = 0.05).

df	two-sided 0.05	two-sided 0.01	one-sided 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
infinity (z)	1,960	2,576	1,645

For cholesterol, with $s = 36.91$ mg/dL instead of σ = 40, SE $= \frac{36.91}{12.649} = 2.918$ mg/dL and $t = \frac{9.39}{2.918} = 3.22$, with 159 degrees of freedom; p = 0.0016. The conclusion does not change. At n = 160, the choice between z and t matters little; at n = 10 it would have mattered a great deal (the critical value would be 2.262 instead of 1.96).

The course rule (the z-test for n ≥ 30 and known σ, the t-test for n < 30 or unknown σ) can be simplified for today's practice: when σ is not known, we use the t-test whatever the sample size. Statistical software does exactly this; JASP, for example, offers in the T-Tests > One Sample T-Test menu Student's t-test and, as a separate option, the z-test, for which it asks you to enter the value of σ.

5.9.3 Conditions for Application

The t-test for a mean assumes that the observations are independent (each patient contributes only once) and that the variable has an approximately normal distribution in the population. The second condition matters mainly in small samples; with large n, the central limit theorem makes the test robust to moderate departures from normality. Normality is checked with the Shapiro-Wilk test [66] and with the Q-Q plot (Chapter 3). If the variable is strongly skewed and the sample is small, the nonparametric alternative is the Wilcoxon signed-rank test for a single median (Chapter 7) or a transformation of the data (for example, log transformation of triglycerides).

For fasting plasma glucose in mg2026, the Shapiro-Wilk test gives p = 0.037, hence a slight departure from normality (skewness 0.37). With n = 160, the t-test remains usable, and the Wilcoxon test leads to the same conclusion (Worked Problem 5.2). For BMI, the skewness is large (1.42) and Shapiro-Wilk p < 0.001; here reporting the median and a nonparametric test is more appropriate, even though n is large.

5.10 Multiple Comparisons and Inflation of the Type I Error

5.10.1 Why the Risk of a False Positive Increases

The α = 0.05 level controls the type I error for a single test. If we perform several independent tests, each at α = 0.05, the probability of obtaining at least one false positive result increases rapidly:

$$P(\text{at least un fals positive}) = 1 - (1 - \alpha)^k = 1 - 0.95^k$$

For k = 3 tests, the probability is 0.143; for k = 5, 0.226; for k = 10, 0.401; for k = 20, 0.642. In other words, a student who compares women and men on 20 variables for which there is in reality no difference at all has almost two chances in three of "discovering" at least one significant difference.

Figure 5.3 illustrates the phenomenon in two ways. In panel A, we simulated 1000 studies in which two groups of 20 patients each come from the same population (SBP with a mean of 130 and an SD of 14 mmHg), so H0 is true every time. The p values are uniformly distributed between 0 and 1, and 4.9% of them fall below 0.05: exactly the proportion of false positive results promised by α. Panel B shows the curve $1 - 0.95^k$ and the effect of the Bonferroni correction, which keeps the overall risk close to 5%. You will repeat the same simulation in Google Sheets in Practical Session 5.

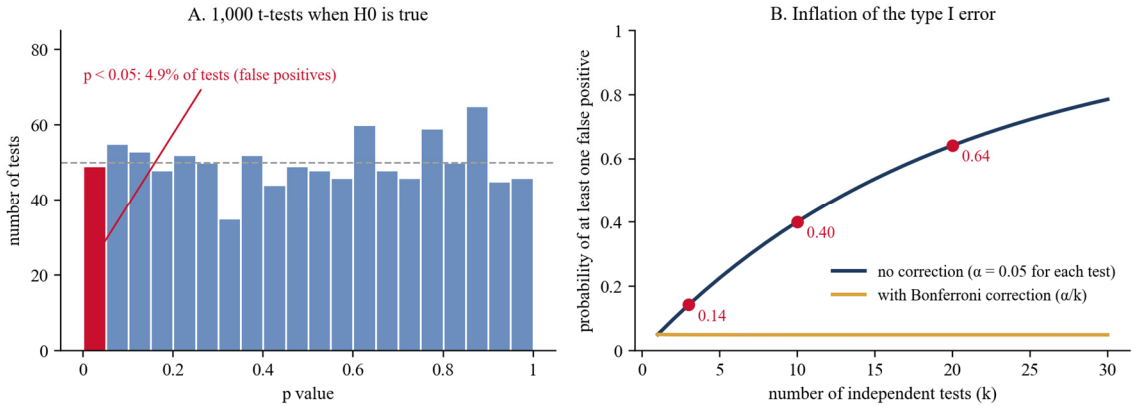


Figure 5.3. A. Distribution of p values in 1000 simulated t-tests when the null hypothesis is true (two groups of 20 values each from the same normal population): the p values are uniformly distributed, and about 5% are below 0.05. B. Probability of obtaining at least one false positive result as a function of the number of independent tests, without correction and with the Bonferroni correction. Simulated data.

5.10.2 The Bonferroni, Holm and Benjamini-Hochberg Corrections

The Bonferroni correction divides the threshold by the number of tests: each test is declared significant only if $p < \alpha/k$. Equivalently, software reports adjusted p values, $p \times k$ (capped at 1), which are compared with the usual α. The method is simple and safe, but conservative: with many tests it loses a great deal of power.

The Holm method (sequential Bonferroni) keeps the same guarantee but is less conservative. We sort the p values in ascending order, $p(1) \leq p(2) \leq ... \leq p(k)$. We compare p(1) with α/k, p(2) with α/(k - 1), p(3) with α/(k - 2), and so on. We stop at the first test that does not pass the threshold; this test and all the following ones are declared nonsignificant.

The Benjamini-Hochberg (BH) method controls a different quantity: not the probability of having any false positive, but the expected proportion of false positives among the results declared significant (the false discovery rate, FDR). We sort the p values in ascending order and compare p(i) with $(i/k) \times \alpha$. We find the largest i for which $p(i) \leq (i/k) \times \alpha$ and declare significant all the tests from 1 to that i. The method is suitable when we perform many exploratory tests and accept that a small fraction of the "discoveries" may be false, provided we know how large that fraction is.

With the mg2026 dataset we can do exactly the tempting exercise described above: we compare women and men on 14 numerical variables measured at enrollment, using the Welch t-test. The results, sorted by p value, are shown in Table 5.7.

Table 5.7. Comparison of women and men on 14 variables in mg2026, with uncorrected and adjusted p values (Welch t-test).

Variable	Mean F	Mean M	Uncorrected p	Bonferroni p	Holm p	BH p
height (cm)	162,60	175,97	< 0,001	< 0,001	< 0,001	< 0,001
weight (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
waist circumference (cm)	86,34	91,09	0,013	0,18	0,14	0,046
triglycerides (mg/dL)	171,21	149,51	0,068	0,95	0,68	0,19
age (years)	55,42	57,31	0,17	1	1	0,40
FINDRISC	10,02	10,77	0,27	1	1	0,55

Variable	Mean F	Mean M	Uncorrected p	Bonferroni p	Holm p	BH p
glicemie$_0$(mg/d	104,81	106,16	0,59	1	1	0,79
DBP (mmHg)	78,88	79,62	0,59	1	1	0,79
cholesterol (mg/dL)	208,08	210,91	0,63	1	1	0,79
HbA1c (%)	5,56	5,59	0,68	1	1	0,79
BMI $\left(\frac{\text{kg}}{m^2}\right)$	28,92	29,18	0,68	1	1	0,79
SBP (mmHg)	131,37	130,85	0,81	1	1	0,88
HR (beats/min)	74,69	74,73	0,98	1	1	0,98

Height, weight and HDL differ between the sexes regardless of the correction; these are well-known biological differences. Waist circumference has an uncorrected p of 0.013: significant without correction, nonsignificant after Bonferroni (the individual threshold becomes 0.05/14 = 0.0036) and after Holm, but still significant after Benjamini-Hochberg (adjusted p 0.046). The difference between the methods is not a contradiction; it reflects different questions. If we want no false positives at all in the whole table, we use Bonferroni or Holm; if we accept that about 5% of the results declared positive may be false, BH is sufficient.

When is a correction needed? When the tests belong to the same question and the conclusion is drawn from "at least one significant": multiple comparisons between groups after an ANOVA (Chapter 6), many outcome variables in one study, analyses in many subgroups, repeated interim looks at the data. It is not needed for a single primary outcome specified in advance in the protocol. The best protection remains choosing a single primary outcome before data collection; secondary outcomes are reported as exploratory.

An example from our colleagues' publications shows how easily tests accumulate. The cross-sectional study of the association between the severity of HIV infection and psychiatric disorders, in 101 patients, reports seven odds ratios, with p values between 0.038 and below 0.001 (for example, OR = 5.81, p < 0.001; OR = 2.81, p = 0.038) [84]. Each result taken separately is significant. With seven tests, the Bonferroni threshold $\frac{\text{ar fi} \quad 0.05}{7} = 0.0071$; results with p = 0.016, 0.024 or 0.038 would not pass it. They may be real, but they should be regarded as hypotheses to be verified, not as established results. Reporting confidence intervals for each OR would also have helped the reader judge the precision of the estimates.

Case study: the Benjamini-Hochberg correction in an emergency department study

A retrospective study from an emergency department analyzed the differences between women and men in 213 patients (100 women, 113 men) treated for suspected bacterial infection [9]. The women were older (median 77.0 versus 67.0 years, p < 0.001), and a urinary source of infection was more frequent in women (51.0% versus 34.5%, p = 0.015). For this association, the abstract reports an OR of 2.14 (95% CI 1.20-3.81; p = 0.010) and, after the Benjamini-Hochberg correction for multiple testing, an adjusted p = 0.039. The authors compared numerous characteristics between the sexes; the correction for multiple testing was therefore justified, and the result remained significant after it, which strengthens its credibility. For the clinical outcomes, the picture is different: in-hospital mortality was 7.0% in women (7 of 100) and 8.0% in men (9 of 113), with an OR of 0.87 (95% CI 0.31-2.43), and another outcome occurred in 14 of 100 women (14.0%) and in 21 of 113 men (18.6%), OR 0.71 (95% CI 0.34-1.49). The abstract also reports an adjusted OR of 0.80 (95% CI 0.27-2.35; p = 0.690). The confidence intervals are very wide, ranging from an almost threefold reduction in the odds to a more than twofold increase. With 16 deaths in total, the study can rule out neither a

protective effect nor an important unfavorable effect; the result is undetermined, not "absent". There are three lessons: the correction for multiple comparisons is applied when many hypotheses are tested in the same study; a result that survives the correction is more convincing; with rare events, the confidence interval says more than the p value. The comparison between the crude OR and the adjusted OR (here, both nonsignificant) will be taken up again in the chapters on risk analysis and logistic regression.

5.11 p-Hacking, HARKing and Preregistration

The p value keeps its meaning only if the analysis was decided before seeing the data. When researchers try several analyses and report only the one "that worked", the nominal 5% threshold is no longer real. This practice is called p-hacking (hunting for p values) and takes seemingly innocent forms: a few more patients are added until p drops below 0.05; "outliers" are excluded only after it has been seen that the result becomes significant; the t-test is tried, then Mann-Whitney, then the logarithmic transformation; patients are split into subgroups (by sex, by age, by place of residence) until a subgroup with an effect appears; the primary outcome is swapped for a secondary one. Each step seems reasonable taken separately, but together they turn five attempts into a single "significant" report, with a real risk of a false positive well above 5%.

HARKing (hypothesizing after the results are known) means formulating the hypothesis after the results have been seen and presenting it as if it had been specified in advance. An association between smoking and some variable, observed by chance, becomes "our main hypothesis" in the introduction of the article. The reader can no longer distinguish confirmation from exploration.

The antidote is preregistration: the study protocol, with the main hypothesis, the primary outcome, the sample size and the statistical analysis plan, is published in a public registry before the first patient is enrolled. For randomized clinical trials, registration is mandatory, and the CONSORT guideline requires that any deviation from the protocol be reported [56]. For observational studies, the STROBE guideline requires a description of all the analyses performed, not just the significant ones [25]. At your level, the practical rule is this: write down the hypotheses and the chosen test (items a-c of the assignment format) before you open the data file in JASP.

Not every unplanned analysis is wrong. Exploring the data generates new hypotheses; the problem arises only when exploration is presented as confirmation. The honest wording is "in exploratory, unplanned analyses, we observed...", and it tells the reader that the result needs to be verified in a new study. Sensitivity analyses are an example of good practice: the authors repeat the analysis under different conditions to see whether the result holds. In the cross-sectional study of disorders of gut-brain interaction, with 136 participants analyzed (52 cases and 84 controls), BMI was higher in cases (28.15 ± 6.49 versus $24.47 \pm \frac{4.60\ \text{kg}}{m^2}$, $p = 0.001$), and a dietary factor had a protective OR of 0.36 (95% CI 0.18-0.74; p = 0.009); the authors repeated the analysis after excluding participants with C-reactive protein above $\frac{10\text{mg}}{L}$ $(n = 98)$ [71]. A result that holds up in the sensitivity analysis is more credible than one that appears in only a single version of the analysis.

Case study: the within-group comparison versus the between-group comparison

A placebo-controlled clinical trial, reported according to the CONSORT guideline, evaluated over 3 months the effect of Aloe ferox extract (2 capsules/day, 460 mg) combined with diet in 20 people with obesity, compared with 20 matched controls treated with diet and placebo [85]. The abstract reports two types of results: between-group comparisons at 3 months and before-after comparisons within the groups, with p values such as 0.03 and 0.002 for BMI, 0.01 and 0.039 for LDL cholesterol, 0.018 and below 0.001 for fasting plasma

> glucose, 0.032 for total cholesterol and 0.006 for diastolic blood pressure. The critical reading exercise does not concern this study in particular, but a classic error in the clinical literature that every reader must be able to avoid. Suppose a study in which glucose decreases significantly in the treated group (p = 0.01 before-after) and nonsignificantly in the placebo group (p = 0.08). The conclusion "the treatment is effective and placebo is not" is wrong: two results on either side of the threshold do not show that the effects differ between the groups. The decrease in the placebo group may be almost as large, merely estimated less precisely. The only test that answers the question about efficacy is the direct comparison between groups: either comparing the decreases (the individual differences) between groups with the independent-samples t-test, or comparing the final values adjusted for the baseline value (ANCOVA, Chapter 6). Within-group comparisons only describe how each group changed, and they are influenced by regression to the mean and by the placebo effect. When you read a controlled study, look first for the between-group comparison; in the Aloe ferox study, the abstract reports it, and the reader only needs to know which of the p values to rely on for the conclusion about efficacy. With 20 patients per group, the confidence intervals of the differences would have been a useful complement to the p values.

5.12 Equivalence and Non-Inferiority, in Brief

The classic test can reject the hypothesis "no difference", but it cannot confirm it. Sometimes, however, that is exactly what we want to show: that a generic drug has the same effect as the original, that a cheap measurement method gives the same results as an expensive one, or that a more convenient treatment is not worse than the standard one. For such questions, the hypotheses are reversed.

In an equivalence trial, the researcher sets in advance an equivalence margin Δ (for example, 5 mmHg for SBP), that is, the largest difference considered clinically unimportant. H0 becomes "the difference is at least Δ in absolute value" (the treatments are not equivalent), and H1 "the difference is smaller than Δ" (the treatments are equivalent). In practice, equivalence is demonstrated if the entire confidence interval for the difference lies between -Δ and +Δ; the procedure is called TOST (two one-sided tests). In a non-inferiority trial only one direction is of interest: the new treatment is not worse than the standard one by more than Δ; this is the rare case in which a one-sided test is justified.

The SBP-by-sex example is instructive. The 95% CI for the difference is -4.9 to 3.8 mmHg. If the equivalence margin set in advance had been 5 mmHg, the interval would lie entirely within the zone -5 to +5, so we could claim equivalence of the two means. If the margin had been 3 mmHg, we could claim neither a difference nor equivalence. For the rigorous calculation, a 90% CI is usually used, corresponding to the two one-sided tests at $\alpha = 0.05$; the basic idea, however, remains the same: equivalence is demonstrated with the confidence interval, never with a $p > 0.05$.

5.13 How to Report a Statistical Test

Reporting a test in an article, an undergraduate thesis or an assignment must allow the reader to retrace the reasoning. Current standards, set out in the ICMJE recommendations and in the CONSORT and STROBE guidelines [25,36,56], require a few mandatory elements: the descriptive statistics for each group (mean and SD, or median and IQR, with n); the name of the test, its variant (for example, Welch t, two-sided) and the α level; the value of the test statistic, with its degrees of freedom; the exact p value; the effect size, with its 95% confidence interval; the clinical interpretation, including the relevance threshold, if there is one.

An example of complete reporting for mg2026: "Fasting plasma glucose decreased from 105.4 mg/dL at enrollment to 95.2 mg/dL at 6 months (mean difference 10.19 mg/dL, 95% CI 8.38-11.99; paired-samples t-test, t(159) = 11.14, p < 0.001; dz = 0.88). The distribution of individual differences did not depart significantly from normality (Shapiro-Wilk p = 0.725)."

An example of reporting a nonsignificant result: "SBP did not differ significantly between women $(131.37 \pm 14.47$ mmHg, $n = 86)$ and men $(130.85 \pm 13.25$ mmHg, $n = 74)$; the difference men minus women was -0.5 mmHg (95% CI -4.9 to 3.8; Welch t-test, $t(157.4) = -0.24$, $p = 0.81$; $d = -0.04$)." Notice that the sentence does not say "SBP is equal in women and men"; the confidence interval lets the reader judge which differences are ruled out.

A few phrasings to avoid: "p = 0.000" (write p < 0.001); "NS" without any figure; "a trend toward significance" for p = 0.07 (either we interpret the confidence interval or we say that the result is nonsignificant); "extremely significant difference" without the size of the difference; "the results demonstrate that..." in an observational study.

The retrospective study of the course of COVID-19 in patients with type 1 and type 2 diabetes mellitus, in 486 patients (200 without diabetes, 62 with type 1 diabetes, 224 with type 2 diabetes), is a good reading exercise [8]. The abstract reports clear proportions (severe forms 33.5% in type 2 versus 25.8% in type 1; mortality 11.6% versus 8.1%) and means with SD (length of hospital stay 11.6 ± 7.0 versus 9.1 ± 5.8 days; glucose on admission 178.3 ± 34.7 versus 164.8 ± 39.6 mg/dL), but the odds ratios for admission to intensive care (OR 2.24) and for mechanical ventilation (OR 2.46) appear without a confidence interval and without a p value. Without an interval, the reader cannot judge whether an OR of 2.24 is estimated precisely (for example, 1.8-2.8) or very imprecisely (for example, 0.9-5.6). In addition, the large standard deviations relative to the mean length of hospital stay suggest a skewed distribution, for which the median and IQR would have been more informative. This observation does not lessen the clinical value of the study; it only shows how much a few extra figures in the abstract matter to the reader.

5.14 Worked Problems

Worked Problem 5.1. From the Question to the Hypotheses

Statement. For each of the questions below, state the research hypothesis and the statistical hypotheses in words and mathematically, and specify whether the test is two-sided or one-sided. (a) In mg2026, does resting heart rate differ between smokers and nonsmokers? (b) Is the proportion of patients with diabetes confirmed by OGTT different from 10%, a hypothetical prevalence in the general adult population? (c) A manufacturer claims that a new glucose test strip does not underestimate values by more than 3 mg/dL compared with the reference laboratory. (d) Does SBP change between enrollment and 6 months in the same patients?

Solution. (a) Research hypothesis: the resting heart rate of smokers differs from that of nonsmokers. H0: in the population, mean HR is the same in smokers and nonsmokers, μsmokers = μnonsmokers. H1: the means differ, μsmokers ≠ μnonsmokers. Two-sided test: a difference in either direction would be of interest. (b) Research hypothesis: the prevalence of diabetes in the population at cardiometabolic risk differs from that in the general population. H0: π = 0.10; H1: π ≠ 0.10. Two-sided test. (c) This is a non-inferiority question: H0: the strip underestimates by at least 3 mg/dL (mean difference strip minus laboratory ≤ -3); H1: the mean underestimation is less than 3 mg/dL (difference > -3). One-sided test, justified by the nature of the question, with the margin set in advance. (d) Research hypothesis: over the 6 months of intervention, SBP changes. H0: the mean of the individual differences (SBP0 - SBP6) is zero in the population, μD = 0; H1: μD ≠ 0. Even if we expect a decrease, the test remains two-sided: an increase in SBP would be an

important result (for example, an adverse effect) that must not be ignored. In mg2026, the paired test shows a mean decrease of 4.78 mmHg (95% CI 3.36-6.19; t = 6.68; p < 0.001).

Interpretation. In all four cases, H0 contains the equals sign (or, for non-inferiority, "equal or worse"), that is, the precise statement that allows the distribution of the test statistic to be calculated.

Worked Problem 5.2. Fasting Plasma Glucose versus 100 mg/dL

Statement. In mg2026, fasting plasma glucose at enrollment has a mean of 105.44 mg/dL, a standard deviation of 15.76 mg/dL, n = 160. The ADA threshold for impaired fasting glucose is 100 mg/dL [80]. Does the mean of the population of patients at cardiometabolic risk differ from 100 mg/dL? Use α = 0.05 and all three decision routes.

Solution. Research hypothesis: mean fasting plasma glucose in the population at cardiometabolic risk differs from the 100 mg/dL threshold. H0: $\mu = 100$ mg/dL; H1: μ ≠ 100 mg/dL; two-sided, α = 0.05. The test: a mean compared with a given value, σ unknown, hence the t-test for a single mean. Conditions: independent observations; Shapiro-Wilk p = 0.037 shows a slight departure from normality (skewness 0.37), but with n = 160 the mean has a practically normal distribution, and the t-test is robust; we will nevertheless check with the Wilcoxon test.

Calculation. $\text{SE} = \frac{15.76}{\sqrt{160}} = \frac{15.76}{12.649} = 1.246$ mg/dL. $t = \frac{105.44-100}{1.246} = \frac{5.44}{1.246} = 4.36$, with df = 159.

The critical value route: critical t (159; 0.05 two-sided) = 1.975; 4.36 > 1.975, so we reject H0. The p value route: p = T.DIST.2T(4.36, 159) = 0.00002, that is, p < 0.001. The CI route: $\text{CI}\,95\% = 105.44 \pm 1.975 \times 1.246 = 105.44 \pm 2.46$, that is, 102.98-107.90 mg/dL; the value 100 does not lie within the interval, so we reject H0. Effect size: $d = \frac{5.44}{15.76} = 0.35$ (a small-to-medium effect). The Wilcoxon test for the median versus 100 gives p < 0.001, the same conclusion.

Statistical conclusion: p < 0.001, H0 is rejected; the difference is extremely significant. Research conclusion: mean fasting plasma glucose in the population of adults at cardiometabolic risk from which the sample comes is higher than the 100 mg/dL threshold, by about 3-8 mg/dL (95% CI for the mean 102.98-107.90 mg/dL). Clinical interpretation: the "average" patient in the study lies in the impaired fasting glucose range, which corresponds to the inclusion criterion (cardiometabolic risk). If we tested the value 105 mg/dL, we would obtain t = 0.35 and p = 0.73: the data are compatible with a mean of 105 mg/dL, which does not mean that the mean is exactly 105.

Worked Problem 5.3. The z-Test and the t-Test for Cholesterol

Statement. A national registry indicates for adults aged 35-79 years a mean total cholesterol of 200 mg/dL, with $\sigma = 40$ mg/dL (hypothetical values). In mg2026, n = 160, $m = 209.39$ mg/dL, $s = 36.91$ mg/dL. (a) Apply the z-test. (b) Apply the t-test. (c) Comment on the difference.

Solution. H0: $\mu = 200$ mg/dL; H1: μ ≠ 200 mg/dL; α = 0.05, two-sided. (a) With known σ: $\text{SE} = \frac{40}{\sqrt{160}} = 3.162$; $z = \frac{9.39}{3.162} = 2.97$. The critical value of 1.96 is exceeded, and 2.97 > 2.576 also exceeds the threshold for α = 0.01. p = 2 × (1 - Φ(2.97)) = 0.003. H0 is rejected; the difference is highly significant. (b) With σ estimated by s: $\text{SE} = \frac{36.91}{12.649} = 2.918$; $t = \frac{9.39}{2.918} = 3.22$, df = 159; p = 0.0016, also highly significant. (c) The two tests differ only in how the noise is estimated. Here s (36.91) is smaller than the registry σ (40), which is why t is slightly larger than z. If the registry σ also describes our population correctly, the z-test is the appropriate one; if our patients are more homogeneous than the general population, s better reflects their variability and

the t-test is preferable. Research conclusion: patients at cardiometabolic risk have a higher mean cholesterol than the reference population, by about 9 mg/dL (95% CI for the mean, calculated with t: 203.6-215.2 mg/dL).

Worked Problem 5.4. A Nonsignificant Result in a Small Sample

Statement. A student had access to only 10 patients from mg2026, with SBP values of: 150, 154, 131, 130, 137, 102, 110, 103, 147, 136 mmHg. Test whether the mean differs from 125 mmHg and interpret the result in the context of the analysis of all 160 patients.

Solution. H0: μ = 125; H1: μ ≠ 125; α = 0.05. The sum of the values is 1300, so m = 130.0 mmHg. The deviations from the mean are 20, 24, 1, 0, 7, -28, -20, -27, 17, 6; their squares are 400, 576, 1, 0, 49, 784, 400, 729, 289, 36, with a sum of 3264. The variance $s^2 = \frac{3264}{9} = 362.67$, so s = 19.04 mmHg. SE $= \frac{19.04}{\sqrt{10}} = \frac{19.04}{3.162} = 6.02$ mmHg. $t = \frac{130.0-125}{6.02} = 0.83$, df = 9. The critical value for df = 9 is 2.262; 0.83 < 2.262, so we do not reject H0. The p value = T.DIST.2T(0.83, 9) = 0.43. CI 95% $= 130.0 \pm 2.262 \times 6.02 = 130.0 \pm 13.62$, that is, 116.38-143.62 mmHg, an interval that contains 125.

Interpretation. Statistical conclusion: p = 0.43 > 0.05, H0 is not rejected; the difference is statistically nonsignificant. The research conclusion cannot be "mean SBP is 125 mmHg": the confidence interval covers values from 116 to 144 mmHg, that is, from optimal blood pressure to grade 1 hypertension. The result is undetermined, not negative. In all 160 patients, with SE = 1.097 (5.5 times smaller), the same question receives a clear answer: t = 5.59, p < 0.001. The difference between the two analyses does not come from contradictory data, but from the power of the test. For an effect of the size observed in the whole sample (d ≈ 0.44), the power of a t-test on 10 patients is very low, of the order of 20-25%; the failure to reject H0 was predictable.

Worked Problem 5.5. Bonferroni, Holm and Benjamini-Hochberg

Statement. An exploratory study compares two groups on six outcomes and obtains the p values: 0.012; 0.001; 0.20; 0.041; 0.008; 0.030. How many results are significant at α = 0.05 (a) without correction, (b) with Bonferroni, (c) with Holm, (d) with Benjamini-Hochberg?

Solution. We sort them in ascending order: 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) Without correction, all values below 0.05 are significant: five results. The probability of having at least one false positive, if all the H0 were true, ar fi $1 - 0.95^6 = 0.265$.

(b) Bonferroni: the threshold is 0.05 / 6 = 0.0083. Only 0.001 and 0.008 pass: two results. The adjusted p values (p × 6) are 0.006; 0.048; 0.072; 0.18; 0.246; 1.

(c) Holm: we compare p(1) = 0.001 with 0.05/6 = 0.0083 (passes); p(2) = 0.008 with 0.05/5 = 0.010 (passes); p(3) = 0.012 with 0.05/4 = 0.0125 (passes); p(4) = 0.030 with 0.05/3 = 0.0167 (does not pass). We stop: three significant results.

(d) Benjamini-Hochberg: the thresholds (i/6) × 0.05 are 0.0083; 0.0167; 0.025; 0.0333; 0.0417; 0.05. We compare: 0.001 ≤ 0.0083; 0.008 ≤ 0.0167; 0.012 ≤ 0.025; 0.030 ≤ 0.0333; 0.041 ≤ 0.0417; 0.20 > 0.05. The largest i for which p(i) ≤ threshold is i = 5, so five results are significant.

Interpretation. The methods answer different questions. Bonferroni and Holm guarantee that the probability of having any false positive in the whole set does not exceed 5%; Holm is always at least as powerful as Bonferroni and should be preferred. BH accepts that, on average, at most 5% of the results declared significant are false; it is suited to exploratory analyses with many tests. The choice of method is made before the analysis and reported in the methods section.

Worked Problem 5.6. Power and Sample Size

Statement. A family physician wants to compare two hypertension education programs in a randomized trial with two groups. The physician considers a difference in mean SBP of 5 mmHg to be clinically relevant and knows from mg2026 that the standard deviation of SBP is about 14 mmHg. (a) What effect size does this correspond to? (b) How many patients are needed in each group for 80% power at α = 0.05 two-sided? (c) What power would a study with 53 patients per group have, like the arms in mg2026?

Solution. $(a)\ d = \frac{\Delta}{\sigma} = \frac{5}{14} = 0.36$, a small-to-medium effect. (b) The approximate formula for comparing two means is n per group = 2 × (z1-α/2 + z1-β)² × σ² / Δ². For α = 0.05 two-sided, $z_1 - \frac{\alpha}{2} = 1.96$; for 80% power, z1-β = 0.84. This gives $n = \frac{2\times(1.96+0.84)^2\times 14^2}{5^2} = \frac{2\times 7.84\times 196}{25} = 122.9$, that is, 123 patients per group; the exact calculation with the t distribution (G*Power, JASP) gives 125. To these, 10-20% is usually added for losses to follow-up. (c) With 53 patients per group, the power for d = 0.36 is only about 45%: the study would miss a real effect of 5 mmHg more often than it would detect it.

Interpretation. The sample size calculation is done before the study and starts from the clinically relevant difference, not from the difference we hope to find. A study with 53 patients per arm is sufficient for large effects (the reduction in glucose in mg2026 differs between arms with standardized effects between about 0.5 and 1.4), but too small for a difference of 5 mmHg in SBP.

Worked Problem 5.7. The p Value from the t Statistic

Statement. Without statistical software, using only Table 5.6, decide at α = 0.05: (a) t = 2.10, df = 20, two-sided; (b) t = 1.80, df = 25, two-sided and then one-sided (in the direction stated in advance); (c) t = 3.10, df = 9, two-sided. Then check the exact p values with Google Sheets.

Solution. (a) The critical value for df = 20 is 2.086; 2.10 > 2.086, so p is slightly below 0.05; H0 is rejected, at the borderline. In Sheets, =T.DIST.2T(2.1, 20) gives 0.049. (b) The two-sided critical value for df = 25 is 2.060; 1.80 < 2.060, so p > 0.05 and H0 is not rejected (two-sided p = T.DIST.2T(1.8, 25) = 0.084). One-sided, the critical value is approximately 1.708 (=T.INV(0.95, 25)); 1.80 > 1.708, so the one-sided p = 0.042 < 0.05. The same result is "significant" one-sided and "nonsignificant" two-sided; this is precisely why the choice must be made and justified before the analysis. (c) For df = 9, the critical values are 2.262 (α = 0.05) and 3.250 (α = 0.01); 3.10 lies between them, so 0.01 < p < 0.05; in Sheets, p = 0.013. The result is significant, but not highly significant.

5.15 Proposed Exercises

1. For HbA1c at enrollment, mg2026 has m = 5.58%, s = 0.45%, n = 160. Test whether the population mean differs from 5.7% (the lower limit of the prediabetes range according to the ADA). Calculate t, decide at α = 0.05 and state the research conclusion.
2. The FINDRISC score in mg2026 has m = 10.37, s = 4.27, n = 160. Does the mean differ from 12 (a threshold chosen for the exercise)? Calculate t, the 95% CI for the mean, and d.
3. Resting heart rate at enrollment: m = 74.71 beats/min, s = 9.30, n = 160. Test H0: μ = 72 beats/min. Calculate t and the effect size and comment on the clinical relevance of the difference.
4. In a group of 36 children from a rural community, the mean hemoglobin is 11.2 g/dL. From large studies it is known that, at this age, $\sigma = 1.2$ g/dL, and the mean of the general population is 11.8 g/dL. Apply the z-test (α = 0.05, two-sided) and calculate the p value.

5. Classify each situation as a type I error, a type II error or a correct decision: (a) a study concludes that a supplement lowers cholesterol, but the supplement has no effect at all; (b) a study with 15 patients per group finds no significant differences between a new antibiotic, which is in reality superior, and the standard one; (c) a large study finds no differences between two diets that in reality have the same effect.
6. A student performs 8 independent tests at α = 0.05, and all the null hypotheses are true. (a) What is the probability of obtaining at least one false positive result? (b) What individual threshold would the Bonferroni correction impose? (c) And for 4 tests?
7. Five comparisons gave the p values: 0.002; 0.009; 0.018; 0.035; 0.25. How many are significant at α = 0.05 without correction, with Bonferroni, with Holm and with Benjamini-Hochberg?
8. Using Table 5.6 and then Google Sheets, find the two-sided p value for: (a) t = 2.5 with df = 12; (b) t = 1.70 with df = 40 (also give the one-sided value); (c) t = -2.9 with df = 159.
9. A study reports for the difference in SBP between two treatments: mean difference 4 mmHg, 95% CI -1 to 9 mmHg. (a) Is p smaller or larger than 0.05? (b) The minimal clinically important difference is 5 mmHg. Is the result an "absent effect" or an "undetermined effect"? (c) What would you recommend for a future study?
10. For BMI at enrollment, $m = \frac{29.04\ \text{kg}}{m^2}$, s = 3.91, n = 160, skewness 1.42, Shapiro-Wilk p < 0.001. (a) Calculate t for H0: $\mu = \frac{30\ \text{kg}}{m^2}$. (b) Which test would you prefer, and why? (c) How would you report the result?
11. How many patients per group are needed to detect, with 80% power and α = 0.05 two-sided, a difference of 10 mmHg in SBP between two groups, if σ = 14 mmHg? Use the approximate formula from Worked Problem 5.6.

Common Pitfalls

- Statistical hypotheses are written with the population parameters (μ, π, ρ), not with the sample means (m1, m2).
- "H0 is accepted" is an incorrect phrase; a nonsignificant result means that the data do not allow H0 to be rejected.
- The p value is not the probability that H0 is true, nor the probability that the result is due to chance.
- "Extremely significant" (p < 0.001) describes the strength of the evidence against H0, not the size or clinical importance of the effect.
- Choosing a one-sided test after seeing the direction of the difference artificially halves the p value.
- An effect that is significant in one group and nonsignificant in another does not prove that the effects differ between groups; compare the groups directly.
- With many tests, a few "significant" results appear by chance alone; without a correction or without a primary outcome specified in advance, they remain hypotheses.
- Write p < 0.001, never p = 0.000.

Key Points

- A statistical test starts from H0 (no effect), calculates how unusual the data would be if H0 were true, and rejects H0 when p < α.
- α = probability of a type I error (false positive); β = probability of a type II error (false negative); 1 - β = statistical power; 1 - α = confidence level.
- Power increases with effect size, with sample size and with α, and decreases with variability; when planning a study, a power of at least 80% is usually sought.

- The critical value, the p value and the confidence interval lead to the same decision; the confidence interval provides the most information.
- The t-test for a mean: $t = \frac{m-\mu_0}{\frac{s}{\sqrt{n}}}$, with n - 1 degrees of freedom; the z-test uses a known σ.
- A completely reported result contains the descriptive statistics, the test, the statistic with its degrees of freedom, the exact p, the effect size with its 95% CI, and the clinical interpretation.
- $1 - 0.95^k$ gives the risk of at least one false positive with k tests; Bonferroni (α/k), Holm and Benjamini-Hochberg are the usual corrections.

Self-Assessment Questions

1. (SC) The p value of a statistical test represents:
    a) the probability that the null hypothesis is true
    b) the probability that the alternative hypothesis is true
    c) the probability of obtaining a result at least as extreme as the one observed, if the null hypothesis is true
    d) the probability of making a type II error
    e) the standardized effect size
2. (SC) A study with 12 patients per group finds no significant difference between a new drug and placebo (p = 0.24), although the drug is in reality effective. The error committed is:
    a) a type I error
    b) a type II error
    c) a systematic sampling error
    d) a measurement error
    e) it is not an error, because p > 0.05
3. (SC) For mean SBP in mg2026 (m = 131.13 mmHg, 95% CI 128.96-133.30 mmHg), the two-sided test at α = 0.05 of the hypothesis H0: μ = 132 mmHg will result in:
    a) rejection of H0, because 132 ≠ 131.13
    b) non-rejection of H0, because 132 lies within the 95% CI
    c) acceptance of H0, because the population mean is certainly 132 mmHg
    d) rejection of H0, because n = 160 is a large sample
    e) the decision cannot be made without the p value
4. (SC) The power of a statistical test is:
    a) α
    b) 1 - α
    c) β
    d) 1 - β
    e) α + β
5. (SC) A researcher performs 10 independent tests at α = 0.05, and all the null hypotheses are true. The probability of obtaining at least one significant result is approximately:
    a) 0,05
    b) 0,10
    c) 0,23
    d) 0,40
    e) 0,50

6. (MC) Which statements about statistical hypotheses are correct?
    a) the null hypothesis usually states the absence of a difference or an association
    b) the hypotheses are formulated about the population parameters
    c) the two-sided alternative hypothesis for two means is written $\mu 1 \neq \mu 2$
    d) the hypotheses are formulated after examining the data, in order to choose the direction of the test
    e) H0 and H1 are mutually exclusive
7. (MC) The power of a t-test for two independent groups increases if:
    a) the number of patients in each group increases
    b) the true size of the effect increases
    c) the standard deviation of the variable increases
    d) we move from $\alpha = 0.01$ to $\alpha = 0.05$
    e) we move from a justified one-sided test to a two-sided one
8. (MC) In mg2026, SBP does not differ significantly between the sexes: difference -0.5 mmHg, 95% CI -4.9 to 3.8, p = 0.81. Which statements are correct?
    a) H0 is accepted, so mean SBP is identical in women and men
    b) H0 is not rejected
    c) the confidence interval rules out mean differences larger than about 5 mmHg
    d) the result demonstrates that sex does not influence SBP in any population
    e) the effect size (d = -0.04) is practically zero
9. (MC) Which of the following are examples of practices that increase the real risk of false positive results?
    a) adding patients until p drops below 0.05
    b) choosing a one-sided test after seeing the direction of the difference
    c) preregistering the protocol with the primary outcome
    d) reporting only the subgroups in which the effect was significant
    e) applying the Holm correction for multiple comparisons
10. (MC) Which statements about statistical and clinical significance are correct?

a) in a very large sample, a clinically unimportant difference can be extremely significant statistically
b) a correlation r = -0.088 with p = 0.01 in 867 patients explains less than 1% of the variability
c) a $p < 0.001$ guarantees that the effect is important for the patient
d) the minimal clinically important difference is established on clinical criteria, before the analysis
e) the effect size and the confidence interval complement the p value in reporting

Chapter 6. Common Parametric Statistical Tests: The t-Test for Paired and Independent Samples, Analysis of Variance (ANOVA)

Objectives

By the end of this chapter, the student will be able to:

- recognize the design of a study (one mean against a given value, two measurements on the same patients, two independent groups, three or more groups) and choose the appropriate parametric test;
- apply and calculate by hand the paired-samples t-test, including the confidence interval for the mean difference and the effect size dz;
- apply the independent-samples t-test in its Student version (pooled variance) and its Welch version, interpret Levene's test and calculate Cohen's d;
- explain the logic of analysis of variance (between-group variation versus within-group variation) and build the ANOVA table by hand (SS, df, MS, F);
- check the assumptions (independence, normality, homogeneity of variances) and choose the alternative when they are not met (Welch, Games-Howell, nonparametric tests);
- choose and interpret post hoc tests (Tukey, Bonferroni, Games-Howell, Dunnett) and explain why multiple t-tests are not performed;
- describe repeated-measures ANOVA (sphericity, Greenhouse-Geisser correction), the idea of two-way ANOVA and the role of ANCOVA in adjusting for the baseline value;
- critically read the results of t-tests and ANOVA in published articles and report their own results correctly.

6.1 The Family of Parametric Tests

The tests in this chapter are called parametric because they make assumptions about the shape of the population distribution (usually the normal distribution) and compare its parameters, the means. In exchange for these assumptions, they use all the information in the data and, when the assumptions are met, have the greatest power. The nonparametric tests in Chapter 7 (Mann-Whitney, Wilcoxon, Kruskal-Wallis, Friedman) make fewer assumptions and work with the ranks of the observations; they are the natural alternative when the distribution is strongly skewed, the sample is small or the variable is ordinal [19,63].

The choice of parametric test depends first of all on the study design. Table 6.1 summarizes the situations we will study; the complete algorithm for choosing a test, including for qualitative variables, appears in Chapter 8.

Table 6.1. Parametric tests for comparing means and their nonparametric alternatives.

Situation	Example in mg2026	Parametric test	Nonparametric alternative
one mean against a given value	mean SBP against 125 mmHg	z-test or t-test for a single mean	Wilcoxon for a median
two measurements on the same patients	blood glucose at 0 and at 6 months	paired-samples t-test	Wilcoxon signed-rank
two independent groups	SBP in women and in men	independent-samples t-test (Student or Welch)	Mann-Whitney U

Situation	Example in mg2026	Parametric test	Nonparametric alternative
three or more independent groups	decrease in blood glucose in arms A, B, C	one-way ANOVA	Kruskal-Wallis
three or more measurements on the same patients	blood glucose at 0, 3 and 6 months	repeated-measures ANOVA	Friedman

All the tests in the table share the same structure, the one from Chapter 5: a null hypothesis of equal means, a test statistic that relates signal to noise, a known distribution of the statistic under H0 (t or F), a p value and a decision. They differ only in how the signal and the noise are measured.

6.2 Comparing a Mean with a Given Value: The z-Test and the t-Test

The first situation in the table was covered at length in Chapter 5. We revisit it briefly here, with the rule from the course. The z-test is used if the sample size is n ≥ 30 and the population variance is known; the statistic is $z = \frac{m-\mu_0}{\frac{\sigma}{\sqrt{n}}}$, and the critical values come from the standard normal distribution (1.96 for α = 0.05, two-sided). The t-test is used if n < 30 or the population variance is unknown; the statistic is $t = \frac{m-\mu_0}{\frac{s}{\sqrt{n}}}$, with n - 1 degrees of freedom. Both tests assume an approximately normal distribution of the variable in the population, a condition that becomes less important as n increases.

In today's practice, σ is rarely known, and software uses the t-test regardless of n. For mean SBP in mg2026 against 125 mmHg we obtained t = 5.59 with 159 degrees of freedom, p < 0.001; for fasting plasma glucose against 100 mg/dL, t = 4.36, p < 0.001 (Chapter 5). What interests us now is that the paired-samples t-test, the first new test of the chapter, reduces exactly to this single-mean test.

6.3 The Paired-Samples t-Test

6.3.1 When the Data Are Paired

Two series of values are paired (dependent) when each value in one series has a natural, unique counterpart in the other. There are four typical clinical situations. The same patient measured before and after a treatment (blood glucose at enrollment and at 6 months). The same patient measured with two methods or two devices (blood pressure measured with an automatic device and with a manual one). Two paired organs of the same patient (the right eye and the left eye). Matched pairs of different patients, chosen to be similar in age, sex and severity, one of whom receives the treatment and the other does not, or twins.

What do these situations have in common? Values from the same pair are correlated: a patient with high blood glucose at enrollment will usually also have relatively high blood glucose at 6 months. The paired test uses this correlation to our advantage, removing the differences between patients from the comparison and keeping only the change within each patient.

6.3.2 The Logic: The Test Works on Differences

The idea is simple and elegant. For each patient we calculate the difference between the two measurements, di = xi,before - xi,after. The two series become a single series of n differences. If the treatment has no effect, the mean of the differences in the population is zero. The paired test is therefore the single-mean t-test applied to the differences, with the reference value μ0 = 0:

H0: μD = 0 (the mean of the differences in the population is zero; there is no change between the two time points)

H1: μD ≠ 0 (two-sided)

Test statistic:

$$t = \frac{m_D}{\text{SED}} = \frac{m_D}{\frac{s_D}{\sqrt{n}}}, \text{ with df = n - 1}$$

where mD is the mean of the differences, sD the standard deviation of the differences, and n the number of pairs (not the total number of measurements). The 95% confidence interval for the mean difference is $m_D \pm t$ critical × SED, and the effect size for paired data is dz $= \frac{m_D}{s_D}$.

The assumptions are: the pairs are independent of each other (different patients do not influence one another), and the differences have an approximately normal distribution. Pay attention to this detail, which often comes up in exams: normality is checked on the differences, not on each series separately. Blood glucose at enrollment in mg2026 is not perfectly normal (Shapiro-Wilk p = 0.037), but the differences glicemie0 - glicemie6 are (p = 0.725), and the paired test is fully justified. If the differences are strongly skewed or have extreme values, especially with small n, the Wilcoxon signed-rank test is used (Chapter 7).

6.3.3 An Example Calculated by Hand

Ten hypertensive patients from a family practice have their SBP measured before and after four weeks of treatment with an angiotensin-converting enzyme inhibitor. The data and calculation are in Table 6.2; the full solution, with all the steps of the a-g format, is Worked Problem 6.1.

Table 6.2. SBP (mmHg) before and after 4 weeks of treatment in 10 patients (fictitious data).

Patient	Before	After	Difference 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 / mean	152,4	144,6	78 / 7,8	0	151,60

The mean of the differences is $m_D = \frac{78}{10} = 7.8$ mmHg. The standard deviation of the differences: $s_D = \sqrt{\frac{151.60}{9}} = \sqrt{16.84} = 4.10$ mmHg. The standard error: SED $= \frac{4.10}{\sqrt{10}} = \frac{4.10}{3.162} =$ 1.30 mmHg. The test statistic: $t = \frac{7.8}{1.30} = 6.01$, with 9 degrees of freedom. The critical value for df = 9 and α = 0.05, two-sided, is 2.262 (Table 5.6), and for α = 0.01 it is 3.250; t = 6.01 exceeds both. The exact p value is 0.0002. 95% CI for the mean decrease: 7.8 ± 2.262 × 1.30 = 7.8 ± 2.94, that is, 4.86-10.74 mmHg. Effect size dz $= \frac{7.8}{4.10} = 1.90$, a very large effect.

What would have happened if we had ignored the pairing and compared the two columns as if they came from different patients? The independent-samples t-test would have given t = 2.96

and p = 0.008: still significant, but with a statistic half as large. The reason is that, in the independent analysis, the noise includes the large differences between patients (baseline SBP ranges from 142 to 165 mmHg, SD 7.11), whereas in the paired analysis the noise is only the variability of the response to treatment (SD of the differences 4.10). On real data, with smaller effects, ignoring the pairing often turns a significant result into a nonsignificant one. The opposite mistake, treating independent groups as paired, is even more serious, because it produces meaningless results: pairing must exist in the design, not be invented at the analysis stage.

6.3.4 The Paired Test on the mg2026 Dataset

In mg2026, fasting plasma glucose was measured at enrollment and at 6 months in all 160 patients. Figure 6.1 shows, in panel A, the individual trajectories of 25 patients (each line joins the values of the same patient) and the trajectory of the mean; in panel B, the histogram of the 160 individual differences.

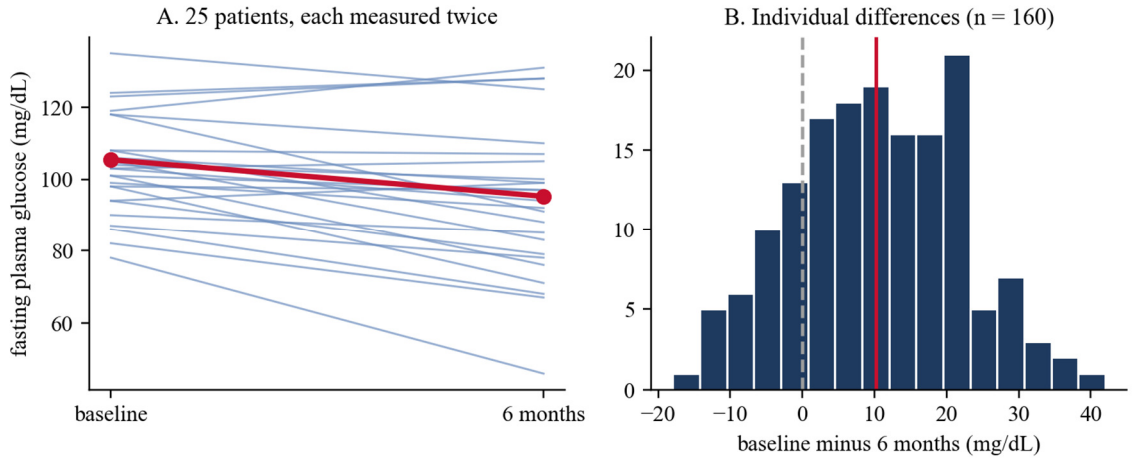


Figure 6.1. Fasting plasma glucose at enrollment and at 6 months. A. Trajectories of 25 patients (thin lines) and the group mean (thick line). B. Histogram of the individual differences (baseline minus 6 months) in the 160 patients; the dotted line marks zero difference, and the solid line the mean of the differences (10.19 mg/dL). Synthetic data (mg2026 dataset).

The graph already tells the story before any calculation. Most lines go down, but not all: 125 patients have lower blood glucose at 6 months, 34 have higher values, and one has the same value. The histogram of the differences is approximately symmetric, centered around 10 mg/dL, with values between -18 and +42 mg/dL.

Numerical results: mean at enrollment 105.4 mg/dL, at 6 months 95.2 mg/dL; mean difference 10.19 mg/dL, with a standard deviation of the differences of 11.57 mg/dL. SED $= \frac{11.57}{\sqrt{160}} = \frac{11.57}{12.649} = 0.915$ mg/dL. $t = \frac{10.19}{0.915} = 11.14$, with 159 degrees of freedom, $p < 0.001$. 95% CI for the mean difference: 10.19 ± 1.975 × 0.915 = 10.19 ± 1.81, that is, 8.38-11.99 mg/dL. Effect size $dz = \frac{10.19}{11.57} = 0.88$. The Shapiro-Wilk test on the differences gives p = 0.725, so the normality assumption is met; the Wilcoxon test, for comparison, also gives $p < 0.001$.

The interpretation has two levels. Statistically, the decrease in blood glucose is highly significant and large in effect. Clinically, we must remember that the paired analysis of all patients has no control group: it describes the course of the cohort, not the effect of a particular intervention. The question of the effectiveness of the app and of teleconsultation is answered only by comparing the arms (Section 6.5). The same logic applies to shorter periods: between enrollment and 3 months, blood glucose decreases on average by 6.08 mg/dL(CI 95% 4.61 − 7.54; $t = 8.22$; $p < 0.001$; dz = 0.65), and between 3 and 6 months by a further 4.11 mg/dL(CI 95% 2.10 − 6.13; $t = 4.03$; $p < 0.001$; dz = 0.32).

A useful example for the discussion of assumptions arises if we look only at arm A (standard counseling, n = 53) and body weight. The mean weight loss between enrollment and 6 months is 0.71 kg (SD 2.48 kg). The paired t-test gives t = 2.07, df = 52, p = 0.043, a borderline significant result; the Wilcoxon test on the same data gives p = 0.052, borderline nonsignificant. The differences have a distribution compatible with normality (Shapiro-Wilk p = 0.696), so the t-test is the justified choice. The example shows how fragile a result close to the threshold is: two reasonable tests give opposite decisions, and a loss of 0.7 kg in 6 months is, in any case, clinically unimportant. The test is chosen according to the assumptions, checked beforehand, not according to the p value that suits us.

Case study: paired data in pharmacology and nutrition

A before-after study of hypertensive patients with endothelial dysfunction followed the serum level of thrombospondin-1 at enrollment (T0) and after one year of treatment with perindopril (T1): 8061.9 ± 3684.80 ng/mL at T0 versus 15380 ± 5887 ng/mL at T1, $p < 0.001$, alongside a control group of non-hypertensive individuals (9221.03 ± 6510.21 ng/mL) [86]. The second marker reported decreased from 0.91 ± 0.51 to 0.50 ± 0.24 ng/mL($p < 0.001$). The design is classic for the paired test (or for Wilcoxon, if the distributions are skewed; the large standard deviations relative to the means suggest this possibility). The critical reading exercise is this: the paired test cannot be reconstructed from the mean and SD at each time point, because the test uses the standard deviation of the individual differences, which depends on the correlation between the two measurements. A complete report would include the mean difference (or the median of the differences) with its confidence interval. An example of this kind of reporting comes from the prospective study of 921 students followed during their first year of university: the mean change in BMI was -0.45 ± 1.07 kg/m² (95% CI -0.52 to -0.38; p < 0.001), and 48.2% of participants had a decrease in BMI, 38.7% remained stable and 13.1% increased [5]. We can check the interval: SED = $\frac{1.07}{\sqrt{921}} = 0.035$, and -0.45 ± 1.96 × 0.035 gives -0.52 to -0.38, exactly the reported values. Note also the clinical relevance: a mean decrease of 0.45 kg/m² is highly significant at n = 921, but small for an individual; the proportions of students who gain or lose weight say more than the mean. A third example, the retrospective study of SGLT-2 inhibitors in 200 patients with type 2 diabetes, reports the before-after changes as medians with IQR (for example, blood glucose from 180.00 to 130.00 mg%, p < 0.001), which suggests the use of the Wilcoxon test, discussed in the next chapter [59].

6.4 The Independent-Samples t-Test

6.4.1 Design and Hypotheses

Two samples are independent when they come from different patients and there is no correspondence between the values of one group and those of the other: women and men, smokers and nonsmokers, patients treated with drug A and those treated with drug B in a parallel-group randomized trial. The groups may have different sizes. The question is whether the means of the two populations differ:

H0: μ1 = μ2 (or μ1 - μ2 = 0)

H1: μ1 ≠ μ2 (two-sided)

6.4.2 The Student Version, with Pooled Variance

The signal is the difference between the sample means, m1 - m2. The noise is the standard error of this difference. In the classic version, Student assumes that the two populations have the same variance and estimates it with a pooled variance, a weighted average of the two variances:

$$s^2_{\text{pooled}} = \frac{(n_1 - 1) \times {s_1}^2 + (n_2 - 1) \times {s_2}^2}{n_1 + n_2 - 2}$$

$$\text{SE}(m_1 - m_2) = \sqrt{\left[s^2_{\text{pooled}} \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]}$$

$t = \frac{m_1 - m_2}{\text{SE}(m_1 - m_2)}$, with df = n1 + n2 - 2

The 95% confidence interval for the difference in means is $(m_1 - m_2) \pm t$ critical × SE(m1 - m2), and the effect size is Cohen's d = (m1 - m2) / spooled.

6.4.3 The Welch Version and Levene's Test

If the variances of the two populations differ, especially when the group sizes are also unequal, the Student version may give wrong p values. The Welch version no longer assumes equal variances: each group keeps its own variance in the calculation of the standard error, and the degrees of freedom are adjusted (the Welch-Satterthwaite formula usually gives a number with decimals):

$$\text{SE}(m_1 - m_2) = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

$$\text{df} = \frac{\left(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}\right)^2}{\frac{\left(\frac{{s_1}^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{{s_2}^2}{n_2}\right)^2}{n_2 - 1}}$$

To decide between the two versions, software reports Levene's test for equality of variances, with H0: $\sigma 1^2 = \sigma 2^2$. If Levene's $p \geq 0.05$, we have no reason to reject equality of variances and the Student version can be used; if Levene's $p < 0.05$, the Welch version is used. A quick rule of thumb: if the larger standard deviation is more than twice the smaller one, a pooled variance is no longer a reasonable assumption.

More and more statisticians recommend using the Welch test by default, without a preliminary Levene's test: when the variances are equal, Welch gives practically the same result as Student, and when they are not, it protects the researcher from an error. JASP offers both versions simply by ticking the Student and Welch options; in Google Sheets, the T.TEST function uses type 2 for pooled variance and type 3 for unequal variances. In this course, report Levene's test and choose the version according to it, but remember that choosing Welch is never a serious mistake.

6.4.4 Groups of Unequal Size

The t-test does not require equal groups; in mg2026 we have 86 women and 74 men, and in observational studies the imbalances can be much larger (for example, 224 patients with type 2 diabetes versus 62 with type 1 diabetes [8]). Three things should nevertheless be kept in mind. First: for the same total number of patients, power is greatest when the groups are equal; a study with 20 + 140 patients has much less power than one with 80 + 80. Second: the combination of unequal groups and unequal variances is the situation in which the Student version goes most wrong; if the small group has the larger variance, the Student test becomes too liberal (too many false positive results), and if the small group has the smaller variance, it becomes too conservative. Exercise 3 in this chapter shows such a situation, in which Student and Welch lead

to opposite decisions. Third: large differences in size between groups raise questions about selection, which no test can resolve.

6.4.5 The Independent Test on the mg2026 Dataset

The nonsignificant result used in Chapter 5 comes from an independent test: SBP in women 131.37 ± 14.47 mmHg$(n = 86)$ and in men 130.85 ± 13.25 mmHg$(n = 74)$. The hand calculation with pooled variance: $s^2_{\text{pooled}} = \frac{85\times14.47^2+73\times13.25^2}{158} = \frac{85\times209.38+73\times175.56}{158} = \frac{17797.4+12816.0}{158} = 193.76$, so spooled = 13.92 mmHg. SE $= \sqrt{\left[193.76 \times \left(\frac{1}{86}+\frac{1}{74}\right)\right]} = \sqrt{193.76 \times 0.02514} = \sqrt{4.871} = 2.207$ mmHg. $t = \frac{130.85-131.37}{2.207} = -\frac{0.52}{2.207} = -0.24$, df = 158, p = 0.81. Levene's test gives p = 0.529, so the variances can be considered equal; the Welch version gives practically the same result (t = -0.24, df = 157.4, p = 0.81). The difference men minus women is -0.5 mmHg (95% CI -4.9 to 3.8), and d = -0.04. The Mann-Whitney test, for comparison, gives p = 0.80. H0 is not rejected; as we discussed in Chapter 5, the confidence interval excludes mean differences larger than about 5 mmHg.

Fasting plasma glucose by sex gives a similar result: men have a mean of 106.2 mg/dL(SD 15.19; $n = 74$), women 104.8 mg/dL(SD 16.30; $n = 86$). Levene's test p = 0.426; Welch t-test t = 0.54, df = 157.0, p = 0.589; difference men minus women 1.35 mg/dL (95% CI -3.57 to 6.27), d = 0.09. Not significant. To see a significant result on the same kind of comparison as well, look at height: women 162.6 ± 6.06 cm, men 175.97 ± 6.18 cm; Levene p = 0.890; Student t-test t = 13.79, df = 158, p < 0.001; difference 13.37 cm (95% CI 11.45-15.28), d = 2.19. Here the difference is obvious even without statistics, but the test quantifies it.

The literature produced by our colleagues contains numerous comparisons of two means. In the study of patients with ischemic heart failure with reduced ejection fraction, compared with 60 healthy controls of the same age, osteoprotegerin was much higher in patients (4.7 ± 0.25 versus 1.3 ± 0.67 ng/mL, $p < 0.001$), and vitamin D3 much lower (20.49 ± 7.31 versus 37.09 ± 4.59 ng/mL, p < 0.001) [87]. In the diagnostic study using contrast-enhanced ultrasound, the CEUS-PAT parameter was 0.83 ± 0.09 in the 37 patients with cirrhosis and 0.49 ± 0.11 in the 27 healthy volunteers (p < 0.0001) [2]; if we calculate d from the published values, we obtain about 3.4, an almost complete separation of the two groups. Such huge differences are typical of case-control designs with perfectly healthy controls, which exaggerate the discriminating ability of a test compared with the real situation, in which the physician has to distinguish cirrhosis from other liver diseases (spectrum effect, Chapter 11). In the cross-sectional study of disorders of gut-brain interaction, BMI was 28.15 ± 6.49 kg/m² in the 52 cases and 24.47 ± 4.60 kg/m² in the 84 controls (p = 0.001) [71]; here the groups are unequal and the standard deviations differ (ratio 1.4), a situation in which the Welch version is the prudent choice.

> **Case study: when the standard deviation approaches the mean**
>
> A diagnostic study carried out in the emergency department on 85 patients with a stroke alert (69 with stroke, 16 without stroke) compared two serum biomarkers [61]. Glial fibrillary acidic protein (GFAP) was 142.91 ± 102.19 pg/mL in patients with stroke and 37.76 ± 19.92 pg/mL in those without stroke (p < 0.001); UCH-L1 was 1307.68 ± 967.54 versus 189.81 ± 92.69 pg/mL$(p < 0.001)$. The differences are large and clinically interesting, and for distinguishing hemorrhagic from ischemic stroke GFAP reached an accuracy of 94.2% at a cutoff of 77.15 pg/mL. For a first-year student, the abstract is an excellent exercise on the assumptions of the t-test. The first warning sign: the standard deviation is almost as large as the mean (102.19 versus 142.91; 967.54 versus 1307.68). A variable that cannot be negative and has an SD close to its mean is almost certainly right-skewed; the mean $\pm$2SDar includes

> negative values, which are biologically impossible. The second warning sign: the variances differ enormously between groups (the ratio of the standard deviations is about 5 for GFAP and 10 for UCH-L1), and the groups are very unequal (69 versus 16). Under these conditions, the Student t-test with pooled variance is not appropriate; the correct options are the Welch test on log-transformed data or the Mann-Whitney test, and a description using the median and IQR would represent the distributions more faithfully. The clinical conclusion of the study (the biomarkers are much higher in stroke) does not change, because the differences are very large; with smaller differences, the choice of test could change the decision. The abstract does not name the test used, so the exercise is not a criticism but a list of things to check before choosing the test for your own data.

6.5 One-Way Analysis of Variance (ANOVA)

6.5.1 Why We Do Not Perform Multiple t-Tests

In mg2026 we have three treatment arms. The question "does the decrease in blood glucose differ between arms?" could be approached with three t-tests: A versus B, A versus C, B versus C. The problem is the one from Chapter 5: each test has a 5% risk of a false positive, and the risk that at least one of the three is a false positive, if there is no real difference, rises to about $1 - 0.95^3 = 0.143$ (the tests are not perfectly independent, but the order of magnitude is correct). For 4 groups there are 6 comparisons, for 5 groups 10, for k groups $k(k - 1)/2$, and the overall risk increases rapidly. Analysis of variance (ANOVA), developed by R. A. Fisher in the 1920s, answers the global question with a single test, keeping the type I error risk at α [42,53].

H0: $\mu 1 = \mu 2 = \ldots = \mu k$ (all population means are equal)

H1: at least one mean differs from the others

Pay attention to the wording of H1: ANOVA does not tell us which mean differs. For that, post hoc tests are needed (Section 6.6).

6.5.2 The Logic: Between-Group Variation Versus Within-Group Variation

The name "analysis of variance" seems odd for a test that compares means. It comes from the way the comparison is made. The total variability of the data (how much each value differs from the grand mean) is partitioned into two parts. Between-group variation shows how much the group means differ from the grand mean; if H0 is true, it is small, produced by chance alone. Within-group variation shows how much each value differs from the mean of its own group; it reflects biological and measurement variability and does not depend on H0. The ratio of the two, after each has been divided by its degrees of freedom, is the F statistic:

F = MS between groups / MS within groups = (variation explained by group) / (unexplained variation)

If H0 is true, both MS estimate the same variance, and F is around 1. If the means differ, MS between groups increases, and F becomes much larger than 1. The distribution of F under H0 is Fisher's F distribution, with two numbers of degrees of freedom: k - 1 in the numerator and N - k in the denominator (N being the total number of observations).

Figure 6.2 shows the logic on a small example: the pain score (0-10) in 15 patients treated with three analgesics, 5 patients per group. The arrows show the distance between each group mean and the grand mean (between-group variation); the thin vertical lines show the distance between each patient and the mean of their group (within-group variation). The means are well separated, and the values in each group are tightly clustered around the mean: F will be large.

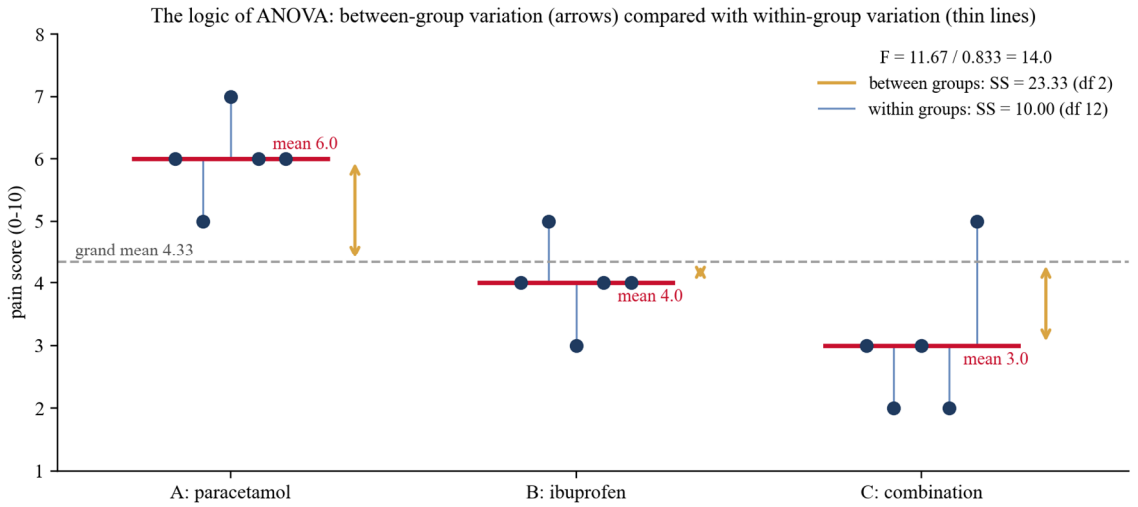


Figure 6.2. The logic of ANOVA in an example with three analgesics (A paracetamol, B ibuprofen, C combination; 5 patients each, pain score 0-10). The red horizontal lines are the group means, the dotted line is the grand mean (4.33). The arrows represent between-group variation, the thin lines represent within-group variation. Fictitious data.

6.5.3 Sums of Squares, Degrees of Freedom and the ANOVA Table

The partition is done with sums of squares (SS). We denote by xij the value of patient j in group i, by mi the mean of group i, by ni the number of patients in group i, by M the grand mean and by N the total number of patients.

SS total $= \sum(x_{ij} - M)^2$, with df = N - 1

SS between groups $= \sum n_i \times (m_i - M)^2$, with df = k - 1

SS within groups (residual) = $\sum(x_{ij} - m_i)^2 = \sum(n_i - 1) \times {s_i}^2$, with df = N - k

The fundamental property is additivity: SS total = SS between + SS within, and df total = df between + df within. The mean squares (MS) are obtained by dividing each SS by its degrees of freedom: MS between = SS between / (k - 1), MS within = SS within / (N - k). Then F = MS between / MS within, and the p value is the area under the F distribution to the right of the calculated value. All these quantities are arranged in a standard table, which all software produces.

Table 6.3. Structure of the one-way ANOVA table.

Source of variation	SS	df	MS	F	p
between groups	$\sum n_i(m_i - M)^2$	k - 1	SS between / (k - 1)	MS between / MS within	from the F distribution
within groups	$\sum(n_i - 1)\,{s_i}^2$	N - k	SS within / (N - k)		
total	$\sum(x_{ij} - M)^2$	N - 1			

The formula SS within = $\sum(n_i - 1) \times {s_i}^2$ is very useful: it allows the ANOVA table to be built only from the n, the mean and the standard deviation of each group, that is, from the figures reported in any article. We will use it on the mg2026 data in Worked Problem 6.6.

One measure of effect size in ANOVA is eta squared: η^2 = SS between / SS total, the proportion of total variability explained by group membership. Cohen's rough benchmarks are 0.01 (small), 0.06 (medium) and 0.14 (large) [76].

6.5.4 The ANOVA Table for the Analgesics Example

For the data in Figure 6.2 (A: 6, 5, 7, 6, 6; B: 4, 5, 3, 4, 4; C: 3, 2, 3, 2, 5), the group means are 6, 4 and 3, and the grand mean is 65/15 = 4.33. SS between = 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 within is obtained by adding, in each group, the squared deviations from the group mean: in group A, 0 + 1 + 1 + 0 + 0 = 2; in group B, 0 + 1 + 1 + 0 + 0 = 2; in group C, 0 + 1 + 0 + 1 + 4 = 6; total 10.00. SS total = 33.33 (check: 23.33 + 10.00).

Table 6.4. ANOVA table for the pain score with three analgesics (fictitious data).

Source of variation	SS	df	MS	F	p
between groups	23,33	2	11,67	14,00	0,0007
within groups	10,00	12	0,833		
total	33,33	14			

The critical value F(2, 12) for α = 0.05 is 3.89, and for α = 0.01 it is 6.93; F = 14.00 exceeds both, and p = 0.0007. We reject H0: at least one analgesic differs from the others in terms of mean pain score. $\eta^2 = \frac{23.33}{33.33} = 0.70$: 70% of the variability in the score is explained by the type of analgesic, a very large effect (naturally, for an example built to be clear). Which analgesics differ from each other we find out with the Tukey test, in Worked Problem 6.3.

6.5.5 Assumptions

One-way ANOVA has three assumptions, similar to those of the independent t-test.

Independence of observations. Each patient appears only once, in a single group, and the patients do not influence one another. This is the most important assumption and it is ensured by the design; no statistical test can check or repair it. Repeated measurements on the same patients require repeated-measures ANOVA (Section 6.7).

Normality in each group (or, equivalently, of the residuals, that is, of the deviations from the group mean). It is checked with Shapiro-Wilk in each group or with the Q-Q plot of the residuals. With large groups of similar size, ANOVA is robust to moderate departures.

Homogeneity of variances (homoscedasticity). It is checked with Levene's test (H0: the variances are equal) or with Bartlett's test, which is more sensitive to departures from normality. Rule of thumb: classic ANOVA remains acceptable if the ratio of the largest to the smallest standard deviation is below 2 and the groups are of similar size.

When the assumptions are not met, we have alternatives. For unequal variances: Welch's ANOVA (F with adjusted degrees of freedom) and the Games-Howell post hoc test. For strongly skewed distributions or ordinal data: the Kruskal-Wallis test (Chapter 7) or a logarithmic transformation.

6.5.6 ANOVA on the mg2026 Dataset: Decrease in Blood Glucose by Arm

We now come to the main question of the mg2026 study. The variable analyzed is the decrease in fasting plasma glucose at 6 months, calculated for each patient as glicemie0 - glicemie6 (positive values mean a decrease). Figure 6.3 shows its distribution in the three arms.

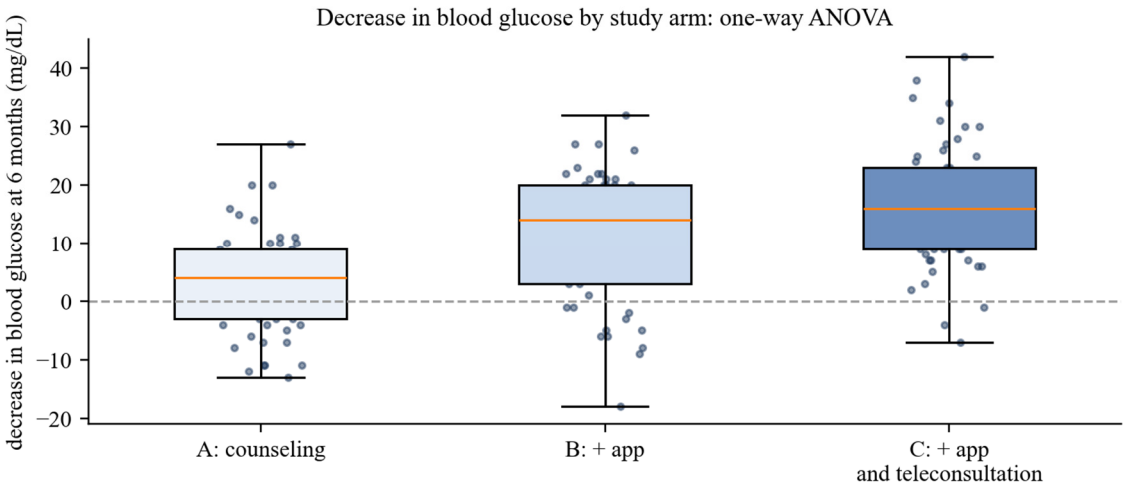


Figure 6.3. Decrease in fasting plasma glucose at 6 months (glicemie0 - glicemie6) by study arm: A standard counseling, B counseling + app, C counseling + app + teleconsultation. Box plot with individual values overlaid; the dotted line marks zero decrease. Synthetic data (mg2026 dataset).

The graph suggests that the decrease in blood glucose grows from arm A to arm C, with similar spreads. Descriptive statistics: A, mean 3.0 mg/dL(SD 8.8; $n = 53$); B, 11.1 mg/dL(SD 11.2; $n = 53$); C, 16.3 mg/dL(SD 10.6; $n = 54$); grand mean 10.19 mg/dL.

Checking the assumptions. The patients are independent and each appears in a single arm (individual randomization). The Shapiro-Wilk test in each arm: A p = 0.575, B p = 0.052, C p = 0.903; normality is acceptable, with arm B borderline, and the groups are large and equal. The ratio of the largest to the smallest SD is 11.2 / 8.8 = 1.27, below 2. Levene's test gives p = 0.163 in the median-centered version (Brown-Forsythe) and p = 0.089 in the mean-centered version; Bartlett's test, p = 0.192. We have no reason to reject homogeneity of variances.

Table 6.5. ANOVA table for the decrease in blood glucose at 6 months by arm (mg2026).

Source of variation	SS	df	MS	F	p
between arms	4806,9	2	2403,4	22,93	< 0,001
within arms	16459,5	157	104,8		
total	21266,4	159			

F(2, 157) = 22.93, with p < 0.001 (the critical value for α = 0.05 is 3.05). $\eta^2 = \frac{4806.9}{21266.4} = 0.226$: almost 23% of the variability in the decrease in blood glucose is explained by the treatment arm, a large effect. Welch's ANOVA, which does not assume equal variances, confirms this: F(2, 103.5) = 26.05, p < 0.001. The nonparametric Kruskal-Wallis test gives H = 36.63, p < 0.001.

The statistical conclusion: H0 is rejected; the mean decrease in blood glucose differs highly significantly between at least two arms. The research conclusion requires post hoc tests, to see which arms differ.

6.5.7 The Second Outcome: Weight Loss

The same analysis for weight loss at 6 months (greutate0 - greutate6, kg) gives: A 0.71 ± 2.48 kg, B 2.84 ± 2.36 kg, C 3.37 ± 2.39 kg. The standard deviations are almost identical (Levene p = 0.79), and the distributions are compatible with normality in all arms (Shapiro-Wilk p between 0.17 and 0.70). ANOVA: SS between = 211.9 (df 2), SS within = 911.3 (df 157), MS between = 105.9, MS within = 5.80, F(2, 157) = 18.25, p < 0.001, $\eta^2 = 0.189$. The arms with the app lost more weight than the standard counseling arm; whether teleconsultation also brings an additional benefit over the app we will again see from the post hoc tests.

6.6 Post Hoc Tests

6.6.1 What Problem They Solve

A significant F tells us only that the means are not all equal. To find out which pairs differ, follow-up comparisons are performed, called post hoc tests (Latin "after this"). They compare pairs of groups, but adjust the threshold or the p values so that the overall type I error risk stays at α, exactly the idea of the corrections for multiple comparisons in Chapter 5. There are many post hoc tests; in clinical medicine, four are used.

Tukey's HSD (honestly significant difference) test compares all possible pairs and is the standard choice when the variances are homogeneous and the groups are of similar size. It uses the "studentized range" distribution q and calculates the minimum significant difference:

$$HSD = q(k; N - k; \alpha) \times \sqrt{(MS\ within / n)}$$

where n is the number of patients in a group (for unequal groups, software uses the Tukey-Kramer version). Two means differ significantly if their difference exceeds HSD.

The Bonferroni correction applied to pairwise t-tests multiplies each p value by the number of comparisons. It is simple, but conservative when there are many comparisons; the Holm version is preferable.

The Games-Howell test assumes neither equal variances nor equal groups. It is the appropriate choice when Levene's test is significant or when the standard deviations differ greatly, usually together with Welch's ANOVA.

Dunnett's test compares each treatment group only with a control group, not all pairs with one another. Because there are fewer comparisons, it has more power. It is appropriate when the question is "which treatments differ from the control?" and the comparisons between treatments are of no interest.

6.6.2 The Tukey Test on the mg2026 Dataset

For the decrease in blood glucose by arm, the Tukey test gives the results in Table 6.6.

Table 6.6. Post hoc comparisons for the decrease in blood glucose at 6 months (mg2026).

Comparison	Mean difference (mg/dL)	Tukey 95% CI	Tukey p	Games-Howell p	Bonferroni p
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

All three pairs differ significantly. The app (arm B) brings an additional decrease of 8.1 mg/dL compared with standard counseling, and monthly teleconsultation (arm C) a further 5.2 mg/dL compared with the app. The C - B difference is the smallest and the closest to the threshold: it remains significant with all three methods, but its confidence interval starts at 0.6 mg/dL, so the additional benefit of teleconsultation could also be very small. Note how the more conservative methods (Games-Howell, Bonferroni) give slightly larger p values for this pair; if Tukey's p had been 0.045, the choice of method could have changed the decision, which is why it is set before the analysis.

For weight loss, Tukey shows that arm B (difference from A 2.13 kg, $p < 0.001$) and arm C (2.66 kg, $p < 0.001$) differ from A, but C does not differ significantly from B (0.53 kg, $p = 0.489$). Combining the two results, the app seems to bring most of the benefit on weight, and teleconsultation adds an additional benefit mainly on blood glucose. We phrase this cautiously, because the absence of a significant C - B difference in weight does not prove that the two arms are equal.

6.6.3 Other Analysis Options and What Not to Do

If the question had been only "are the app and teleconsultation better than standard counseling?", Dunnett's test, with arm A as the control, would have been the natural choice. What not to do: do not apply simple, uncorrected t-tests between all pairs; do not perform post hoc tests after a nonsignificant F, searching for a pair "that comes out" (except for comparisons planned in advance and justified in the protocol); do not interpret only the significant pairs, ignoring the size of the differences and the confidence intervals.

A published example of an analysis across several ordered groups comes from liver ultrasound. The UGAP parameter increased monotonically with the steatosis grade established by the reference method: 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 $270.8 \pm \frac{31.62 d_B}{m}$ ($p < 0.001$), with groups of 48, 56, 14 and 59 subjects [58]. The standard deviations are very close (ratio 1.2), so classic ANOVA is appropriate; group S2, however, is small, and comparisons involving it have less power. When the groups are naturally ordered (stages, grades), in addition to ANOVA one can also test for a linear trend, which directly answers the question "does the parameter increase with grade?". Two other studies by our colleagues compare three groups without naming the test in the abstract: a prospective study of 150 pregnant women divided into three equal groups of 50 found significant differences in arterial stiffness ($p < 0.001$), heart rate ($p = 0.006$), aortic systolic pressure ($p = 0.02$), brachial augmentation index ($p = 0.01$) and pulse pressure ($p = 0.049$) [88], and a study with three parallel groups of patients after stroke (22, 18 and 30 patients) reported better outcomes in the groups receiving combined therapy ($p < 0.05$) and improvements from baseline at 12 months ($p < 0.05$) [89]. In both cases, the reader would need the name of the test (ANOVA or Kruskal-Wallis), the post hoc tests used and the means with their confidence intervals to judge the size of the differences; when the same patients are assessed at several time points, an analysis for repeated measurements is also needed.

Case study: ANOVA, Tukey and eta² in two studies by our colleagues

The cross-sectional study of smart hearing aids compared self-esteem (SES scale) in 245 people divided into three groups: 73 with normal hearing, 86 hearing aid users and 86 patients with sensorineural hearing loss without a hearing aid [82]. One-way ANOVA gave $F = 299.00$, $p < 0.001$, with $\eta^2 = 0.712$, and the Tukey HSD post hoc test showed that hearing aid users had higher scores than those without a hearing aid (35.41 ± 5.32 versus 22.99 ± 4.53; $p < 0.001$; d Cohen = 2.515). An η^2 of 0.712 means that 71% of the variability in self-esteem is associated with the group, an exceptionally large effect in the behavioral sciences. A few critical reading observations: the standard deviations are close (5.32 and 4.53), so the homogeneity assumption seems to be met and Tukey is the appropriate test; the groups are of similar size; the design is cross-sectional, so the result shows a strong association, but does not establish that the hearing aid produced the increase in self-esteem (people who choose to wear a hearing aid may also differ in other characteristics). The second study compared left ventricular global longitudinal strain (GLS) in 148 patients from three groups: 62 with heart failure with preserved ejection fraction, 46 with hypertension and 40 age-matched controls [90]. The overall difference was significant ($p < 0.001$), and the pairwise comparisons showed that the controls (-20.2 ± 1.4%) differed from the hypertensive patients ($-18.4 \pm 3.0\%$, $p = 0.031$) and from the patients with heart failure ($-17.6 \pm 2.3\%$, $p < 0.001$). Here the standard deviation of the hypertensive patients is more than twice that of the controls; in a reanalysis, Welch's ANOVA with the Games-Howell test would be a useful check. Note also that the controls-hypertensive difference, although significant, has a p much closer to the threshold than the controls-heart failure difference, which corresponds to a smaller mean difference (1.8 versus 2.6 percentage points).

6.7 Repeated-Measures ANOVA

6.7.1 When the Same Patients Are Measured Several Times

In mg2026, blood glucose was measured at enrollment, at 3 months and at 6 months in the same patients. The three series are not independent, so one-way ANOVA is not appropriate; nor are three uncorrected paired t-tests (inflation of the type I error). The appropriate test is repeated-measures ANOVA, the generalization of the paired t-test to three or more time points. Like the paired test, it removes the constant differences between patients from the noise and compares only the changes within each patient.

H0: μ0 months = μ3 months = μ6 months (mean blood glucose does not change over time)

H1: the mean differs at one or more time points

6.7.2 Sphericity and the Greenhouse-Geisser Correction

In addition to normality, repeated-measures ANOVA has a specific assumption, sphericity: the variances of the differences between all pairs of time points (0-3, 0-6, 3-6) must be equal. It is checked with Mauchly's test (H0: sphericity holds). If Mauchly's test is significant, the degrees of freedom of the F test are multiplied by a correction factor ε (epsilon) smaller than one, most often estimated with the Greenhouse-Geisser method (more conservative) or the Huynh-Feldt method. The correction does not change the F value but the degrees of freedom, so the p value increases slightly. For two time points, sphericity is automatically satisfied, and repeated-measures ANOVA gives exactly the same result as the paired t-test ($F = t^2$).

6.7.3 The Result on mg2026

The mean blood glucose values are 105.4 / 99.4 / 95.2 mg/dL at 0 / 3 / 6 months. Repeated-measures ANOVA: SS time = 8405.5 (df 2), SS error = 20570.5 (df 318), MS time = 4202.8, MS error = 64.69, F(2, 318) = 64.97, $p < 0.001$. Effect size, partial eta² = 8405.5 / (8405.5 + 20570.5) = 0.29. Mauchly's test is significant (W = 0.873, $p < 0.001$), so sphericity does not hold; the Greenhouse-Geisser correction has ε = 0.887, the degrees of freedom become 1.77 and 282.2, and p remains < 0.001. Post hoc comparisons with Bonferroni correction show that all three pairs of time points differ (0-3 months, 0-6 months and 3-6 months, all $p < 0.001$). The equivalent nonparametric test, Friedman, gives $\chi^2 = 74.67$, $p < 0.001$ (Chapter 7).

The conclusion: mean blood glucose decreases significantly and progressively over the 6 months, with a larger decrease in the first 3 months (6.1 mg/dL) than in the next 3 (4.1 mg/dL). As with the paired test, the analysis of the whole cohort describes the change over time, not the effect of a particular intervention; the question "does the change over time differ between arms?" requires a mixed model, with the time factor (within patients) and the arm factor (between patients), whose time × arm interaction term answers exactly this question. JASP allows this model from the same menu (Repeated Measures ANOVA, with arm added to "Between Subject Factors").

> **Case study: repeated measurements in an uncontrolled pediatric pilot study**
>
> A prospective, single-arm pilot study followed 21 children with atopic dermatitis and laboratory-confirmed intestinal dysbiosis, treated with a personalized therapy targeting the gut microbiota, with assessments at enrollment, at 30 and at 90 days [91]. The POEM score decreased from 14.67 ± 5.22 at enrollment to 7.10 ± 3.82 at 90 days, and the SCORAD score from 41.66 ± 15.28 to 17.93 ± 11.68 (both $p < 0.001$); about 76% of the children achieved a reduction of at least 50% in the POEM score and 71% in the SCORAD score. The statistical analysis used, repeated-measures ANOVA, is the correct choice for three assessments of the same children; a complete report would include the sphericity test and, if applicable, the

> Greenhouse-Geisser correction, as well as post hoc comparisons between time points. The main lesson, however, concerns the design. In a single-arm study, the decrease in scores cannot be attributed with certainty to the treatment: atopic dermatitis has a fluctuating course, children are often enrolled during a flare (so regression to the mean is likely), and taking part in a study changes the care received. The authors correctly describe the study as a pilot; the highly significant results justify a randomized controlled trial, not a therapeutic recommendation. This is exactly the difference between the paired test on the whole mg2026 cohort (change over time) and the comparison between the randomized arms (effect of the intervention).

6.8 Two-Way ANOVA and ANCOVA, Briefly

6.8.1 Two-Way ANOVA

Sometimes we are interested in the effect of two factors on a numerical variable at the same time, for example the treatment arm and the patient's sex on the decrease in blood glucose. Two-way ANOVA tests three hypotheses: the main effect of the first factor (arm), the main effect of the second factor (sex) and the interaction between them (does the effect of the arm differ between women and men?). The interaction is often the most interesting clinical question: a treatment that works only in women or only in older people.

In mg2026, the brat × sex ANOVA for the decrease in blood glucose gives: effect of arm $F(2, 154) = 23.03$, $p < 0.001$; effect of sex $F(1, 154) = 0.66$, $p = 0.416$; interaction $F(2, 154) = 0.53$, $p = 0.591$. The cell means, for women and men, are: A 2.6 and 3.6 mg/dL; B 12.7 and 9.6 mg/dL; C 17.2 and 15.4 mg/dL. The conclusion: the effect of the intervention does not differ significantly between the sexes, so we can report the effect of the arm for the whole cohort. Be careful, as always: interactions have low power, and the absence of a significant interaction does not prove that the effect is identical in women and men.

6.8.2 ANCOVA: Adjusting for the Baseline Value

Analysis of covariance (ANCOVA) combines ANOVA with regression (Chapter 9): it compares group means after adjusting for a numerical variable, called a covariate. In randomized trials, the most common application is adjustment for the baseline value of the outcome. Even with randomization, the arms may start from slightly different values: in mg2026, mean blood glucose at enrollment was 107.8 mg/dL in arm A, 106.3 in B and 102.3 in C. ANCOVA compares blood glucose at 6 months as if all patients had started from the same baseline blood glucose.

The result on mg2026: after adjustment for baseline blood glucose, blood glucose at 6 months differs between arms, $F(2, 156) = 22.26$, $p < 0.001$; compared with arm A, adjusted blood glucose is lower by 8.07 mg/dL in arm B (95% CI 4.12-12.01) and by 13.29 mg/dL in arm C (95% CI 9.32-17.25). The estimates are almost identical to the differences between the mean decreases from ANOVA (8.1 and 13.3 mg/dL), which shows that the small imbalances at enrollment did not influence the conclusion. ANCOVA is the recommended analysis for randomized trials with a numerical outcome measured before and after, because it generally has more power than a simple comparison of the final values or of the differences. The CONSORT guideline recommends that adjustments be specified in the protocol, not chosen after looking at the data, and discourages statistical testing of baseline differences between randomized arms [56]; if randomization was done correctly, any baseline difference is, by definition, due to chance.

6.9 Reporting t-Tests and ANOVA

A complete report of a t-test or an ANOVA includes: descriptive statistics by group (n, mean, SD) or by time point; the check of the assumptions (normality, homogeneity of variances, sphericity,

with the test used); the version of the test (Student or Welch; classic or Welch ANOVA); the test statistic with its degrees of freedom (t(159) = 11.14; F(2, 157) = 22.93); the exact p value; the effect size (d, dz, eta²) and the confidence interval for the differences between means; the post hoc test and its results; the clinical interpretation [48,53].

Example for mg2026: "The decrease in fasting plasma glucose at 6 months differed between arms (A: 3.0 ± 8.8; B: 11.1 ± 11.2; C: 16.3 ± 10.6 mg/dL; ANOVA (one-way), $F(2;\ 157) = 22.93$, $p < 0.001$, $\eta^2 = 0.23$; Levene's test p = 0.163). Tukey post hoc comparisons showed larger decreases in arm B than in A (difference 8.1 mg/dL, 95% CI 3.4-12.8; p < 0.001), in arm C than in A (13.3 mg/dL, 95% CI 8.6-18.0; p < 0.001) and in arm C than in B (5.2 mg/dL, CI 95% $0.6 - 9.9$; $p = 0.024$)."

For graphs, the box plot with individual points (Figure 6.3) or the plot of means with 95% CI is preferable to bar charts with SD "error bars", which hide the distribution of the data. For paired data, the plot with individual lines (Figure 6.1 A) shows what no table can show: how many patients improve, how many get worse and by how much.

6.10 Worked Problems

Worked Problem 6.1. The Paired t-Test Calculated by Hand

Statement. In 10 hypertensive patients, SBP was measured before and after 4 weeks of treatment with an angiotensin-converting enzyme inhibitor (Table 6.2). Did SBP decrease significantly? Solve in the a-g format, with α = 0.05.

Solution. (a) Research hypothesis: 4 weeks of treatment changes the SBP of hypertensive patients. (b) H0: the mean of the differences SBP before - SBP after is zero in the population, μD = 0; H1: μD ≠ 0 (two-sided). (c) Method: the same patients measured twice, hence paired data, numerical variable; the paired-samples t-test, provided the differences are approximately normal. (d) Calculation: the differences are 12, 6, 12, 5, 8, 3, 15, 4, 9, 4; sum 78, mean mD = 7.8 mmHg. The sum of squared deviations from the mean is 151.60 (Table 6.2), so $s_D{}^2 = \frac{151.60}{9} = 16.84$ and sD = 4.10 mmHg. SED = $\frac{4.10}{\sqrt{10}} = 1.30$ mmHg. $t = \frac{7.8}{1.30} = 6.01$, df = 9. Shapiro-Wilk test on the differences: p = 0.347, normality is acceptable. (e) Result: critical t (9; 0.05) = 2.262 < 6.01; p = 0.0002. 95% CI for the mean decrease: 7.8 ± 2.262 × 1.30 = 4.86-10.74 mmHg. dz = $\frac{7.8}{4.10} = 1.90$. The appropriate graph: individual before-after lines, as in Figure 6.1 A. (f) Statistical conclusion: p < 0.001, H0 is rejected; the difference is highly significant. (g) Research conclusion: after 4 weeks of treatment, SBP decreased on average by 7.8 mmHg (95% CI 4.9-10.7 mmHg); all 10 patients had a decrease, between 3 and 15 mmHg. The main limitation: the lack of a control group; part of the decrease may reflect regression to the mean (the patients were enrolled with high values) or habituation to the measurement.

Checking with software: in Google Sheets, =T.TEST(B2:B11, C2:C11, 2, 1) gives 0.0002 (type 1 = paired). If type 2 (independent samples) were wrongly used, the result would be p = 0.008; the analysis would ignore the fact that the values come from the same patients.

Worked Problem 6.2. The Independent-Samples t-Test Calculated by Hand

Statement. Fasting plasma glucose (mg/dL) after 3 months of treatment was measured in 8 patients treated with metformin and diet and in 10 patients treated with diet alone (fictitious data). Group 1 (metformin + diet): 112, 118, 105, 120, 110, 115, 108, 124. Group 2 (diet): 125, 130, 118, 135, 128, 122, 140, 126, 132, 119. Does mean blood glucose differ between the groups?

Solution. (a) Research hypothesis: fasting plasma glucose at 3 months differs between patients treated with metformin and diet and those treated with diet alone. (b) H0: μ1 = μ2; H1: μ1 ≠ μ2. (c) Two independent groups, numerical variable: the independent-samples t-test; we check normality in each group (Shapiro-Wilk p = 0.965 and p = 0.932) and equality of variances.

(d) Calculation. Group 1: sum 912, m1 = 114.0; the deviations from the mean are -2, 4, -9, 6, -4, 1, -6, 10, and the sum of their squares is 4 + 16 + 81 + 36 + 16 + 1 + 36 + 100 = 290; $s_1{}^2 = \frac{290}{7} = 41.43$; s1 = 6.44. Group 2: sum 1275, m2 = 127.5; the deviations are -2.5, 2.5, -9.5, 7.5, 0.5, -5.5, 12.5, -1.5, 4.5, -8.5, with a sum of squares of 440.5; $s_2{}^2 = \frac{440.5}{9} = 48.94$; s2 = 7.00. The variance ratio is 48.94 / 41.43 = 1.18, and Levene's test gives p = 0.887: the variances can be considered equal.

Pooled variance: $s^2_{\text{pooled}} = \frac{290+440.5}{8+10-2} = \frac{730.5}{16} = 45.66$; spooled = 6.76. SE = $\sqrt{\left[45.66 \times \left(\frac{1}{8} + \frac{1}{10}\right)\right]} = \sqrt{45.66 \times 0.225} = \sqrt{10.27} = 3.205$. $t = \frac{127.5-114.0}{3.205} = \frac{13.5}{3.205} = 4.21$, df = 16.

(e) Critical t (16; 0.05) = 2.120 < 4.21; p = 0.0007. 95% CI for the difference: 13.5 ± 2.120 × 3.205 = 13.5 ± 6.79, that is, 6.71-20.29 mg/dL. Cohen's d = 13.5 / 6.76 = 2.00. The Welch version, for comparison: SE = $\sqrt{\frac{41.43}{8} + \frac{48.94}{10}} = \sqrt{5.179 + 4.894} = \sqrt{10.07} = 3.174$; t = 4.25; df = 15.6; p = 0.0006, practically the same result. (f) p < 0.001, H0 is rejected; the difference is highly significant. (g) At 3 months, patients treated with metformin and diet have a fasting plasma glucose on average 13.5 mg/dL lower than those treated with diet alone (95% CI 6.7-20.3 mg/dL). In a real study, one would need to check whether the groups were comparable at enrollment (ideally, through randomization); otherwise, the difference may reflect patient selection.

Worked Problem 6.3. The ANOVA Table and the Tukey Test by Hand

Statement. Pain score (0-10) 2 hours after administration of three analgesics, in 5 patients each (fictitious data): A (paracetamol) 6, 5, 7, 6, 6; B (ibuprofen) 4, 5, 3, 4, 4; C (combination) 3, 2, 3, 2, 5. Does the mean pain score differ between analgesics? If so, which ones differ from each other?

Solution. (a) Research hypothesis: analgesic efficacy (pain score at 2 hours) differs between the three treatments. (b) H0: μA = μB = μC; H1: at least one mean differs. (c) Three independent groups, numerical variable: one-way ANOVA, followed by the Tukey test if F is significant. Assumptions: independence through design; normality is hard to assess with 5 values per group (in practice, with such small groups, the Kruskal-Wallis test would also be reported); homogeneity of variances: $s^2A = \frac{2}{4} = 0.5$, $s^2B = \frac{2}{4} = 0.5$, $s^2C = \frac{6}{4} = 1.5$; Levene's test p = 0.557.

(d) Calculation. The means: $m_A = \frac{30}{5} = 6$, $m_B = \frac{20}{5} = 4$, $m_C = \frac{15}{5} = 3$; grand mean $M = \frac{65}{15} = 4.33$. SS between = 5 × [(6 - 4.33)² + (4 - 4.33)² + (3 - 4.33)²] = 23.33, df = 2, MS between = 11.67. SS within = 2 + 2 + 6 = 10.00 (equivalently: 4 × 0.5 + 4 × 0.5 + 4 × 1.5), df = 12, MS within = 0.833. $F = \frac{11.67}{0.833} = 14.0$. The complete table is Table 6.4.

(e) Critical F (2; 12; 0.05) = 3.89 < 14.0; p = 0.0007; $\eta^2 = 0.70$. The Tukey test: the q value for k = 3 groups and 12 residual degrees of freedom, at α = 0.05, is 3.77 (studentized range tables). HSD = $3.77 \times \sqrt{\frac{0.833}{5}} = 3.77 \times 0.408 = 1.54$. The differences: A - B = 2 > 1.54, significant (Tukey p = 0.012); A - C = 3 > 1.54, significant (p = 0.0006); B - C = 1 < 1.54, not significant (p = 0.23). (f) The global H0 is rejected; the differences are highly

significant. (g) Ibuprofen and the combination reduce pain more than paracetamol (by 2 and 3 points, respectively, on the 0-10 scale), and the 1-point difference between the combination and ibuprofen could not be demonstrated with 5 patients per group. Clinically, a 2-point difference on a 0-10 pain scale is usually considered perceptible to the patient.

Worked Problem 6.4. The Paired Test from Summary Statistics (mg2026)

Statement. Between enrollment and 3 months, the individual differences in fasting plasma glucose (glicemie0 - glicemie3) in the 160 patients in mg2026 have a mean of 6.08 mg/dL and a standard deviation of 9.35 mg/dL. Test whether blood glucose changed and calculate the 95% CI and dz.

Solution. H0: μD = 0; H1: μD ≠ 0. SED $= \frac{9.35}{\sqrt{160}} = \frac{9.35}{12.649} = 0.739$ mg/dL. $t = \frac{6.08}{0.739} = 8.22$, df = 159; critical t 1.975; p < 0.001. 95% CI: 6.08 ± 1.975 × 0.739 = 6.08 ± 1.46, that is, 4.62-7.54 mg/dL. dz $= \frac{6.08}{9.35} = 0.65$ (medium to large effect). Normality of the differences: Shapiro-Wilk p = 0.133. The conclusion: blood glucose decreased highly significantly in the first 3 months, on average by about 6 mg/dL (95% CI 4.6-7.5). Compare with the decrease at 6 luni (10.19 mg/dL, dz = 0.88): about three fifths of the total decrease occurred in the first three months. Note that the calculation requires the standard deviation of the differences; the standard deviations of blood glucose at each time point (15.76 at enrollment) are not sufficient.

Worked Problem 6.5. The Independent Test from Summary Statistics (mg2026)

Statement. In mg2026, the age of the women is 55.42 ± 8.95 years (n = 86), and that of the men 57.31 ± 8.50 years (n = 74). Does mean age differ between the sexes? Why would this matter for other comparisons between the sexes?

Solution. H0: μF = μM; H1: μF ≠ μM. Assumptions: Shapiro-Wilk p = 0.804 (F) and p = 0.311 (M); Levene p = 0.617. Pooled variance: $s^2_{\text{pooled}} = \frac{85\times8.95^2+73\times8.50^2}{158} = \frac{85\times80.10+73\times72.25}{158} = \frac{6808.5+5274.3}{158} = 76.47$; spooled = 8.74 years. SE $= 8.74 \times \sqrt{\frac{1}{86}+\frac{1}{74}} = 8.74 \times 0.1586 = 1.387$ years. $t = \frac{57.31-55.42}{1.387} = \frac{1.89}{1.387} = 1.36$, df = 158; critical t 1.975; p = 0.17 (Welch version: t = 1.37, df = 156.4, p = 0.173). 95% CI for the difference: -0.84 to 4.62 years. $d = \frac{1.89}{8.74} = 0.22$.

The conclusion: H0 is not rejected; mean age does not differ significantly between the sexes. The confidence interval, however, allows the men to be older by up to about 4.6 years. Why does it matter? Because SBP increases with age (in mg2026, by about 0.4 mmHg per year, Chapter 9). An age difference between the sexes could mask or create an SBP difference between the sexes. Here the age difference is small and nonsignificant, but in an observational study the comparability of the groups with respect to confounders must be checked and, if necessary, controlled through multiple regression or ANCOVA.

Worked Problem 6.6. The ANOVA Table Reconstructed from n, Means and SD (mg2026)

Statement. An article reports the decrease in blood glucose at 6 months in three arms: 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$(n = 54)$. Reconstruct the ANOVA table.

Solution. N = 160, k = 3. The grand mean is the weighted mean: $M = \frac{53\times3.0+53\times11.1+54\times16.3}{160} = \frac{159.0+588.3+880.2}{160} = \frac{1627.5}{160} = 10.17$ mg/dL (with unrounded means, 10.19).

SS between = 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; df = 2; MS between = 2399.9.

SS within = 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; df = 157; MS within = 105.1.

$F = \frac{2399.9}{105.1} = 22.8$; p < 0.001. $\eta^2 = \frac{4799.7}{4799.7+16504.8} = 0.225$.

Interpretation. The reconstructed result ($F = 22.8$, $\eta^2 = 0.225$) is very close to the one calculated by the software from the individual data ($F = 22.93$, $\eta^2 = 0.226$, Table 6.5); the small differences come from rounding the means and standard deviations to one decimal place. The technique is useful in the critical reading of articles: if an abstract reports means, SD and n by group, you can check the order of magnitude of the reported F and p value.

Worked Problem 6.7. SBP by Smoking Status: Reading an ANOVA Result

Statement. In mg2026, baseline SBP by smoking category is: former smokers 127.55 ± 14.39 mmHg($n = 33$), smokers 137.04 ± 14.63 mmHg($n = 53$), nonsmokers 128.50 ± 11.76 mmHg($n = 74$). The software reports: SS between = 2785.6, SS within = 27846.6, Levene p = 0.114, Shapiro-Wilk p between 0.23 and 0.86 across groups; Tukey: smokers - former smokers 9.49 mmHg (p = 0.0045), smokers - nonsmokers 8.54 mmHg (p = 0.0014), nonsmokers - former smokers 0.95 mmHg (p = 0.94). Complete the ANOVA table and interpret it.

Solution. df between = 3 - 1 = 2; df within = 160 - 3 = 157. MS between = 2785.6 / 2 = 1392.8; MS within = 27846.6 / 157 = 177.4. $F = \frac{1392.8}{177.4} = 7.85$; p = 0.0006 (critical F 3.05). $\eta^2 = \frac{2785.6}{2785.6+27846.6} = \frac{2785.6}{30632.2} = 0.091$, a medium to large effect. The assumptions are met (Levene p > 0.05; acceptable normality; ratio $\frac{\text{SD } 14.63}{11.76} = 1.24$).

The statistical conclusion: H0 is rejected, mean SBP differs very significantly between smoking categories. The Tukey test shows that current smokers have an SBP about 9 mmHg higher than former smokers and nonsmokers, and former smokers do not differ from nonsmokers. The research conclusion must be phrased cautiously: for this question, the study is a cross-sectional analysis; smokers may also differ in other factors (age, BMI, alcohol consumption). In Chapter 9 you will see that, after adjustment for BMI, age and sex in multiple regression, the smokers - former smokers difference remains significant (6.74 mmHg, p = 0.012), but smaller. The result is compatible with the known pressor effect of active smoking and with its reversibility after quitting, without proving it.

6.11 Proposed Exercises

1. In 8 patients with type 2 diabetes mellitus, HbA1c (%) was measured before and after 3 months of therapeutic education: before 7.8, 8.2, 7.5, 9.0, 8.4, 7.9, 8.8, 8.1; after 7.2, 7.9, 7.4, 8.1, 7.8, 7.6, 8.0, 7.7. Apply the paired t-test (α = 0.05): calculate the differences, mD, sD, t, the 95% CI and dz.
2. A study compares SBP in two independent groups: group 1, n = 25, m = 142 mmHg, s = 12; group 2, n = 30, m = 135 mmHg, s = 14. Calculate the pooled-variance t-test, the p value, the 95% CI and d. Comment on the result in relation to the 0.05 threshold.
3. Two groups: group 1, n = 12, m = 4.8, s = 1.1; group 2, n = 40, m = 6.1, s = 3.2 (a serum concentration, in arbitrary units). Calculate the t-test in the Student version and in the Welch version. Explain why the conclusions differ.

4. Three groups of 10 patients each have mean SBP of 120, 128 and 131 mmHg, with standard deviations of 10, 12 and 11 mmHg. Build the ANOVA table, calculate F and eta² and decide at α = 0.05 (critical F (2; 27) = 3.35).
5. Complete the ANOVA table: 4 groups, N = 40 patients, SS between = 360, SS within = 1440. Calculate df, MS, F and eta² and decide at α = 0.05 (critical F (3; 36) = 2.87).
6. In the study of disorders of gut-brain interaction [71], BMI was 28.15 ± 6.49 kg/m² in 52 cases and 24.47 ± 4.60 kg/m² in 84 controls. Using only these values, calculate the Welch t-test and Cohen's d. Is the result compatible with the reported p value (p = 0.001)?
7. In the study of COVID-19 in patients with diabetes [8], length of hospital stay was 11.6 ± 7.0 days in 224 patients with type 2 diabetes and 9.1 ± 5.8 days in 62 patients with type 1 diabetes. As an exercise, calculate the Welch t-test from these values. What does the ratio of SD to mean tell you about the shape of the distribution, and which test would you prefer on the individual data?
8. In mg2026, the individual differences tas0 - tas6 have a mean of 4.78 mmHg and a standard deviation of 9.05 mmHg (n = 160). Calculate t, the 95% CI for the mean difference and dz. Is the decrease clinically relevant?
9. For each situation, name the appropriate parametric test and its nonparametric alternative: (a) the heart rate of patients with COPD before and after bronchodilator treatment; (b) the height of students by sex; (c) blood glucose at 0, 3 and 6 months in the same patients; (d) blood glucose at 6 months in insured and uninsured patients; (e) the bronchial patency (Tiffeneau) index in patients treated with three different drugs; (f) the mean BMI of a group against the value 25 kg/m².
10. For the data in Exercise 4, calculate Tukey's minimum significant difference (q(3; 27; 0.05) = 3.51) and check whether any pair of groups differs. Is the result consistent with the ANOVA?
11. An abstract reports, for a marker measured in the same patients: 8061.9 ± 3684.80 ng/mL before and 15380 ± 5887 ng/mL after treatment, p < 0.001 [86]. Explain why the paired t-test cannot be recalculated from these figures and what additional information would be needed.

Common Pitfalls

- Paired data analyzed with the independent-samples test lose power; independent groups analyzed as paired produce meaningless results.
- Normality for the paired test is checked on the differences, not on each series separately.
- The number of degrees of freedom in the paired test is n - 1, where n is the number of pairs, not the total number of measurements.
- With unequal groups and unequal variances, the Student test may give wrong p values; the Welch test is used.
- Three or more groups are not compared with multiple uncorrected t-tests; an ANOVA is performed, followed by a post hoc test.
- A significant F does not tell us which groups differ; post hoc tests are not performed after a nonsignificant F in search of a significant pair.
- Repeated measurements on the same patients require repeated-measures ANOVA (with a check of sphericity), not one-way ANOVA.
- An SD close to the mean for a positive variable signals skewness; mean ± SD and the t-test may be inappropriate.

Key Points

- The paired t-test: $t = \frac{m_D}{\frac{s_D}{\sqrt{n}}}$, df = n - 1; it is the single-mean t-test applied to the differences; dz $= \frac{m_D}{s_D}$.
- The Student independent t-test: the variance is pooled $s^2_{\text{pooled}} = \frac{(n_1-1)\,{s_1}^2+(n_2-1)\,{s_2}^2}{n_1+n_2-2}$, df = n1 + n2 - 2; the Welch version does not assume equal variances; Levene's test checks the equality of variances.
- ANOVA compares between-group variation with within-group variation: F = MS between / MS within, with df k - 1 and N - k; SS total = SS between + SS within.
- $\eta^2 =$ SS between / SS total is the effect size in ANOVA; Cohen's d is the effect size for two means.
- ANOVA assumptions: independence, normality within groups, homogeneity of variances; alternatives: Welch's ANOVA with Games-Howell, Kruskal-Wallis.
- Post hoc: Tukey (all pairs, equal variances), Bonferroni or Holm (few comparisons), Games-Howell (unequal variances), Dunnett (each group versus the control).
- In mg2026: paired t-test for blood glucose 0-6 months t(159) = 11.14, p < 0.001; SBP by sex p = 0.81; ANOVA for the decrease in blood glucose by arm F(2, 157) = 22.93, p < 0.001, $\eta^2 = 0.226$; repeated-measures ANOVA F = 64.97, p < 0.001.

Self-Assessment Questions

1. (SC) Fasting plasma glucose was measured at enrollment and at 6 months in the same 160 patients. The appropriate parametric test for comparing the two time points is:
    a) the independent-samples t-test
    b) the paired-samples t-test
    c) one-way ANOVA
    d) the chi-square test
    e) the z-test for a proportion
2. (SC) In a paired t-test with 25 patients, the number of degrees of freedom is:
    a) 48
    b) 50
    c) 24
    d) 25
    e) 23
3. (SC) In a one-way ANOVA with 4 groups and a total of 60 patients, the degrees of freedom of the numerator and denominator of the F statistic are:
    a) 4 and 60
    b) 3 and 56
    c) 3 and 59
    d) 4 and 56
    e) 59 and 3
4. (SC) Levene's test has p = 0.003 when comparing two independent groups of 15 and 60 patients. The appropriate version of the t-test is:
    a) the paired t-test
    b) the Student t-test with pooled variance
    c) the Welch t-test
    d) the z-test
    e) the single-mean t-test
5. (SC) In an ANOVA, SS between groups = 300 and SS total = 1200. Eta² is:

a) 0,20
b) 0,25
c) 0,33
d) 4,00
e) it cannot be calculated without the degrees of freedom

6. (MC) Regarding the paired-samples t-test, the following statements are correct:
   a) it works on the individual differences between the two measurements
   b) normality is checked on the differences
   c) it generally has more power than the independent test applied to the same paired data
   d) it requires the two series to have equal variances
   e) the nonparametric alternative is the Wilcoxon signed-rank test
7. (MC) The assumptions of one-way ANOVA are:
   a) independence of observations
   b) an approximately normal distribution in each group
   c) homogeneity of variances
   d) groups that must be of equal size
   e) at least 30 patients in each group
8. (MC) In mg2026, the ANOVA for the decrease in blood glucose by arm gives $F(2, 157) = 22.93$, $p < 0.001$, $\eta^2 = 0.226$, and the Tukey test shows significant differences for all three pairs. The following statements are correct:
   a) the hypothesis that the three means are equal is rejected
   b) about 23% of the variability in the decrease in blood glucose is explained by the arm
   c) ANOVA alone shows that arm C differs from arm B
   d) the Tukey test keeps the overall type I error risk at the level α
   e) the result is due solely to the large number of patients
9. (MC) Appropriate choices when the group variances differ greatly are:
   a) the Welch t-test for two groups
   b) Welch's ANOVA for three or more groups
   c) the Games-Howell post hoc test
   d) the Student t-test with pooled variance
   e) the classic Tukey test, without any check
10. (MC) Regarding repeated-measures ANOVA, the following statements are correct:

a) it is applied when the same patients are measured at three or more time points
b) sphericity is checked with Mauchly's test
c) the Greenhouse-Geisser correction changes the degrees of freedom of the F test
d) the nonparametric alternative is the Kruskal-Wallis test
e) for two time points, the result is equivalent to that of the paired t-test

Chapter 7. Common Nonparametric Statistical Tests: Mann-Whitney U, Wilcoxon, Kruskal-Wallis, Friedman and the Chi-Square Test

Objectives

By the end of this chapter, the student will be able to:

- recognize the situations in which the assumptions of parametric tests are not met (ordinal data, marked skewness, small samples, extreme values) and choose the appropriate nonparametric alternative;
- assign ranks to a series of values, including when there are tied values, and check the sum of the ranks;
- calculate by hand the Mann-Whitney U statistic, the Wilcoxon T statistic, the Kruskal-Wallis H statistic and the Friedman statistic on small samples and compare them with the critical values;
- correctly interpret the hypotheses of rank-based tests (distributions, medians) and report an effect size (rank-biserial correlation, epsilon squared, Kendall's W);
- build a contingency table, calculate the expected frequencies, the chi-square statistic and the degrees of freedom, and check the assumptions;
- choose between the chi-square test, the Yates correction, Fisher's exact test, McNemar's test, the test for trend and the goodness-of-fit test;
- calculate and interpret the phi coefficient and Cramér's V, understanding that the p value does not measure the strength of the association;
- critically read the methods section and the results of articles published by colleagues in the discipline who have used these tests.

7.1 When We Can No Longer Use Parametric Tests

In Chapter 6 we worked with the t-test and with ANOVA. Both compare means and rest on a few assumptions: the variable is numerical, the distribution in each group is approximately normal (or the sample is large enough for the mean to behave normally), and the group variances do not differ too much. In medical practice, these assumptions fail more often than we would like. The patient does not answer the question "how satisfied are you?" with a number measured in milligrams, but ticks a box from 1 to 5. Length of hospital stay has a long right tail, because a single complicated patient may stay 40 days while the others stay 5. And in a pilot study with 12 patients we have no way of checking normality convincingly.

Nonparametric tests were built precisely for these situations. Their common idea is simple: instead of working with the actual values, we work with their position in a ranking, that is, with the ranks. A patient who stayed 40 days in hospital no longer pulls the mean up with all their weight; they simply become "last in the ranking". We lose part of the information, but gain robustness.

7.1.1 Four Typical Situations

The first situation is that of **ordinal data**. The Likert satisfaction scale, the NYHA class of heart failure, the grade of liver fibrosis F0-F4, the mMRC dyspnea score or the level of education have ordered categories, but the distance between categories is not necessarily equal. Moving from "dissatisfied" to "neutral" is not necessarily equivalent to moving from "satisfied" to "very

satisfied". The mean of such codes has a questionable meaning, and the t-test assumes precisely that the differences between values make arithmetic sense. For ordinal data, rank-based tests are the natural choice.

The second situation is **marked skewness** of a numerical variable. In the mg2026 dataset, triglycerides have a skewness of 0.95 and the Shapiro-Wilk test gives $p < 0.001$, and body mass index (BMI) has a skewness of 1.42 and an excess kurtosis of 3.68. When the groups are large, the t-test holds up fairly well against skewness, thanks to the central limit theorem. When the groups are small or the skewness is strong, the mean becomes an unrepresentative summary, and the rank-based test is safer. An alternative discussed in Chapter 3 is the logarithmic transformation, which for triglycerides brings the skewness to -0.28 and SW $p = 0.200$.

The third situation is the **small sample**. With 6-8 patients per group, the normality test has very low power: a "$p > 0.05$" result on Shapiro-Wilk does not prove normality, it only shows that we do not have enough data to refute it. In such studies, many authors choose a nonparametric test from the outset, and the choice is reasonable if it was specified in the protocol.

The fourth situation is the presence of **extreme values** (outliers) that are not measurement errors. A patient with triglycerides of 442 mg/dL is real and must not be deleted just because they spoil the mean. Rank-based tests keep this patient but limit their influence: whether the value is 442 or 4420, the rank stays the same.

7.1.2 What "Nonparametric" Means and What It Does Not Mean

The term "nonparametric" comes from the fact that these tests do not assume a particular shape of the distribution (for example, the normal distribution with parameters μ and σ). This does not mean that they have no assumptions. All of them assume independent observations (or pairs that are independent of each other), and some assume that the distributions being compared have the same shape, differing only in location, if we want to interpret the result as a difference in medians. The chi-square test, in turn, requires sufficiently large expected frequencies.

A common confusion in exams is the idea that a nonparametric test "compares medians" under any conditions. More precisely, the Mann-Whitney test checks whether the values in one group tend to be larger than those in the other group. If the two distributions have the same shape and the same spread, this is equivalent to a difference between medians. If the shapes differ greatly, the test can be significant even with equal medians. In reporting, we therefore say "the distributions differ" or "values were higher in group C", and give the medians and interquartile ranges as a description. Table 7.1 sets the parametric tests from Chapter 6 side by side with their nonparametric equivalents.

Table 7.1. Correspondence between parametric and nonparametric tests.

Situation	Parametric test	Nonparametric test	What is compared in the nonparametric version
2 independent groups	unpaired t-test (Welch)	Mann-Whitney U (Wilcoxon rank-sum)	the ranks of the two groups
2 measurements on the same subjects	paired t-test	Wilcoxon (signed-rank); sign test	the signed ranks of the differences
≥ 3 independent groups	one-way ANOVA	Kruskal-Wallis, then Dunn	the mean ranks of the groups
≥ 3 measurements on the same subjects	repeated-measures ANOVA	Friedman, then Wilcoxon with correction	the ranks within each subject
association of two variables	Pearson correlation	Spearman correlation	the ranks of both variables
two qualitative variables	(no parametric version exists)	chi-square, Fisher, McNemar	observed versus expected frequencies

7.1.3 The Relative Efficiency of Nonparametric Tests

The price paid for robustness is a loss of power when the data really are normal. This loss is measured by the asymptotic relative efficiency: the ratio between the number of subjects needed by the parametric test and the number needed by the nonparametric test for the same power. For Mann-Whitney compared with the t-test, with perfectly normal data, the efficiency is $3/\pi$, that is, about 0.955 (95.5%) [53,63]. In concrete terms, what the t-test does with 95 patients, Mann-Whitney does with about 100. The loss is small. For heavy-tailed distributions (many extreme values), the efficiency of the rank-based test can exceed 1, meaning that the nonparametric test becomes more powerful than the t-test, and for no continuous distribution does it fall below about 0.86. The Wilcoxon test for paired data has the same 95.5% efficiency relative to the paired t-test.

The sign test, which keeps only the direction of the difference (plus or minus) and ignores its size, is much more wasteful: its efficiency with normal data is $2/\pi$, about 0.64. In Section 7.4 we will see an example from mg2026 in which the same question receives $p = 0.043$ with the paired t-test, $p = 0.052$ with Wilcoxon and $p = 0.092$ with the sign test, exactly in order of efficiency.

The practical conclusion is reassuring. If you have a real reason to choose a nonparametric test (ordinal data, strong skewness with small n, extreme values), you lose almost nothing. If the data are numerical and approximately normal, the parametric test remains the first choice, because it directly estimates the difference in means with a confidence interval, which is easier to interpret clinically.

7.2 Ranks: The Raw Material of Nonparametric Tests

7.2.1 Assigning Ranks, with Tied Values

The rank of a value is its position in the series sorted in ascending order: the smallest value receives rank 1, the next one rank 2, and so on. The problem arises with tied values (ties), which are very common with ordinal scores. The rule is that tied values receive the **mean of the ranks** they would have occupied if they had been slightly different.

Let us take eight pain scores on the 0-10 scale, recorded by patients after a tooth extraction: 3, 5, 5, 2, 7, 5, 9, 3. We sort them: 2, 3, 3, 5, 5, 5, 7, 9. The value 2 occupies position 1, so rank 1. The two values of 3 occupy positions 2 and 3, so each receives $(2 + 3)/2 = 2.5$. The three values of 5 occupy positions 4, 5 and 6, so each receives $(4 + 5 + 6)/3 = 5$. The value 7 receives rank 7, and the value 9 rank 8.

Table 7.2. Assigning ranks with tied values (pain scores, n = 8).

Original value, in order of collection	**3**	**5**	**5**	**2**	**7**	**5**	**9**	**3**
Rank	2,5	5	5	1	7	5	8	2,5

7.2.2 The Check: The Sum of the Ranks

Before any calculation, always check the sum of the ranks. For N observations, the sum of the ranks 1, 2,..., N is:

$$\frac{N(N+1)}{2}$$

For N = 8, the sum must be $\frac{8\times9}{2} = 36$. We add: 2.5 + 5 + 5 + 1 + 7 + 5 + 8 + 2.5 = 36. Correct. The averaging rule for tied values preserves the sum, which is why the check also works

when there are ties. In an exam, a wrong sum of ranks is the first sign that a value was skipped or a mean rank was miscalculated, and the check takes ten seconds.

Tied values also have an effect on the variability of the ranks: they reduce it slightly. That is why the test formulas include a "correction for ties", which software applies automatically. On the small samples calculated by hand we will show it once, so that you know where the small differences between your result and the one from JASP come from.

7.3 The Mann-Whitney U Test

The test was proposed in the form with the U statistic by Mann and Whitney in 1947 [92], starting from the rank-sum test described by Wilcoxon two years earlier [93]. That is why some software lists it as the "Wilcoxon rank-sum test"; it is the same test. We use it to compare two independent groups when the variable is ordinal, or numerical without a normal distribution.

7.3.1 The Hypotheses

In words, the null hypothesis states that the values in the two populations come from the same distribution: if we randomly choose a patient from group 1 and one from group 2, the probability that the first has the larger value is 0.5. The two-sided alternative hypothesis states that this probability differs from 0.5, that is, the values in one group tend to be systematically larger.

H0: P(X1 > X2) = 0.5 (the distributions are identical) H1: P(X1 > X2) ≠ 0.5

If we accept the assumption that the distributions have the same shape, the hypotheses can be written in terms of medians: H0: Me1 = Me2, H1: Me1 ≠ Me2. This is the formulation used in the course assignments and it is acceptable, provided you know the assumption behind it.

7.3.2 Calculating U by Hand

The steps are the same for any small sample:

1. all the values from the two groups are put into a single series and ranked, with the mean rank for tied values;
2. the sum of the ranks in each group, R1 and R2, is calculated, and $R_1 + R_2 = \frac{N(N+1)}{2}$ is checked;
3. $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$ and $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$ are calculated; check: U1 + U2 = n1 x n2;
4. the test statistic is U = the smaller of U1 and U2;
5. U is compared with the critical value from the table for n1, n2 and α; **H0 is rejected if U is less than or equal to the critical value** (note: the opposite of t and chi-square).

An example: in a pilot study, 5 surgical patients received a phone app with breathing exercises and reminders for analgesics, and 6 patients received usual care. The pain score at 24 hours (0-10) was: app 2, 3, 3, 5, 6; control 4, 6, 7, 7, 8, 9. Figure 7.1 shows the values and their ranks in the combined sample.

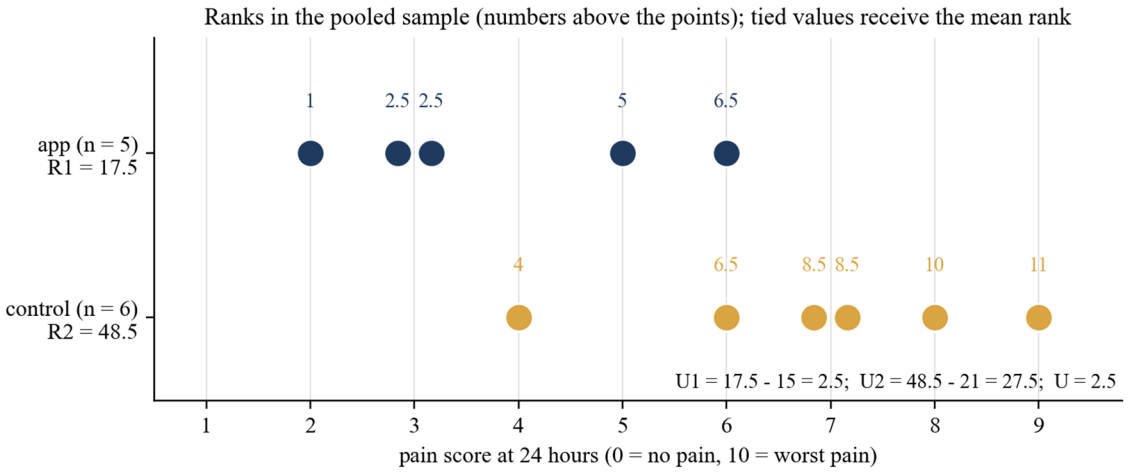


Figure 7.1. Calculating the Mann-Whitney U statistic in a small example: pain score at 24 hours in 5 patients with the app and 6 control patients. The numbers above the points are the ranks in the combined sample of 11 values; tied values (3 and 3, 6 and 6, 7 and 7) receive the mean rank. Fictitious data, constructed for hand calculation.

The combined sorted series is 2, 3, 3, 4, 5, 6, 6, 7, 7, 8, 9 (N = 11). The ranks: 2 → 1; the two values of 3 → 2.5; 4 → 4; 5 → 5; the two values of 6 (one from each group) → 6.5; the two values of 7 → 8.5; 8 → 10; 9 → 11.

$$R_1(\text{ app }) = 1 + 2.5 + 2.5 + 5 + 6.5 = 17.5 R_2(\text{ control }) = 4 + 6.5 + 8.5 + 8.5 + 10 + 11 = 48.5$$

Check: $17.5 + 48.5 = 66 = \frac{11 \times 12}{2}$. Correct.

$$U_1 = 17.5 - \frac{5 \times 6}{2} = 17.5 - 15 = 2.5 U_2 = 48.5 - \frac{6 \times 7}{2} = 48.5 - 21 = 27.5$$

Check: 2.5 + 27.5 = 30 = 5 x 6. The test statistic is U = 2.5.

What does U mean, intuitively? If we form all 5 x 6 = 30 pairs (one patient with the app, one control patient), U1 counts how many times the patient with the app has the higher score (ties count as one half). Only 2.5 of the 30 pairs "favor" the app group as having more pain. In other words, in 27.5 of the 30 comparisons the patient with the app has less pain.

The critical value for n1 = 5, n2 = 6, two-sided test, α = 0.05, is 3 (Table 7.15, at the beginning of Section 7.11). Since U = 2.5 ≤ 3, we reject H0. The exact p value given by the software is 0.030. The statistical conclusion: the difference between the distributions of the pain score is significant at the 0.05 level. The research conclusion: patients who used the app reported less pain at 24 hours (median 3 versus 7). As this is a pilot study with 11 patients, the result must be confirmed in a larger sample.

7.3.3 The Normal Approximation

For larger groups (in practice, both groups above 10-20 subjects), the tables are no longer needed. Under H0, U has an approximately normal distribution with mean and standard deviation:

$$\mu_U = \frac{n_1 \times n_2}{2} \sigma_U = \sqrt{\left[\frac{n_1 \times n_2 \times (N+1)}{12}\right]}$$

and the statistic

$$z = \frac{U - \mu_U}{\sigma_U}$$

is compared with 1.96 for α = 0.05, two-sided. If there are many tied values, σU is corrected downward, and software does this automatically. In the example above, μU = 15, σU = √(30 x 12/12) = 5.48 (5.44 with the correction for ties), so $z = \frac{2.5-15}{5.44} = -2.30$, with a two-sided p of about 0.02. With such small groups the approximation is only indicative; the exact value (0.030) is the correct one.

7.3.4 Effect Size: The Rank-Biserial Correlation

The p value tells us how surprising the difference would be if H0 were true, but not how large it is. For Mann-Whitney, the simplest measure of effect is the rank-biserial correlation:

$$r_{\text{rb}} = 1 - \frac{2U}{n_1 \times n_2}$$

where U is the smaller of U1 and U2. Its value ranges from 0 (no difference) to 1 (all values in one group are larger than all values in the other). In the app example, $r_{\text{rb}} = 1 - \frac{2\times 2.5}{30} = 0.83$, a very large effect. Another version, used by some authors, is $r = \frac{z}{\sqrt{N}}$. As a rough guide, 0.1 is considered a small effect, 0.3 medium and 0.5 large [76], but the thresholds are conventions, not laws.

7.3.5 Application to the mg2026 Dataset: Satisfaction in Arm C Versus Arm A

At 6 months, the patients in the mg2026 study rated their satisfaction with the program they received on a scale from 1 (not at all satisfied) to 5 (very satisfied). The frequencies by arm are in Table 7.3, and the percentage distribution is in Figure 7.2.

Table 7.3. Satisfaction at 6 months by arm (number of patients), mg2026 dataset.

Arm	1	2	3	4	5	Total	Median (IQR)
A: counseling	1	13	28	9	2	53	3 (2-3)
B: + app	3	3	14	24	9	53	4 (3-4)
C: + app and teleconsultation	0	3	12	22	17	54	4 (3-5)
Total	4	19	54	55	28	160	4 (3-4)

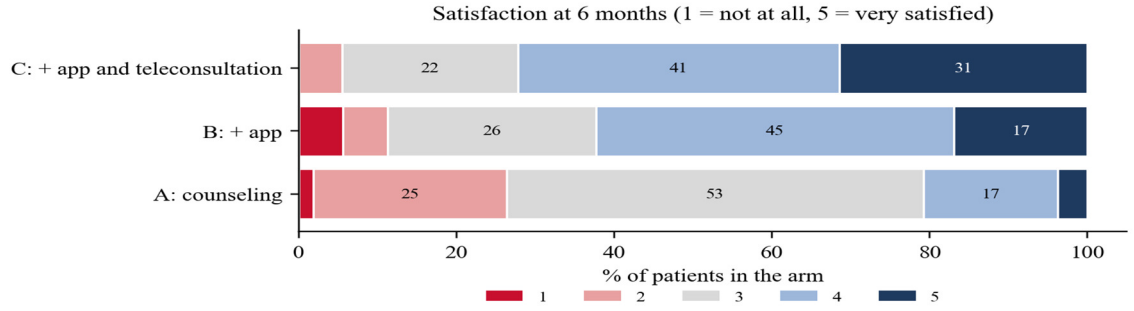


Figure 7.2. Distribution of satisfaction at 6 months (1 = not at all satisfied, 5 = very satisfied), as percentages of the patients in each arm. In arm C, ratings of 4 and 5 predominate; in arm A, a rating of 3. Synthetic data (mg2026 dataset).

The research question: does the satisfaction of patients in arm C differ from that of patients in arm A? The variable is ordinal, so we use Mann-Whitney. There is no point in testing the normality of a variable with five possible values.

The result from the software: the medians are 4 (IQR 3-5) in arm C and 3 (IQR 2-3) in arm A, U = 2265.5, p < 0.001. Let us see where the figure comes from. The rank sums in the combined sample of 107 patients are RC = 3750.5 and RA = 2027.5. Then $UC = 3750.5 - \frac{54\times55}{2} = 3750.5 - 1485 = 2265.5$, and UA = 54 x 53 - 2265.5 = 596.5. Software sometimes reports the U of one group, sometimes that of the other; both lead to the same p value. The normal approximation: $\mu_U = \frac{54\times53}{2} = 1431$, σU = 153.5 (with the correction for the numerous ties), so $z = \frac{2265.5-1431}{153.5} = 5.44$, well above 1.96.

Effect size: $r_{\text{rb}} = 1 - \frac{2\times596.5}{2862} = 0.58$, a large effect. The direct interpretation: if we randomly choose a patient from arm C and one from arm A, the probability that the first is more satisfied (with ties split in half) is (1 + 0.58)/2 = 0.79.

The statistical conclusion: we reject H0; satisfaction differs significantly between arms C and A (p < 0.001). The clinical conclusion: patients who received the app and monthly teleconsultation were more satisfied than those who received counseling only, and the difference is statistically significant and, in addition, large. Be careful, however: satisfaction is not the same thing as effectiveness; it may also reflect the extra attention received in arm C.

Case study: the mobile app for asthma, three nonparametric tests in a single article

Munteanu and colleagues, including Dr. Frandeş from our discipline, evaluated in a controlled pilot study a phone app designed to improve treatment adherence and disease self-management in 93 patients with asthma, followed for one year [94]. The design combines independent groups (standard treatment versus treatment plus app) with repeated measurements on the same patients. The authors consistently chose nonparametric tests: Mann-Whitney U for comparing the groups (p < 0.001), the Friedman test for the repeated measurements, with Bonferroni correction for the pairwise comparisons (p < 0.001), and Pearson's chi-square test for a 2 x 2 table in which the proportions compared were 10.30% and 46.30%, with $\chi^2(1) = 13.707$, p < 0.001. Reading this abstract is a good exercise in matching the test to the question: a comparison between independent groups, a comparison over time in the same patients, a comparison of proportions. What would have added value: the effect size for Mann-Whitney (for example, the rank-biserial correlation) and, for the 2 x 2 table, the relative risk with its confidence interval, which shows the clinician how large the difference is, not only that it exists. As this is a pilot study, the results generate hypotheses for a larger randomized trial.

7.4 The Wilcoxon Test for Paired Data and the Sign Test

The Wilcoxon signed-rank test [93] is the nonparametric alternative to the paired t-test. We use it when we have two measurements on the same subjects (before and after treatment, the right eye and the left eye, two methods applied to the same sample) and the differences do not have a normal distribution or the variable is ordinal.

7.4.1 The Logic of the Test

For each subject we calculate the difference d = before - after. If the treatment has no effect, positive and negative differences should be equally frequent and equally large. The test sorts the absolute values of the differences, assigns them ranks and then puts the sign back. Under H0, the sum of the positive ranks (W+) should be close to the sum of the negative ranks (W-). Differences equal to zero are dropped (classic version), and n becomes the number of nonzero differences.

H0: the distribution of the differences is symmetric around 0 (median of the differences = 0)
H1: median of the differences ≠ 0

7.4.2 The Hand Calculation

Ten overweight patients followed a three-month nutritional counseling program. Weight (kg) before and after is given in Table 7.4.

Table 7.4. Calculation of the Wilcoxon test in 10 patients (fictitious data).

Patient	Before	After	d	abs(d)	Rank of abs(d)	Signed rank
1	92,0	89,5	2,5	2,5	6	+6
2	85,5	85,5	0,0	0	dropped	-
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

Patient 2 has a difference of zero and is dropped, so n = 9. The sorted absolute values: 0.8 (rank 1), 1.2 (rank 2), 1.5 (rank 3), 2.2 and 2.2 (ranks 4 and 5, so 4.5 each), 2.5 (rank 6), 3.2 (rank 7), 4.2 (rank 8), 6.2 (rank 9).

$$W^+ = 6 + 8 + 4.5 + 4.5 + 9 + 7 + 3 = 42\, W^- = 2 + 1 = 3$$

Check: $W^+ + W^- = 45 = \frac{9\times10}{2}$. Correct. The test statistic is T = the smaller of W+ and W-, that is, T = 3.

The critical value for n = 9, two-sided, α = 0.05, is 5 (Table 7.16). As with Mann-Whitney, **H0 is rejected if T is less than or equal to the critical value**. Since 3 ≤ 5, we reject H0; exact p = 0.020. For larger n, the normal approximation is used:

$$\mu_T = \frac{n(n+1)}{4} = 22.5 \sigma_T = \sqrt{\left[\frac{n(n+1)\,(2n+1)}{24}\right]} = \sqrt{\frac{9\times10\times19}{24}} = 8.44 z = \frac{3-22.5}{8.44} = -2.31,\ p \approx 0.021$$

The conclusion: weight decreased significantly after the program (median of the differences 2.2 kg; 7 of the 9 patients with a nonzero difference lost weight).

7.4.3 The Sign Test

The sign test is the simplest test for paired data. It only counts how many differences are positive and how many are negative, ignoring their size. Under H0, each nonzero difference has a probability of 0.5 of being positive, so the number of positive differences follows a binomial distribution with n and p = 0.5.

In the example above we have 7 positive differences out of 9. The two-sided probability of obtaining 7 or more (or, symmetrically, 2 or fewer) in a binomial distribution with n = 9 and p = 0.5 is 0.18. The sign test does not find a significant difference where Wilcoxon finds p = 0.020. The explanation: the two patients who gained weight gained very little (1.2 and 0.8 kg), while those who lost weight lost much more. Wilcoxon uses this information; the sign test throws it away. The sign test remains useful when we only have the direction of change ("better", "the same", "worse"), for example in the patient's global assessment.

7.4.4 Application to mg2026: Weight in Arm A, a Borderline Result

In arm A (counseling only, n = 53), the median weight difference between enrollment and 6 months is 1.1 kg. 32 patients lost weight, 19 gained weight and 2 stayed the same. The Wilcoxon test gives W- = 456 and W+ = 870 for n = 51 nonzero differences, with p = 0.052.

Here it is clear why it makes no sense to treat the 0.05 threshold as a wall. With p = 0.052, we do not reject H0, but this does not mean that the program in arm A has no effect at all: it means that the data are not convincing enough at the chosen threshold. We will phrase the conclusion cautiously: "at 6 months, in the standard counseling arm, weight loss (median 1.1 kg) did not reach the threshold of statistical significance (Wilcoxon, p = 0.052)". Absence of evidence is not evidence of absence [79].

The same example illustrates the relative efficiency discussed in Section 7.1.3. The weight differences in arm A actually have a distribution compatible with the normal (Shapiro-Wilk on the differences p = 0.696), so the appropriate test was the paired t-test. It gives a mean difference of 0.71 kg (SD 2.48), t = 2.07, df 52, p = 0.043. The sign test (32 pluses out of 51) gives p = 0.092. Three tests, the same question, three p values ordered exactly by the amount of information used: the mean and variance (t), the ranks (Wilcoxon), only the signs (sign test). The lesson is not "choose the test that comes out significant", but "choose the appropriate test before seeing the result"; here, the paired t-test was justified by the distribution of the differences. In the other two arms there is no ambiguity: Wilcoxon gives p < 0.001 both in arm B (median 2.3 kg) and in arm C (median 3.55 kg).

An example from our colleagues' research shows what paired data reported as medians look like. Gherbon and colleagues, with Dr. Frandeș, retrospectively followed 200 Romanian patients with type 2 diabetes mellitus treated with SGLT-2 inhibitors, comparing the values at the start of treatment with those after one year [59]. The data are paired (the same patients, two time points), and the results are reported as median and interquartile range: blood glucose decreased from 180.00 mg% (IQR 154.50-207.00) to 130.00 mg% (IQR 117.50-150.00), p < 0.001; HbA1c from 8.40% (IQR 7.98-9.15%) to 7.30% (IQR 6.90-7.95%), p < 0.001; weight from 90.50 kg (82.00-106.50) to 89.00 kg (77.50-100.00), p = 0.018; systolic blood pressure from 140.00 mmHg (130.00-160.00) to 120.00 mmHg (115.50-130.00), p < 0.001. Waist circumference, on the other hand, appears as mean ± SD, from 107.05 (± 16.39) to 102.50 (± 15.11), p = 0.042. The abstract does not name the test, but the choice of medians suggests a nonparametric paired test (Wilcoxon) for the skewed variables and the paired t-test for waist circumference. Two critical reading observations: the abstract calls these medians "means", a terminological confusion that an attentive student will spot immediately, and for paired data the most informative figure would have been the median of the individual differences (with IQR or a confidence interval), alongside the medians at the two time points. The difference between the medians (for example 90.50 - 89.00 = 1.50 kg) is not necessarily equal to the median of the differences.

7.5 The Kruskal-Wallis Test

When we have three or more independent groups and the variable is ordinal, or numerical without a normal distribution, we use the Kruskal-Wallis test [95], the nonparametric alternative to one-way ANOVA. It can be seen as a generalization of the Mann-Whitney test to k groups.

7.5.1 The Hypotheses and the H Statistic

H0: the k populations have the same distribution (if the shapes are similar: Me1 = Me2 =... = Mek) H1: at least one population differs from the others (at least one median differs)

All N values are sorted together and ranked. For each group i, the sum of the ranks Ri is calculated. If H0 is true, the mean rank in each group should be close to the overall mean rank, (N + 1)/2. The test statistic measures how much the mean ranks of the groups deviate from this value:

$$H = \left[\frac{12}{N(N+1)}\right] \times \sum\left(\frac{{R_i}^2}{n_i}\right) - 3(N+1)$$

For groups that are not very small (as a rule, at least 5 subjects in each group), H approximately follows a chi-square distribution with k - 1 degrees of freedom. H0 is rejected if H exceeds the critical value: 5.99 for 2 degrees of freedom, 7.81 for 3, 9.49 for 4 (α = 0.05). When there are tied values, H is divided by the correction factor

$$C = 1 - \frac{\sum(t^3 - t)}{N^3 - N}$$

where t is the number of values in a group of ties. The correction increases H slightly.

7.5.2 The Hand Calculation

Fifteen surgical patients received one of three analgesic regimens: regimen 1 (paracetamol), regimen 2 (paracetamol + anti-inflammatory drug), regimen 3 (multimodal analgesia). The pain score at 24 hours (0-10) was:

- regimen 1: 6, 7, 5, 8, 7
- regimen 2: 4, 5, 3, 5, 6
- regimen 3: 2, 3, 4, 1, 3

We sort the 15 values: 1, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 6, 7, 7, 8. The ranks: 1 → 1; 2 → 2; the three values of 3 occupy positions 3, 4, 5 → rank 4; the two values of 4 occupy positions 6, 7 → 6.5; the three values of 5 occupy positions 8, 9, 10 → 9; the two values of 6 → 11.5; the two values of 7 → 13.5; 8 → 15.

Table 7.5. Ranks for the Kruskal-Wallis test (fictitious data).

Regimen 1	Rank	Regimen 2	Rank	Regimen 3	Rank
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

Check: $62.5 + 40 + 17.5 = 120 = \frac{15 \times 16}{2}$. 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 for ties: we have two groups of three tied values each (the values 3 and 5) and three groups of two each (the values 4, 6 and 7). $\sum(t^3 - t) = 2 \times 24 + 3 \times 6 = 66$, and $C = 1 - \frac{66}{3375-15} = 0.980$. H corrected $= \frac{10.125}{0.980} = 10.33$.

With 2 degrees of freedom, the critical value is 5.99. Since 10.33 > 5.99, we reject H0 (p = 0.006). The medians are 7, 5 and 3. Conclusion: the pain score differs between the three regimens. The test does not tell us, however, which regimens differ from each other; for that we need post hoc comparisons.

7.5.3 Post Hoc Comparisons: Dunn's Test

After a significant Kruskal-Wallis test, the most widely used procedure is Dunn's test. For each pair of groups, it compares the mean ranks (calculated in the common ranking of all N subjects) using a z statistic:

z = (mean rank i - mean rank j)/√{[N(N + 1)/12] x (1/ni + 1/nj)}

The p values obtained are corrected for the number of comparisons. The simplest correction is Bonferroni: each p is multiplied by the number of comparisons (3 for three groups), or, equivalently, the 0.05 threshold is divided by 3. The Holm correction, applied stepwise, is less conservative. A common mistake is to run three separate Mann-Whitney tests, each at the 0.05 threshold, without any correction: the probability of obtaining at least one false positive result then rises from 5% to almost 14%.

7.5.4 Application to mg2026: Decrease in Blood Glucose and Satisfaction Across Three Arms

In Chapter 6 we compared the decrease in fasting plasma glucose at 6 months between the three arms with ANOVA: F(2, 157) = 22.93, p < 0.001. The Kruskal-Wallis test on the same data gives H = 36.63, df 2, p < 0.001. The calculation follows exactly the steps above, only on 160 patients. The rank sums are 2696 (A), 4554.5 (B) and 5629.5 (C), that is, mean ranks of 50.9, 85.9 and 104.3, compared with the overall mean rank (160 + 1)/2 = 80.5.

$$H = \left[\frac{12}{160\times161}\right] \times \left(\frac{2696^2}{53} + \frac{4554.5^2}{53} + \frac{5629.5^2}{54}\right) - 3 \times 161 = 0.000466 \times (137\,139.9 + 391\,386.2 + 586\,875.4) - 483 = 36.60$$

, and with the correction for ties 36.63.

The medians of the decrease in blood glucose are 4 mg/dL (IQR -3 to 9) in arm A, 14 mg/dL (3 to 20) in arm B and 16 mg/dL (9 to 23) in arm C. Dunn's test with Bonferroni correction shows significant differences between A and B (p = 0.0003) and between A and C (p < 0.001), but not between B and C (adjusted p = 0.12; unadjusted 0.041). It is interesting to compare this with the Tukey test in Chapter 6, which had also found the C-B difference significant (p = 0.024). This is not a contradiction: rank-based tests are slightly more conservative when the data are approximately normal, and the Bonferroni correction is stricter than Tukey. For this numerical variable, with a reasonable distribution in each arm and Levene p = 0.163, the main analysis remains ANOVA with Tukey; Kruskal-Wallis serves as a sensitivity analysis that confirms the direction of the result.

Satisfaction, being ordinal, is analyzed directly with Kruskal-Wallis: H = 32.16, df 2, p < 0.001. The mean ranks are 53.5 (A), 86.7 (B) and 100.9 (C). Dunn's comparisons with Bonferroni: A versus B p = 0.0003, A versus C p < 0.001, B versus C p = 0.29. Thus, both arms with the app produced more satisfied patients than counseling alone, and adding teleconsultation did not significantly change satisfaction compared with the app alone. As a check, the direct Mann-Whitney test between B and C gives p = 0.079.

7.5.5 Effect Size for Kruskal-Wallis

The analogue of eta squared from ANOVA is epsilon squared:

$$\varepsilon^2 = \frac{H}{N-1}$$

For the decrease in blood glucose, $\varepsilon^2 = \frac{36.63}{159} = 0.23$; for satisfaction, $\varepsilon^2 = \frac{32.16}{159} = 0.20$. Both are large effects, comparable to the $\eta^2 = 0.226$ obtained by ANOVA for blood glucose. Reporting this measure is good practice and appears more and more often in articles, as in the HIV study presented below.

> **Case study: Kruskal-Wallis in three articles by our colleagues**
>
> Gherbon and colleagues, with Dr. Frandeș, performed a cross-sectional analysis of 453 adults with type 2 diabetes, in whom severe diabetes-related distress was identified in 19 participants (4.2%) [69]. To compare scores across three ordered categories they used the Kruskal-Wallis test, with p = 0.003, followed by a multivariable logistic regression in which age had an aOR = 1.163 per year (95% CI 1.055-1.282; p = 0.002). The logic is clear: distress

scores are ordinal and skewed (few patients with high values), so a rank-based test is appropriate. Hasan and colleagues, with Dr. Tudor, studied 90 patients one month after thoracic surgery and compared anxiety and pain intensity across three groups defined by coping style, using Kruskal-Wallis: p = 0.028, p = 0.022, p = 0.042 and p = 0.007 for four variables [96]. Scores from validated questionnaires are ordinal variables, and the test is well chosen; the question an attentive reader asks is how many comparisons were made in total and whether they were taken into account, because three of the four p values lie between 0.02 and 0.05. Marin and colleagues, with Dr. R. Moleriu and Dr. Muntean, studied 462 adults with HIV on antiretroviral treatment [68]. Discordant immunovirological responses occurred in 30.7% of patients (14.5% and 16.2% for the two types of discordance), and the comparisons between the three response groups used Kruskal-Wallis with effect size: H = 153.62, p < 0.001, $\varepsilon^2 = 0.33$ and H = 115.10, p < 0.001, $\varepsilon^2 = 0.25$. This is the model of complete reporting: the statistic, the p value and the effect size. An ε^2 of 0.33 means that about one third of the variability in the ranks is explained by group membership.

7.6 The Friedman Test

The Friedman test [97] is the nonparametric alternative to repeated-measures ANOVA. We use it when the same subjects are assessed at three or more time points (or under three or more conditions, or with three or more methods), and the variable is ordinal, or numerical without a normal distribution.

7.6.1 The Logic of the Test

The difference from Kruskal-Wallis is that the ranks are assigned **within each subject**, not in the combined sample. For each patient, the k measurements receive ranks 1, 2,..., k. In this way, the differences between patients (one has blood glucose values around 90, another around 140) disappear, and only the order of the measurements within the same patient remains. If there is no effect of time, each time point should receive, on average, the rank (k + 1)/2.

H0: the distribution of values is the same at all k time points H1: at least one time point differs from the others

With n subjects and k time points, and Rj the sum of the ranks for time point j:

$$\chi^2 F = \left[\frac{12}{n \times k \times (k+1)}\right] \times \sum {R_j}^2 - 3n(k+1)$$

The statistic approximately follows a chi-square distribution with k - 1 degrees of freedom.

7.6.2 The Hand Calculation

Six patients with low back pain rated their pain (0-10) at the start of a physiotherapy program, at 2 weeks and at 6 weeks.

Table 7.6. Ranks within each patient for the Friedman test (fictitious data).

Patient	Start	2 wk	6 wk	Ranks (start; 2 wk; 6 wk)
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
Rank sum Rj				17,5; 12; 6,5

Check: each row has a rank sum of 1 + 2 + 3 = 6, so the total is 6 x 6 = 36 = 17.5 + 12 + 6.5. Correct.

$$\chi^2 F = \left[\frac{12}{6 \times 3 \times 4}\right] \times (17.5^2 + 12^2 + 6.5^2) - 3 \times 6 \times 4$$
$$= \left(\frac{12}{72}\right) \times (306.25 + 144 + 42.25) - 72 = 0.1667 \times 492.5 - 72 = 82.08 - 72$$
$$= 10.08$$

With 2 degrees of freedom, the critical value is 5.99. We reject H0. Here too, software applies a correction for the two ties (patients 3 and 5) and reports 11.00, with p = 0.004. Pain decreased significantly during the program.

7.6.3 Application to mg2026: Blood Glucose at 0, 3 and 6 Months

Fasting plasma glucose was measured at enrollment, at 3 months and at 6 months in all 160 patients. The medians are 104.5, 99.0 and 92.5 mg/dL (means 105.4, 99.4 and 95.2). Figure 7.3 shows the median and interquartile range by arm.

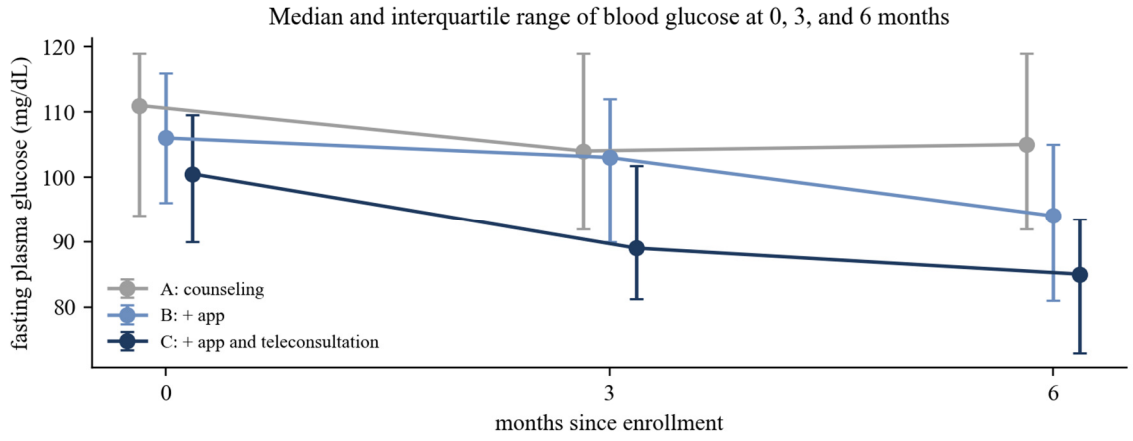


Figure 7.3. Median and interquartile range of fasting plasma glucose at 0, 3 and 6 months, by arm. In arm C the decrease is the steepest and continues up to 6 months; in arm A the median decreases at 3 months and then stays flat. Synthetic data (mg2026 dataset).

The Friedman test on all patients: the rank sums are 406 (0 months), 295 (3 months) and 259 (6 months), that is, mean ranks of 2.54, 1.84 and 1.62. Without correction, $\chi^2 F = \left[\frac{12}{160 \times 3 \times 4}\right] \times (406^2 + 295^2 + 259^2) - 3 \times 160 \times 4 = 73.39$; with the correction for ties, 74.67, df 2, p < 0.001. For comparison, repeated-measures ANOVA gives F = 64.97, p < 0.001. Both tests say the same thing.

The effect size for Friedman is Kendall's coefficient of concordance W:

$$W = \frac{\chi^2 F}{n(k-1)} = \frac{74.67}{160 \times 2} = 0.23$$

W ranges from 0 (no common trend) to 1 (all patients have exactly the same order of measurements). A value of 0.23 indicates a moderate common trend: many patients show decreases, but not all in the same order.

Post hoc comparisons are made with the Wilcoxon test for each pair of time points, with Bonferroni correction (3 comparisons, so each p is multiplied by 3 or the threshold becomes 0.05/3 = 0.0167):

Table 7.7. Post hoc comparisons after the Friedman test (blood glucose, n = 160).

Comparison	Median of the differences (mg/dL)	Wilcoxon p	Bonferroni-adjusted p
0 months versus 3 months	7,0	< 0,001	< 0,001
3 months versus 6 months	4,0	0,0002	0,0007
0 months versus 6 months	9,5	< 0,001	< 0,001

All three comparisons remain significant after correction. The analysis by arm is more instructive: in arm C, Friedman gives $\chi^2F = 64.63$, $p < 0.001$ (medians 100.5, 89.0 and 85.0 mg/dL), in arm B $\chi^2F = 21.07$, $p < 0.001$, and in arm A only $\chi^2F = 6.04$, $p = 0.049$ (medians 111, 104 and 105 mg/dL). In arm A the overall result just crosses the 0.05 threshold, but none of the Wilcoxon comparisons remains significant after the Bonferroni correction (the smallest adjusted p is 0.075, for 0 versus 6 months). This situation is common in practice: the global test "sees" a weak trend, but the power is not enough to locate the difference after correction.

The Friedman test also appears in studies comparing measurement methods. Al Barri and colleagues, with Dr. Tudor, compared three keratometry and ocular biometry devices in patients who were candidates for cataract surgery: 170 eyes from 102 patients, mean age 69.12 ± 10.26 years [81]. Each eye was measured with all three devices, so we have three related measurements, exactly the situation for the Friedman test, followed by pairwise Wilcoxon comparisons. For K1 keratometry, the means were 43.45 ± 1.64 D (Argos) and 43.41 ± 1.70 D (Aladdin), with $p < 0.001$ overall and $p = 0.019$ for Argos versus Aladdin. Anterior chamber depth was 3.19 ± 0.42 versus 3.13 ± 0.41mm, $p < 0.001$, and corneal thickness was practically identical between Aladdin and Sirius (540.27 ± 33.44 versus 540.47 ± 33.78 µm, $p = 0.169$). The vector components of astigmatism J0 ($p = 0.277$) and J45 ($p = 0.084$) did not differ. The article shows very well the difference between statistical and clinical significance: a difference of 0.04 D in K1 or of 0.06 mm in anterior chamber depth can be "significant", because paired measurements remove the variability between eyes, but it is far below the threshold that would change the calculation of the intraocular lens. The second observation concerns independence: the 170 eyes come from 102 patients, so some patients contribute two eyes, and the eyes of the same patient are not completely independent. We return to this topic in Chapter 8, under the unit of analysis.

7.7 The Chi-Square Test for Contingency Tables

So far we have worked with numerical or ordinal variables. When both variables are qualitative (nominal, or ordinal with few categories) and we want to know whether they are associated, the data are summarized in a contingency table, and the classic tool is the chi-square (χ^2) test proposed by Karl Pearson in 1900 [98]. It is probably the test most often encountered in medical articles, because many clinical outcomes are of the "yes/no" type: cured or not, hospitalized or not, with or without a complication.

7.7.1 The Contingency Table and the Hypotheses

An r x c contingency table has r rows (the categories of one variable) and c columns (the categories of the other). Each cell contains the number of subjects who simultaneously have the category of the row and that of the column. The row and column totals are called marginal totals.

The working example comes from the cross-sectional analysis of the mg2026 patients at enrollment: is there an association between obesity (BMI ≥ 30 kg/m²) and diabetes confirmed by the oral glucose tolerance test (OGTT)?

Table 7.8. Obesity and OGTT-confirmed diabetes at enrollment, mg2026 dataset (observed frequencies).

	OGTT diabetes yes	**OGTT diabetes no**	**Total**
Obese (BMI ≥ 30)	18	36	54
Non-obese (BMI < 30)	11	95	106
Total	29	131	160

The prevalence of diabetes is 18/54 = 33.3% in obese patients and 11/106 = 10.4% in non-obese patients. The hypotheses:

H0: obesity and diabetes are independent (the prevalence of diabetes is the same in obese and non-obese patients: p1 = p2) H1: there is an association between obesity and diabetes (p1 ≠ p2)

7.7.2 Expected Frequencies

The idea of the test is to compare what we observed with what we would expect to observe if H0 were true. If obesity has nothing to do with diabetes, then the overall proportion of patients with diabetes, 29/160 = 18.1%, should be found equally among obese and non-obese patients. Among the 54 obese patients we would then expect $\frac{54\times29}{160} = 9.79$ patients with diabetes. The general rule for any cell:

E = (row total x column total)/grand total

Table 7.9. Expected frequencies under the hypothesis of independence.

	OGTT diabetes yes	**OGTT diabetes no**	**Total**
Obese	$\frac{54 \times 29}{160} = 9.79$	$\frac{54 \times 131}{160} = 44.21$	54
Non-obese	$\frac{106 \times 29}{160} = 19.21$	$\frac{106 \times 131}{160} = 86.79$	106
Total	29	131	160

Check: the expected frequencies have the same marginal totals as the observed ones (9.79 + 44.21 = 54; 9.79 + 19.21 = 29). If the totals do not match, you have made a calculation error. Expected frequencies can be decimals; they are not rounded to whole numbers.

7.7.3 The Statistic, the Degrees of Freedom, the Decision

The chi-square statistic adds up, over all cells, the relative squared deviation between observed (O) and expected (E):

$$\chi^2 = \frac{\sum(O-E)^2}{E}$$

Table 7.10. Contributions of the cells to χ^2 (obesity x diabetes).

Cell	**O**	**E**	**O - E**	$\frac{(O-E)^2}{E}$
obese, diabetes	18	9,79	8,21	6,891
obese, no diabetes	36	44,21	-8,21	1,525
non-obese, diabetes	11	19,21	-8,21	3,510
non-obese, no diabetes	95	86,79	8,21	0,777
Total	160	160	0	$\chi^2 = 12.70$

Note that in a 2 x 2 table all the O - E deviations have the same absolute value (here 8.21). For 2 x 2 tables there is also a short formula, useful in exams:

$$\chi^2 = \frac{N(\text{ad} - \text{bc})^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)}$$

= 160 x (18 x 95 - 36 x 11)²/(54 x 106 x 29 x 131) = 160 x 1314²/21,745,116 = 12.70

The number of degrees of freedom for an r x c table is:

$$\text{df} = (r-1)\,(c-1)$$

For 2 x 2, df = 1. The reasoning: once the marginal totals are fixed, only one cell can be chosen freely; the others follow by subtraction. The critical value for df = 1 and α = 0.05 is 3.84 (for df = 2 it is 5.99; for df = 3, 7.81; for df = 4, 9.49). Since 12.70 > 3.84, we reject H0; the software gives p = 0.0004 (we report p < 0.001).

The statistical conclusion: there is a significant association between obesity and OGTT-confirmed diabetes. The clinical conclusion: in this population at cardiometabolic risk, diabetes was about three times more common in obese patients (33.3% versus 10.4%; prevalence ratio 3.21, 95% CI 1.64-6.31; OR 4.32, 95% CI 1.86-10.03). As this is a cross-sectional study, we cannot say that obesity "caused" diabetes; we only say that the two occur together more often than would be expected by chance. Risk measures and study designs are covered in detail in Chapter 10.

The chi-square test is always a "two-sided" test by construction: it detects deviations in any direction. We read the direction of the association from the table (the row percentages), not from the χ^2 value.

7.7.4 A 3 x 2 Table on mg2026: Arm and Weight Loss of at Least 5%

The secondary objective of the mg2026 study was the proportion of patients who lost at least 5% of their initial weight in 6 months. Figure 7.4 shows the proportions by arm.

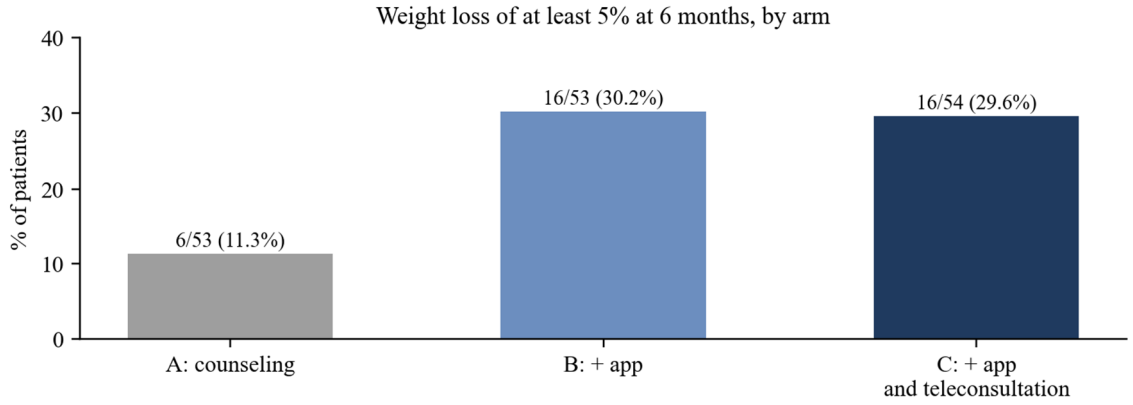


Figure 7.4. Proportion of patients with weight loss of at least 5% at 6 months, by arm: 6/53 (11.3%) in arm A, 16/53 (30.2%) in arm B and 16/54 (29.6%) in arm C. Chi-square test on the 3 x 2 table: χ^2 = 6.77, df 2, p = 0.034. Synthetic data (mg2026 dataset).

Table 7.11. Study arm and weight loss ≥ 5%: observed frequencies, expected frequencies (in parentheses) and contributions to χ^2.

Arm	Loss ≥ 5%	No loss ≥ 5%	Total	% with loss
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

The expected frequencies are calculated in the same way: for arm A and the column "loss ≥ 5%", $E = \frac{53 \times 38}{160} = 12.59$. The sum of the contributions is χ^2 = 3.447 + 1.074 + 0.925 + 0.288 + 0.786 + 0.245 = 6.77. The degrees of freedom: (3 - 1)(2 - 1) = 2. The critical value of 5.99 is exceeded, p = 0.034. The smallest expected frequency is 12.6, so the assumptions are met.

The table of contributions also tells us **where** the association lies: almost half of the χ^2 (3.447 of 6.77) comes from a single cell, arm A with loss ≥ 5%, where we observed 6 patients instead of the 12.6 expected. Arms B and C are very close to each other.

A significant χ^2 in a table larger than 2 x 2 is, as with ANOVA, a global result. If we want to compare the arms pairwise, we build 2 x 2 tables and correct for multiple comparisons. Without correction: A versus B χ^2 = 5.74, p = 0.017; A versus C χ^2 = 5.49, p = 0.019; B versus C χ^2 = 0.00, p = 0.95. With Bonferroni for three comparisons, the threshold becomes 0.0167, and A versus B and A versus C end up right at the limit. For the main question of the study, which had been specified in the protocol as the comparison of C versus A, we do not need a correction, and the result is most usefully expressed through the relative risk: 29.6% versus 11.3%, RR 2.62 (95% CI 1.11-6.17), absolute difference 18.3 percentage points (95% CI 3.4-33.2), NNT about 6.

7.7.5 Assumptions

The chi-square test is an approximate test: the distribution of the statistic approaches the theoretical χ^2 distribution only when the expected frequencies are large enough. The rules used in practice [42,53]:

- the observations are independent: each subject appears in a single cell, only once;
- we work with absolute frequencies (numbers of subjects), never with percentages or means;
- for 2 x 2 tables, all expected frequencies must be at least 5;
- for larger tables, at most 20% of the cells may have an expected frequency below 5 (the 80% rule) and no cell may have an expected frequency below 1.

Note: the condition concerns the **expected** frequencies, not the observed ones. A cell with 0 observations can be perfectly acceptable if the expected frequency there is 6. When the condition is not met, we have three solutions: Fisher's exact test (for 2 x 2 tables and, in many programs, for larger tables), merging some categories in a clinically meaningful way (for example "fost fumator" (former smoker) with "fumator" (smoker), if the question allows it) or increasing the sample size.

Unequal group size is not, in itself, a problem for the chi-square test; only the expected frequencies matter. An example from our colleagues' research: Marin and colleagues, with Dr. L. Moleriu, retrospectively compared the virological response in 462 patients with HIV-1 subtype F treated with darunavir boosted either with ritonavir (384 patients) or with cobicistat (78 patients) [99]. The abstract reports a percentage of 84.6% in the cobicistat group, compared with 76.8%. The groups are very unequal (384 and 78), but with dozens of patients in each cell the expected frequencies are large, and comparing two proportions with chi-square (or Fisher, for cells with few events) is perfectly valid. The real problem with such a design is a different one: in an observational study, the physician chose the regimen according to the patient's characteristics, so the difference may reflect the type of patients treated with each regimen (confounding by indication), not only the effect of the drug.

7.7.6 The Yates Correction

For 2 x 2 tables, Frank Yates proposed a continuity correction: 0.5 is subtracted from each absolute deviation before squaring.

$$\chi^2\text{Yates} = \frac{\sum(|O-E|-0.5)^2}{E}$$

The correction compensates for the fact that the observed frequencies are whole numbers, whereas the χ^2 distribution is continuous. In the obesity x diabetes example, χ^2Yates = 11.20 (p = 0.0008), compared with 12.70 without correction. The correction makes the test more conservative. Many statisticians today consider that it overcorrects and prefer to go straight to Fisher's test for small tables; JASP offers it as an option ("continuity correction"), and Epi Info displays it automatically alongside the uncorrected version. The practical rule for the exam: for a 2 x 2 table with expected frequencies between 5 and 10, report the Yates-corrected version or Fisher's test; when all expected frequencies are large, the difference between the versions does not change the conclusion.

7.7.7 Fisher's Exact Test

Fisher's exact test uses no approximation. Treating the marginal totals as fixed, it calculates the exact probability of the observed table and of all tables "at least as extreme", using the hypergeometric distribution. For a table with cells a, b, c, d and N subjects:

$$P = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{N!\ a!\ b!\ c!\ d!}$$

Example: in an audit of postoperative infections, 8 patients received a new antibiotic prophylaxis regimen (1 infection) and 10 patients the old regimen (6 infections).

Table 7.12. Postoperative infection by prophylaxis regimen (fictitious data).

	Infection	No infection	Total
New regimen	1	7	8
Old regimen	6	4	10
Total	7	11	18

The expected frequencies in the "infection" column are $\frac{8\times7}{18} = 3.11$ and $\frac{10\times7}{18} = 3.89$, both below 5. The chi-square test is not valid here, although the software calculates it without complaint: χ^2 = 4.22, p = 0.040. Fisher's test calculates the probability of each possible table with the same totals, according to the number a of infections with the new regimen (a can range from 0 to 7): P(a = 0) = 0.0038; P(a = 1) = 0.0528 (our table); 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. The two-sided p value adds up the probabilities of the tables that are at most as probable as the observed table: 0.0038 + 0.0528 + 0.0088 + 0.0003 = 0.066.

The conclusion changes: with the appropriate test (Fisher), p = 0.066, we do not reject H0. The chi-square test used wrongly would have given a "significant" result (0.040), and the Yates correction (p = 0.117) goes in the same direction as Fisher. With only 18 patients, we cannot claim that the new regimen reduces infections, although the observed difference (12.5% versus 60%) deserves a larger study. Calculating Fisher's test by hand is laborious; in the exam you are usually asked to recognize when it must be used and to interpret the p value given by the software.

Case study: small tables in the emergency department, in patients with stroke and long COVID

Popa and colleagues, with Dr. Tudor, retrospectively analyzed 132 patients who arrived at the Emergency Department with a stroke alert, of whom 26 had long COVID and 106 did not [100]. The group of 26 is small, and some symptoms are rare, so several 2 x 2 tables have small expected frequencies. The authors used Fisher's exact test (the abstract contains the wording "Fisher's Exact Test chi-squared", p = 0.027, which mixes up the names of two different tests). Headache occurred in 19.2% versus 5.7% (OR = 3.97, p = 0.040), and facial

asymmetry in 42.3% versus 19.8% (OR = 2.97, p = 0.022). Let us do a rough calculation from the percentages: 19.2% of 26 means about 5 patients, and 5.7% of 106 about 6 patients, so about 11 patients with headache in total. The expected number of patients with headache in the long COVID group would then be 26 x 11/132, that is, about 2.2, well below 5: Fisher's test was exactly the right choice. Mechanical thrombectomy was more common in the long COVID group (30.8% versus 16.0%), and another comparison gave 26.9% versus 14.2%, p = 0.143, without statistical significance. With only 26 patients in one group, a difference of almost 13-15 percentage points can remain nonsignificant for lack of power; here "nonsignificant" does not mean "no difference". What would have completed the reporting: the confidence intervals of the ORs, which would show how imprecise the estimates obtained from small groups are.

7.8 Variants of the Chi-Square Test

7.8.1 McNemar's Test for Paired Proportions

The classic chi-square test assumes independent groups. When the same dichotomous variable is measured twice on the same subjects (before and after, or with two diagnostic tests applied to the same patient), the data are paired and McNemar's test is used. It is the equivalent, for binary variables, of the paired t-test.

The table is built differently: the rows are the status at the first time point, the columns the status at the second time point, and each patient appears only once. Patients who did not change status (the cells on the diagonal) provide no information about change. The test uses only the discordant cells, b and c:

χ^2McNemar = (b - c)²/(b + c), with 1 degree of freedom; with continuity correction: (|b - c| - 1)²/(b + c)

Application to mg2026: how many patients had fasting plasma glucose ≥ 100 mg/dL (impaired fasting glucose or higher) at enrollment, and how many at 6 months?

Table 7.13. Fasting plasma glucose ≥ 100 mg/dL at enrollment and at 6 months, the same 160 patients (mg2026 dataset).

	At 6 months ≥ 100	**At 6 months < 100**	**Total**
At enrollment ≥ 100	61	36 (b)	97
At enrollment < 100	1 (c)	62	63
Total	62	98	160

The proportion with blood glucose ≥ 100 decreased from 97/160 = 60.6% to 62/160 = 38.8%. The discordant cells: 36 patients moved below 100, and a single patient moved above 100. $\chi^2 = \frac{(36-1)^2}{36+1} = \frac{1225}{37} = 33.11$; with correction, (35 - 1)²/37 = 31.24; p < 0.001. It would be wrong to apply the ordinary chi-square test here to a "time point x status" table with 320 observations: the same patient would appear twice, and independence would be violated. In our colleagues' research, Iovanescu and colleagues, with Dr. Frandeş, used McNemar's test in a colorectal neoplasia screening study in which several investigations (the fecal immunochemical test, sigmoidoscopy, colonoscopy) were evaluated in the same individuals (McNemar, P < 0.001); 252 individuals subsequently underwent colonoscopy, and advanced neoplasia was found in 75 of them (29.76%) [101]. When two diagnostic tests are applied to the same individuals, comparing their proportions of positive results requires a test for paired data.

A less spectacular example: systolic blood pressure ≥ 140 mmHg at enrollment and at 6 months. 16 patients moved below 140, and 8 moved above 140. χ^2McNemar = (16 - 8)²/24 = 2.67, p = 0.10 (with correction 2.04, p = 0.15). Although the proportion with SBP ≥ 140 decreased from

41/160 to 33/160, the change is not statistically significant, because it is based on only 24 discordant patients.

7.8.2 The Chi-Square Test for Trend

When one of the variables has ordered categories (age groups, stages, doses, or arms with increasing intensity of intervention), the ordinary chi-square test ignores the order: it would give the same result if we shuffled the rows. The test for trend (Cochran-Armitage) asks a more precise question: does the proportion of "yes" increase (or decrease) with the category? Each category is assigned a score (for example 0, 1, 2), and the test has a single degree of freedom, which makes it more powerful than the global test when there is a monotonic trend.

In mg2026, arms A, B and C have increasing intensity (counseling; + app; + app and teleconsultation). The test for trend with scores 0, 1, 2 gives χ^2trend = 4.92, df 1, p = 0.027. The proportions are 11.3%, 30.2% and 29.6%: the trend exists, but it is driven almost entirely by the jump from A to B; from B to C the proportion no longer increases. The test for trend is available in Epi Info (StatCalc, "Chi square for trend") and in MedCalc.

7.8.3 The Chi-Square Goodness-of-Fit Test

The goodness-of-fit test compares the distribution of a single qualitative variable with a theoretical distribution specified in advance. The expected frequencies no longer come from the marginal totals but from the hypothesis: E = N x theoretical proportion. The degrees of freedom are k - 1, where k is the number of categories (if no parameter was estimated from the data).

The first example: randomization in mg2026 aimed at equal allocation to the three arms. Observed 53, 53, 54; expected 53.33 in each. χ^2 = 0.0125, df 2, p = 0.99: the allocation matches the 1:1:1 ratio very well. The second: the sample contains 86 women and 74 men. Against a theoretical 50:50 ratio (80 and 80), $\chi^2 = \frac{(86-80)^2}{80} + \frac{(74-80)^2}{80} = 0.45 + 0.45 = 0.90$, df 1, p = 0.34. The third: suppose that, in a regional registry, the distribution of educational level among adults aged 40-70 years were 15% lower secondary school, 35% high school or vocational school, 35% university and 15% postgraduate (hypothetical proportions, used only for the exercise). Observed in mg2026: 20, 57, 64, 19; expected 24, 56, 56, 24. χ^2 = 0.67 + 0.02 + 1.14 + 1.04 = 2.87, df 3, p = 0.41. The sample does not differ significantly from the reference distribution, which supports, without proving, its representativeness for this characteristic.

The test for trend arises naturally when the groups have a natural order. Piroș and colleagues, with Dr. Lupușoru, Dr. L. Moleriu and Dr. Muntean, conducted a cross-sectional study of 693 students aged 18-24 years [6]. The prevalence of prediabetes was 21.1% (diabetes 1.4%). The students were grouped by cluster analysis into three dietary patterns: health-oriented (prediabetes 15.4%), mixed (20.0%) and fast-food-oriented (27.3%), and the chi-square test on the 3 x 2 table gave p = 0.003. The categories are ordered by diet quality, and the authors went further exactly in the direction discussed above: a test for trend (p-trend < 0.001) and a logistic regression in which each pattern category increased the odds of prediabetes with OR = 1.78 per category (95% CI 1.38-2.30; p < 0.001). The correct reading: there is a graded association between dietary pattern and prediabetes, compatible with a dose-response relationship. As this is a cross-sectional study, the association does not demonstrate causality, and the clusters were built from the same data, so the result must be confirmed in another population.

7.9 Beyond the p Value: Phi, Cramér's V and Other Measures of Association

The χ^2 value and the p value increase with sample size. If we doubled all the frequencies in a table, the percentages would remain the same, but χ^2 would double and p would decrease. That is why χ^2 and p do not tell us how **strong** the association is. For that, we use measures that do not depend on N.

For a 2 x 2 table, the phi coefficient:

$$\varphi = \sqrt{\frac{\chi^2}{N}}$$

For larger tables, Cramér's V coefficient:

$$V = \sqrt{\left[\frac{\chi^2}{N \times (m-1)}\right]}$$, where m is the smaller of r and c

Both take values between 0 (independence) and 1 (perfect association). As a rough guide, for tables with m - 1 = 1, 0.1 means a weak association, 0.3 moderate and 0.5 strong [76]. On mg2026:

Table 7.14. The chi-square test and the strength of association for several tables from mg2026.

Table	χ^2	df	p	φ or V
obesity x OGTT diabetes (2 x 2)	12,70	1	< 0,001	$\varphi = 0.28$
arm x loss ≥ 5% (3 x 2)	6,77	2	0,034	$V = 0.21$
sex x smoking status (2 x 3)	0,54	2	0,76	$V = 0.06$
age ≥ 50 years x OGTT diabetes (2 x 2)	4,36	1	0,037	$\varphi = 0.17$

The last row deserves a comment. Diabetes occurred in 2 of 34 patients under 50 years of age (5.9%) and in 27 of 126 patients aged 50 years or over (21.4%). The uncorrected test gives p = 0.037, Fisher p = 0.044, and the Yates correction p = 0.066, because one of the expected frequencies is only 6.2. The result is borderline, and the association, measured by φ = 0.17, is weak. The honest conclusion: a signal compatible with an increase in the risk of diabetes with age, but fragile in this sample.

For 2 x 2 tables from clinical and epidemiological studies, the most clinically useful measures are not φ or V but those with interpretable units: the difference in proportions (with 95% CI), the relative risk or the prevalence ratio, the odds ratio and the number needed to treat. The chi-square test answers the question "is there an association?", and these measures answer "how large is it and in which direction?". A good article reports both. Chapter 10 is devoted to them.

> **Case study: convenience eating behaviors and BMI classes in students**
>
> Moleriu and colleagues, with Dr. L. Moleriu, Dr. Lupușoru and Dr. Muntean, analyzed 921 students aged 18-24 years [4]. Of these, 24.1% were overweight or obese (19.4% overweight, 4.7% obese). The methods combine exactly the tools from this chapter and from Chapter 10: the chi-square test for the association between eating behaviors and BMI class, Cramér's V for the strength of these associations, then multivariable logistic regression with adjusted odds ratios (for example aOR 7.20, 95% CI 4.26-12.17 and aOR 1.98, 95% CI 1.26-3.10). A composite dietary risk score shows a dose-response relationship: the proportion with excess weight increases from 19.8% at a score of 0 to 50.0% at a score of 4, with an adjusted OR of 1.43 for each point (95% CI 1.23-1.66; p < 0.001). In a sample of 921 people, even weak associations can come out statistically significant, and Cramér's V is precisely the

> measure that helps us tell the weak ones from the strong ones. The design is cross-sectional: the association does not establish the causal direction, and some behaviors may be the consequence, not the cause, of excess weight.

7.10 Reporting the Results of Nonparametric Tests and the Chi-Square Test

A well-reported result allows the reader to retrace the reasoning. For rank-based tests, the description uses the median and interquartile range (or minimum-maximum, for very small samples), not the mean and standard deviation. Then the name of the test, the statistic, the exact p value (to three decimal places; below 0.001, write "p < 0.001") and, ideally, an effect size are given. For chi-square, give the absolute frequencies and percentages (for the rows corresponding to the groups compared), χ^2, the degrees of freedom, p and a measure of effect with a confidence interval.

A few templates, using the data from this chapter:

- "Satisfaction at 6 months was higher in arm C than in arm A (median 4, IQR 3-5, versus 3, IQR 2-3; Mann-Whitney U = 2265.5, p < 0.001; rank-biserial correlation 0.58)."
- "The decrease in fasting plasma glucose differed between arms (Kruskal-Wallis H = 36.63, df 2, p < 0.001, ε^2 = 0.23); Dunn's comparisons with Bonferroni correction showed differences between A and B (p < 0.001) and between A and C (p < 0.001), but not between B and C (p = 0.12)."
- "Fasting plasma glucose decreased during the study (medians 104.5, 99.0 and 92.5 mg/dL at 0, 3 and 6 months; Friedman χ^2 = 74.67, df 2, p < 0.001; Kendall's W = 0.23)."
- "The proportion of patients with weight loss ≥ 5% differed between arms (11.3%, 30.2% and 29.6%; χ^2 = 6.77, df 2, p = 0.034; Cramér's V = 0.21)."

What we do not do: we do not write "p = 0.000" (the software rounds, but the probability is not zero), we do not report only "p < 0.05" when we have the exact value, we do not present the mean of Likert scores as if they were physical measurements, and we do not write "the test was nonsignificant, so the groups are equal".

An example of correct reporting of ordinal scores comes from a study that concerns you directly. Lungeanu and colleagues, with Dr. Lupușoru, asked 121 medical students, through an anonymous questionnaire, how they rated the courses related to digital skills and data science [16]. The ratings, on a scale from 1 to 10, are ordinal variables, and the authors reported them exactly as this chapter recommends: median and IQR. One course had a median of 9 (IQR 7-10), the medical informatics course 7 (5-9), and biostatistics 8 (5-9). The relationships between ratings were assessed with Spearman correlation, which also works on ranks: a very strong correlation, R = 0.869 (p < 0.001), and a practically absent correlation between measured and self-assessed knowledge, R = 0.107 (p = 0.246). The second result is probably the most interesting one for you: the fact that a student feels prepared says almost nothing about how much they actually know. Here, a well-interpreted "nonsignificant" result is valuable information, not a failure of the study.

7.11 Worked Problems

For hand calculations on small samples you need the critical values of the Mann-Whitney and Wilcoxon tests, which are not found in the classic distribution tables in the appendix. Tables 7.15 and 7.16 give them for the two-sided test at α = 0.05; they were calculated from the exact distributions of the statistics. The decision rule is the same for both: **H0 is rejected if the calculated statistic is less than or equal to the value in the table**. A dash means that, for those group sizes, no value can reach significance at this level.

Table 7.15. Critical values of the Mann-Whitney U (two-sided test, α = 0.05); n1 and n2 are the sizes of the two groups.

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

Table 7.16. Critical values of the Wilcoxon T for paired data (two-sided test, α = 0.05); n is the number of nonzero differences.

n	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Critical T	0	2	3	5	8	10	13	17	21	25	29	34	40	46	52

Worked Problem 7.1. Length of Hospital Stay in Pneumonia, Mann-Whitney by Hand

In an internal medicine ward, 6 patients with community-acquired pneumonia were switched early from intravenous to oral antibiotics, and 7 patients received the standard intravenous regimen. The length of hospital stay (days) was: early switch 4, 5, 5, 6, 8, 12; standard regimen 6, 7, 7, 9, 10, 11, 15. Does the length of stay differ between the two strategies (α = 0.05)?

Research hypothesis. An early switch to oral antibiotics shortens the hospital stay.

Statistical hypotheses. H0: the lengths of stay come from the same distribution in the two groups (Me1 = Me2). H1: the distributions differ (Me1 ≠ Me2), two-sided test.

Choice of test. The variable is numerical, but length of hospital stay is typically skewed (the values of 15 days and 12 days pull the mean), the groups are very small (6 and 7) and independent. We choose the Mann-Whitney U test.

Calculation. The combined sorted series (N = 13): 4, 5, 5, 6, 6, 7, 7, 8, 9, 10, 11, 12, 15. The ranks: 4 → 1; 5 and 5 → 2.5; 6 and 6 (one from each group) → 4.5; 7 and 7 → 6.5; 8 → 8; 9 → 9; 10 → 10; 11 → 11; 12 → 12; 15 → 13.

R1 (early switch) = 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

Check: $30.5 + 60.5 = 91 = \frac{13\times14}{2}$.

$U_1 = 30.5 - \frac{6\times7}{2} = 9.5$; $U_2 = 60.5 - \frac{7\times8}{2} = 32.5$; check 9.5 + 32.5 = 42 = 6 x 7. U = 9.5.

The critical value for n1 = 6, n2 = 7, two-sided, α = 0.05 is 6. Since 9.5 > 6, we do not reject H0. The exact p value is 0.138. Effect size: $r_{\text{rb}} = 1 - \frac{2\times9.5}{42} = 0.55$.

Statistical conclusion. There is no statistically significant difference between the distributions of length of hospital stay (U = 9.5, p = 0.138).

Clinical conclusion. The medians are 5.5 days versus 9 days, and the estimated effect is large (0.55), but the study is too small to confirm it. The result does not show that the strategies are equivalent; it justifies a study with more patients, whose sample size is calculated in advance (Chapters 4 and 5).

Worked Problem 7.2. Pain Before and After Physiotherapy, Wilcoxon and the Sign Test

Ten patients with knee osteoarthritis rated their pain (0-10) before and after 4 weeks of physiotherapy: before 7, 6, 8, 5, 9, 6, 7, 8, 6, 7; after 4, 5, 5, 4, 6, 6, 3, 5, 4, 7. Did the pain decrease?

Hypotheses. H0: the median of the differences (before - after) is 0. H1: the median of the differences ≠ 0.

Choice of test. The same patients, two time points, hence paired data; the score is ordinal. We choose the Wilcoxon test for paired data.

Calculation. The differences: 3, 1, 3, 1, 3, 0, 4, 3, 2, 0. Two differences are zero and are dropped, n = 8. The sorted absolute values: 1, 1 (ranks 1 and 2, so 1.5 each); 2 (rank 3); 3, 3, 3, 3 (ranks 4, 5, 6, 7, so 5.5 each); 4 (rank 8). All nonzero differences are positive, so W+ = 1.5 + 1.5 + 3 + 4 x 5.5 + 8 = 36 and W- = 0. Check: $36 + 0 = \frac{8\times9}{2}$. T = 0.

The critical value for n = 8, two-sided, α = 0.05 is 3. Since 0 ≤ 3, we reject H0; exact p = 0.008. The sign test gives the same result here: 8 pluses out of 8, p = 2 x 0.5□ = 0.008.

Conclusions. Pain decreased significantly after physiotherapy (median from 7 to 5; median of the differences 2.5 points; Wilcoxon T = 0, p = 0.008). Clinically, a decrease of 2-3 points on the 0-10 scale is usually perceived by the patient as relevant. The lack of a control group, however, does not allow the effect of physiotherapy to be separated from the natural course or from the placebo effect.

Worked Problem 7.3. Wound Healing with Three Types of Dressing, Kruskal-Wallis

The number of days to complete healing of a superficial wound was recorded in 14 patients: conventional dressing (n = 4) 12, 14, 15, 18; hydrocolloid dressing (n = 5) 9, 10, 11, 13, 14; silver dressing (n = 5) 7, 8, 8, 10, 12. Does healing time differ between dressings?

Hypotheses. H0: the distribution of healing time is the same for the three dressings. H1: at least one dressing differs.

Choice of test. Three independent groups, very small samples, a numerical variable of the "duration" type: Kruskal-Wallis.

Calculation. The combined sorted series (N = 14): 7, 8, 8, 9, 10, 10, 11, 12, 12, 13, 14, 14, 15, 18. The ranks: 7 → 1; 8 and 8 → 2.5; 9 → 4; 10 and 10 → 5.5; 11 → 7; 12 and 12 → 8.5; 13 → 10; 14 and 14 → 11.5; 15 → 13; 18 → 14.

R1 (conventional) = 8.5 + 11.5 + 13 + 14 = 47 R2 (hydrocolloid) = 4 + 5.5 + 7 + 10 + 11.5 = 38 R3 (silver) = 1 + 2.5 + 2.5 + 5.5 + 8.5 = 20

Check: $47 + 38 + 20 = 105 = \frac{14\times15}{2}$.

$$H = \left[\frac{12}{14 \times 15}\right] \times \left(\frac{47^2}{4} + \frac{38^2}{5} + \frac{20^2}{5}\right) - 3 \times 15 = 0.05714 \times (552.25 + 288.8 + 80) - 45$$
$$= 0.05714 \times 921.05 - 45 = 7.63$$

With the correction for ties, H = 7.70. With 2 degrees of freedom, 7.70 > 5.99, so we reject H0 (p = 0.021). Note: one group has only 4 patients, so the chi-square approximation is at the limit of acceptability; software that offers the exact version should be preferred.

Post hoc (Dunn, Bonferroni). The mean ranks are 11.75 (conventional), 7.6 (hydrocolloid) and 4.0 (silver). Only the conventional versus silver difference remains significant after correction (z = 2.76, adjusted p = 0.017); conventional versus hydrocolloid adjusted p = 0.42, hydrocolloid versus silver adjusted p = 0.52.

Conclusions. Healing time differed between dressings (medians 14.5, 11 and 8 days; H = 7.70, p = 0.021), and the difference lies between the conventional dressing and the silver dressing. With 14 patients, the result is only indicative.

Worked Problem 7.4. Fatigue After COVID-19 at Three Time Points, Friedman

Seven patients with persistent symptoms after COVID-19 rated their fatigue (0-10) at 1, 3 and 6 months after infection:

Table 7.17. Fatigue score and ranks within each patient (fictitious data).

Patient	1 month	3 months	6 months	Ranks
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

Hypotheses. H0: the distribution of the fatigue score is the same at the three time points. H1: at least one time point differs.

Choice of test. The same patients, three time points, ordinal score: Friedman.

Calculation. Check: 20.5 + 14 + 7.5 = 42 = 7 x 6.

$$\chi^2 F = \left[\frac{12}{7 \times 3 \times 4}\right] \times (20.5^2 + 14^2 + 7.5^2) - 3 \times 7 \times 4$$
$$= 0.1429 \times (420.25 + 196 + 56.25) - 84 = 0.1429 \times 672.5 - 84 = 12.07$$

With the correction for ties, 13.00; df 2; 13.00 > 5.99, so we reject H0 (p = 0.0015). The medians: 8, 6 and 5.

Post hoc. Wilcoxon for each pair: 1 versus 3 months p = 0.031; 3 versus 6 months p = 0.031; 1 versus 6 months p = 0.016. With Bonferroni (x 3): 0.094, 0.094 and 0.047. Only the 1 versus 6 months comparison remains significant. With 7 patients, the smallest p value the exact two-sided Wilcoxon test can give is 0.016, so the post hoc comparisons have very limited power.

Conclusions. Fatigue decreased significantly in the first 6 months (Friedman χ^2 = 13.00, p = 0.0015); the difference between 1 and 6 months is the one that survives correction. Without a comparison group, we cannot attribute the decrease to any treatment.

Worked Problem 7.5. Smoking and COPD, a Complete 2 x 2 Chi-Square Test

In a cross-sectional study in family practices, 300 people over 40 years of age underwent spirometry: of the 120 smokers, 30 had COPD; of the 180 nonsmokers, 15 had COPD. Is smoking associated with COPD?

Hypotheses. H0: the proportion with COPD is the same in smokers and nonsmokers (p1 = p2). H1: p1 ≠ p2.

Choice of test. Two dichotomous nominal variables, independent groups: chi-square on a 2 x 2 table (we check the assumptions).

Table 7.18. Smoking and COPD: observed (expected).

	COPD yes	COPD no	Total
Smokers	30 (18)	90 (102)	120
Nonsmokers	15 (27)	165 (153)	180
Total	45	255	300

Expected frequencies. $E_{11} = \frac{120\times45}{300} = 18$; $E_{12} = \frac{120\times255}{300} = 102$; $E_{21} = \frac{180\times45}{300} = 27$; $E_{22} = \frac{180\times255}{300} = 153$. All ≥ 5, so the test is valid.

The statistic. |O - E| = 12 in each cell.

$$\chi^2 = \frac{144}{18} + \frac{144}{102} + \frac{144}{27} + \frac{144}{153} = 8.000 + 1.412 + 5.333 + 0.941 = 15.69$$

df = 1; critical value 3.84; p < 0.001. With the Yates correction: (12 - 0.5)² = 132.25, and χ^2Yates = $132.25 \times \left(\frac{1}{18} + \frac{1}{102} + \frac{1}{27} + \frac{1}{153}\right) = 14.41$, p < 0.001. Fisher's test also gives p < 0.001. The coefficient φ = √(15.69/300) = 0.23.

Effect measures. Prevalence of COPD: 25.0% in smokers, 8.3% in nonsmokers. Prevalence ratio = $\frac{0.250}{0.083} = 3.0(;\ \text{CI}\,95\%\ 1.69 - 5.33)$; OR = $\frac{30\times165}{90\times15} = 3.67(;\ \text{CI}\,95\%\ 1.87 - 7.17)$.

Conclusions. There is a significant association between smoking and COPD (χ^2 = 15.69, p < 0.001); the prevalence of COPD is three times higher in smokers. The association is weak to moderate in strength as measured by φ, but clinically important, because COPD is common. The cross-sectional design does not by itself establish causality, but the result agrees with the evidence from cohort studies.

Worked Problem 7.6. Smoking Status and Diabetes in mg2026, a 3 x 2 Table

In the mg2026 dataset, OGTT-confirmed diabetes occurred in 3 of 33 former smokers, 10 of 53 smokers and 16 of 74 nonsmokers. Is smoking status associated with diabetes at enrollment?

Hypotheses. H0: the proportion with diabetes is the same in the three smoking categories. H1: at least one proportion differs.

Calculation of the expected frequencies (29 patients with diabetes out of 160, so 18.1%): former smokers $\frac{33\times29}{160} = 5.98$; smokers $\frac{53\times29}{160} = 9.61$; nonsmokers $\frac{74\times29}{160} = 13.41$; in the "no diabetes" column: 27.02, 43.39 and 60.59. All ≥ 5 (the smallest 5.98), so the test is valid.

The contributions: former smokers (3 - 5.98)²/5.98 = 1.486 and (30 - 27.02)²/27.02 = 0.329; smokers (10 - 9.61)²/9.61 = 0.016 and (43 - 43.39)²/43.39 = 0.004; nonsmokers (16 - 13.41)²/13.41 = 0.499 and (58 - 60.59)²/60.59 = 0.111. χ^2 = 2.44, df = (3 - 1)(2 - 1) = 2, p = 0.29. V Cramér $= \sqrt{\frac{2.44}{160}} = 0.12$.

Conclusions. We did not find a significant association between smoking status and diabetes (9.1%, 18.9% and 21.6%; $\chi^2 = 2.44$, p = 0.29). The lower percentage among former smokers is based on only 3 cases; with 33 people in that category, the estimate is very imprecise. Smoking remains a proven cardiovascular risk factor; this small sample lacks the power to reveal an association with diabetes.

Worked Problem 7.7. Two Rapid Tests on the Same Patients, McNemar

In 100 patients with pharyngitis, two rapid streptococcal tests (A and B) were applied. Both positive: 40; A positive and B negative: 15; A negative and B positive: 5; both negative: 40. Does the proportion of positive results differ between the two tests?

Choice of test. Each patient has two binary results, hence paired data: McNemar. The ordinary chi-square test would be wrong, because the same patients would appear in both rows.

Calculation. The proportions of positives: $\frac{A\,55}{100} = 55\%$, $\frac{B\,45}{100} = 45\%$. The discordant cells: b = 15, c = 5.

$\chi^2 = \frac{(15-5)^2}{15+5} = \frac{100}{20} = 5.00$, df 1, p = 0.025; with continuity correction: (10 - 1)²/20 = 4.05, p = 0.044; exact (binomial) version p = 0.041

Conclusions. Test A gives significantly more positive results than test B. McNemar's test does not tell us which test is better; for that, both would have to be compared with the reference standard (throat swab culture), that is, sensitivity and specificity would have to be calculated (Chapters 11 and 12).

Worked Problem 7.8. Postoperative Delirium in a Small Study, Fisher's Exact Test

In a pilot study of older surgical patients, postoperative delirium occurred in 1 of 12 patients who received a new sedative and in 6 of 10 patients who received usual care. Does the frequency of delirium differ?

Checking the assumptions. The total with delirium is 7 of 22. The expected frequencies: $\frac{12\times7}{22} = 3.82$ and $\frac{10\times7}{22} = 3.18$ in the "delirium" column; 8.18 and 6.82 in the "no delirium" column. Two of the four cells have E < 5, so we use Fisher's exact test.

Calculation. With the totals fixed (12 and 10 in the rows, 7 and 15 in the columns), the number a of delirium cases in the sedative group can range from 0 to 7. The hypergeometric probabilities are: P(0) = 0.0007; P(1) = 0.0148 (the observed table); P(2) = 0.0975; P(3) = 0.2709; P(4) = 0.3483; P(5) = 0.2090; P(6) = 0.0542; P(7) = 0.0046. We add the tables that are at most as probable as the observed one: 0.0007 + 0.0148 + 0.0046 = 0.020.

Conclusions. Fisher p = 0.020: delirium was significantly less frequent in the new sedative group (8.3% versus 60.0%). The chi-square test applied wrongly would have given p = 0.010, and with the Yates correction p = 0.033; the conclusion does not change here, but the correct value is the exact one. With 22 patients, the effect must be confirmed in an adequately sized randomized trial.

7.12 Proposed Exercises

1. Assign ranks to the values 12, 15, 15, 9, 20, 15, 11, 9 (in the order given) and check the sum of the ranks.
2. The anxiety score (0-10) before an endoscopy was: group with video information (n = 4) 3, 4, 4, 6; group without information (n = 5) 5, 7, 8, 8, 9. Calculate R1, R2, U1, U2 and U. The critical value for n1 = 4, n2 = 5, two-sided, α = 0.05, is 1. What do you decide?

3. Systolic blood pressure (mmHg) in 8 patients, before and after 4 weeks of treatment: before 140, 152, 138, 160, 145, 150, 148, 155; after 132, 150, 140, 148, 139, 150, 141, 146. Apply the Wilcoxon test by hand (the critical value for n = 7 is 2). What do you notice about patient 6?
4. In a study with three groups of 6 patients each (N = 18), the rank sums are 40, 60 and 71. Check the sum of the ranks, calculate H and decide ($\alpha = 0.05$).
5. Ten patients were assessed at three time points; the within-patient rank sums are 14, 20 and 26. Calculate the Friedman statistic and decide ($\alpha = 0.05$). Calculate Kendall's W.
6. During an influenza season, 10 of 200 vaccinated people and 24 of 150 unvaccinated people had confirmed influenza. Build the table and calculate the expected frequencies, χ^2 (with and without Yates), the relative risk and φ.
7. A 3 x 3 table has the observed frequencies: row 1: 8, 12, 5; row 2: 10, 20, 6; row 3: 2, 4, 3. Calculate the expected frequencies for row 3. Is the chi-square test applicable? What solutions do you propose?
8. In 80 patients, the presence of hypertension was assessed before and after a weight loss program. 12 patients moved from "hypertensive" to "normotensive", and 4 from "normotensive" to "hypertensive". Which test do you use? Calculate the statistic, with and without correction.
9. In a sample of 200 donors, blood groups O, A, B and AB were observed 70, 85, 30 and 15 times. Test the goodness of fit with the hypothetical proportions 35%, 40%, 17%, 8%.
10. An adverse effect occurred in 3 of 12 patients on drug X and in 0 of 12 patients on placebo. Calculate the expected frequencies, choose the test and interpret the p value given by the software (two-sided Fisher p = 0.217).
11. Using the mg2026 dataset in JASP, compare satisfaction at 6 months between arms B and A. Report the medians, U, p and the rank-biserial correlation.

Common Pitfalls

- Applying the Mann-Whitney test or the ordinary chi-square test to paired data (the same patients measured twice). For paired data: Wilcoxon and McNemar, respectively.
- Reporting the mean and standard deviation for Likert scores and then applying a rank-based test: the description must match the test (median and IQR).
- Checking the assumption of the chi-square test on the observed frequencies instead of the expected ones.
- Entering percentages instead of absolute frequencies in the contingency table.
- Concluding "the medians are equal" after a nonsignificant test; the correct wording: "we did not find a significant difference".
- Several Mann-Whitney tests between three groups, each at the 0.05 threshold, without a global test and without correction.
- Comparing U or T with the critical value in the wrong direction: for Mann-Whitney and Wilcoxon, H0 is rejected when the statistic is **less than or equal to** the critical value.
- Interpreting a large χ^2 as a "strong association": the strength is read from φ, Cramér's V or the RR/OR, not from p.

Key Points

- Nonparametric tests work on ranks; they are suitable for ordinal data, skewed distributions in small samples and genuine extreme values. Their efficiency with normal data is about 95%.

- Mann-Whitney: 2 independent groups; Wilcoxon: 2 paired measurements; Kruskal-Wallis: ≥ 3 independent groups (Dunn post hoc); Friedman: ≥ 3 repeated measurements (Wilcoxon post hoc with correction).
- Tied values receive the mean rank; the sum of the ranks is always N(N + 1)/2.
- Expected frequency: E = row total x column total/N; df = (r - 1)(c - 1).
- Chi-square is valid when E ≥ 5 in all cells of a 2 x 2 table; in larger tables, at most 20% of the cells may have E < 5 and none may have E < 1. Otherwise: Fisher or merging categories.
- McNemar for paired proportions, the test for trend for ordered categories, the goodness-of-fit test for a single variable compared with a theoretical distribution.
- Complete reporting: description (median and IQR, or n and %, respectively), the test statistic, df, the exact p and an effect size (r_rb, ε^2, W, φ, V, RR or OR with 95% CI).

Self-Assessment Questions

1. (SC) A study compares a quality-of-life score (ordinal scale 0-100, strongly skewed distribution) between 15 patients who underwent laparoscopic surgery and 15 patients who underwent open surgery. The appropriate test is:
    a) the paired t-test
    b) the Wilcoxon test for paired data
    c) the Mann-Whitney U test
    d) the Friedman test
    e) McNemar's test
2. (SC) The values 4, 7, 7, 2, 7, 9 receive, in the order given, the ranks:
    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. (SC) A contingency table has 4 rows and 3 columns. The number of degrees of freedom for the chi-square test is:
    a) 12
    b) 11
    c) 7
    d) 6
    e) 5
4. (SC) In a 2 x 2 table, the total of the first row is 40, the total of the first column is 30, and the grand total is 200. The expected frequency in the cell at their intersection is:
    a) 6
    b) 7,5
    c) 12
    d) 15
    e) 30
5. (SC) In 60 patients, the presence of chest pain was recorded before and after treatment. The appropriate test for comparing the proportions is:
    a) Pearson's chi-square on a "time point x pain" table with 120 observations

b) Fisher's exact test
c) McNemar's test
d) the Kruskal-Wallis test
e) the Cochran-Armitage test for trend

6. (MC) Regarding the Kruskal-Wallis test, the following statements are true:
   a) it is the nonparametric alternative to one-way ANOVA
   b) the ranks are assigned within each subject
   c) the H statistic is compared with the chi-square distribution with k - 1 degrees of freedom
   d) a significant result shows exactly which groups differ from each other
   e) after a significant result, Dunn's comparisons with correction for multiple comparisons can be made
7. (MC) The chi-square test on a 2 x 2 table is NOT appropriate (without other measures) if:
   a) one expected frequency is 3.2
   b) one observed frequency is 0, and all expected frequencies are above 5
   c) the same patients are counted both in the "before" row and in the "after" row
   d) the table contains percentages instead of numbers of patients
   e) the groups are of unequal size (300 and 80 patients)
8. (MC) In the mg2026 dataset, the Mann-Whitney test for satisfaction, arm C versus A, gave U = 2265.5, $p < 0.001$, rank-biserial correlation 0.58. The following are correct:
   a) satisfaction was higher in arm C
   b) the effect is large
   c) the medians are 4 (IQR 3-5) and 3 (IQR 2-3)
   d) the result proves that arm C is more effective in controlling blood glucose
   e) the probability that a patient from C is more satisfied than one from A is about 0.79
9. (MC) Regarding the Friedman test, the following are true:
   a) it is applied to ≥ 3 measurements on the same subjects
   b) the ranks are assigned in the combined sample of all measurements
   c) Kendall's W coefficient can be used as an effect size
   d) post hoc comparisons can be made with the Wilcoxon test and Bonferroni correction
   e) it is the nonparametric alternative to repeated-measures ANOVA
10. (MC) For the obesity x diabetes table in mg2026 ($\chi^2 = 12.70$, $p < 0.001$, $\varphi = 0.28$), the following are correct:

a) the association is statistically significant
b) φ shows a weak to moderate association
c) the small p value shows that the association is very strong
d) the cross-sectional design does not allow the conclusion that obesity caused diabetes
e) the expected frequency for the cell "obese with diabetes" is 9.79

Chapter 8. Selecting and Applying Statistical Tests: Selection Criteria, Problem Solving and Case Studies

Objectives

By the end of this chapter, the student will be able to:

- identify the aim of an analysis (difference, association, agreement, prediction) starting from the research question;
- classify the outcome variable and the grouping variable and recognize the design (independent groups or paired data, the number of groups or time points);
- apply step by step the algorithm for choosing a statistical test and justify the choice, including the nonparametric or categorical alternative;
- recognize the correct unit of analysis and the situations in which observations are not independent;
- explain why the test must be specified before analyzing the data and what happens when it is chosen according to the p value;
- recognize the most common mistakes in test choice in "Methods" descriptions and propose the correct test;
- write up the solution to a problem in the standardized format of the discipline (a-g), from the research hypothesis to the clinical conclusion;
- critically evaluate the fit between design, type of variable and the reported test in articles published by colleagues in the discipline.

8.1 Choosing the Test Is a Design Decision, Not a Calculation Decision

Modern statistical software calculates any test in a few seconds. JASP, MedCalc or Epi Info will not object if you ask them for a t-test on data that are actually paired or for a chi-square test on a table with expected frequencies of 1.5. The result will look impeccable, with three decimal places and a colorful graph, and it will be wrong. For this reason, the hard part of statistical analysis is not the calculation but the choice: what question we are asking, what type of data we have, which test corresponds to this combination.

In Chapters 5, 6 and 7 we looked at the tests one by one. This chapter brings them together. We will build a decision algorithm, apply it to 22 worked problems on the mg2026 dataset, practice recognizing wrong tests and read the methods sections of articles by colleagues in the discipline. In the practical exam, the typical task is exactly this: "Which statistical test would you apply to find out whether...?", followed by the complete solution in the a-g format. Experience in the lab shows that students who make mistakes in the exam do not usually go wrong in reading the p value, but at the first step: they confused paired data with independent groups or an ordinal variable with a numerical one.

Ideally, the test is chosen before data collection, in the study protocol, together with the definition of the primary variable and the sample size calculation [18,56]. The test follows from the design, and the design follows from the question. If the question is well formulated ("does the decrease in blood glucose at 6 months differ between the three arms?"), the test practically chooses itself.

8.2 The Algorithm for Choosing the Test

Figure 8.1 summarizes the algorithm used in this book. It is not the only possible one, but it covers the vast majority of situations in basic clinical research.

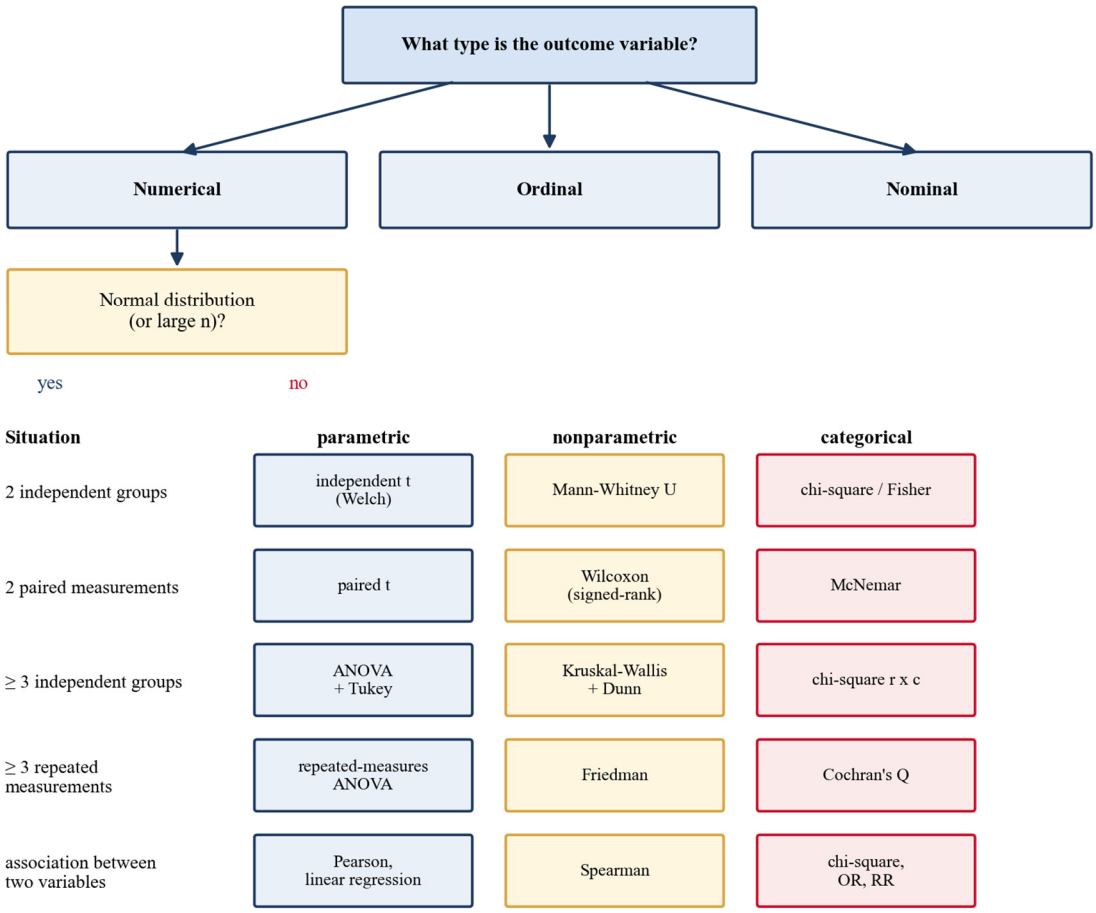


Figure 8.1. The algorithm for choosing a statistical test: the type of outcome variable (numerical, ordinal, nominal), then the situation (2 independent groups, 2 paired measurements, ≥ 3 independent groups, ≥ 3 repeated measurements, association of two variables) and, for numerical variables, the distribution (normal or large n versus skewness with small n). The columns give the parametric, nonparametric and categorical versions.

The algorithm is worked through in five steps. The order matters: a wrong step at the beginning cannot be repaired later.

8.2.1 Step 1: What Is the Aim of the Analysis?

The first question is not about the data but about the intention. There are four broad aims, and each requires a different family of methods:

- **difference** between groups or between time points: "Is blood glucose lower in arm C than in arm A?" (t-tests, ANOVA, Mann-Whitney, Kruskal-Wallis, Wilcoxon, Friedman, chi-square, McNemar);
- **association** between two variables measured on the same subjects: "Does systolic blood pressure increase with BMI?" (Pearson or Spearman correlation, chi-square for qualitative variables, OR and RR);
- **agreement** between two methods or two raters measuring the same thing: "Do the automatic and the manual blood pressure monitor give the same values?" (the Bland-Altman method [102], the intraclass correlation coefficient, Cohen's kappa [103]);

- **prediction** of an outcome from one or more factors: "How well does FINDRISC predict diabetes?" (linear or logistic regression, the ROC curve).

The most common confusion is between association and agreement. Two blood pressure monitors can have a correlation of 0.95 and yet one of them may systematically read 10 mmHg higher. Correlation measures whether the values increase together; agreement measures whether they are the same. For agreement, neither the paired t-test alone nor correlation is used, but the analysis of differences (the mean of the differences and the limits of agreement).

8.2.2 Step 2: What Type Is the Outcome Variable?

The outcome (dependent) variable is the one we want to compare or explain: blood glucose, satisfaction, presence of diabetes. The grouping (independent) variable defines the groups: arm, sex, area of residence. The type of the outcome variable determines the column in Figure 8.1.

- **Numerical** (continuous, or discrete with many values): blood glucose, SBP, weight, BMI, age, FINDRISC score 0-26.
- **Ordinal**: satisfaction 1-5, educational level 1-4, self-rated physical activity 1-5, stage of a disease.
- **Nominal** (including dichotomous): diabetes yes/no, loss ≥ 5% yes/no, smoking status, sex.

Two situations require attention. A score with many values (FINDRISC 0-26, a quality-of-life score 0-100) is usually treated as numerical, if the distribution allows it. A dichotomized numerical variable (blood glucose ≥ 100 mg/dL yes/no) becomes nominal, but loses information and power; dichotomization is justified only when the cutoff is clinically meaningful and was set in advance.

8.2.3 Step 3: How Many Groups or How Many Time Points?

We count the groups or time points being compared: one (comparison with a reference value), two, or three or more. The distinction between "two" and "three or more" is not a formality. With three groups, three separate t-tests raise the overall type I error risk from 5% to 14.3% ($1 - 0.95^3$); with five groups, the ten possible comparisons bring it to 40%. That is why, for three or more groups, we use a global test (ANOVA, Kruskal-Wallis, r x c chi-square) and only then post hoc comparisons with correction.

8.2.4 Step 4: Independent Groups or Paired Data?

Data are **paired** (dependent) when each value in one series has a natural counterpart in the other: the same patient before and after, the right eye and the left eye, two methods on the same sample, matched pairs in case-control studies. Data are **independent** when the groups are made up of different, unrelated subjects. The control question is simple: "can I join each value in the first column with a line to a specific value in the second?" If yes, the data are paired.

The difference in the result can be large. In mg2026, blood glucose decreased from 105.44 mg/dL at enrollment to 95.25 mg/dL at 6 months. The paired t-test (correct) gives t = 11.14, with a standard error of the difference of 0.91 mg/dL. If we wrongly treat the two series as two independent groups (Welch test), the standard error increases to 2.03 mg/dL, and t drops to 5.03. The reason is the strong correlation between a patient's blood glucose at 0 and at 6 months (r = 0.82): the paired test removes the variability between patients and is left only with the variability of the change. Here both versions come out significant, but in a smaller sample the wrong test could have missed a real effect.

8.2.5 Step 5: The Distribution and the Sample Size

For numerical variables only, the last question is whether we can use a parametric test. The criteria, discussed in Chapters 3 and 6:

- the shape of the distribution in each group (histogram, Q-Q plot, skewness) and, as an indication, the Shapiro-Wilk test [66]; for the paired t-test, the distribution of the **differences** is checked, not that of the two series;
- the sample size: with at least 30-40 subjects per group and moderate skewness, the t-test and ANOVA remain robust; with small groups, the distribution matters much more;
- homogeneity of variances (Levene's test); if it is violated, the Welch version of the t-test or of ANOVA is used, not automatically a nonparametric test;
- genuine extreme values and bounded variables (scores with a ceiling effect).

A "significant" Shapiro-Wilk test in a large sample does not force us to use a nonparametric test. In mg2026, blood glucose at enrollment has SW p = 0.037 with n = 160, but the skewness is only 0.37, and the histogram shows an almost symmetric distribution. For 160 patients, the t-test is perfectly adequate. Conversely, a "nonsignificant" SW in 8 patients does not guarantee normality; it only shows the low power of the normality test.

8.2.6 The Algorithm in Brief

For quick work in the lab, the algorithm can be written as a checklist:

1. Write the research question in one sentence. What do you want to find out: a difference, an association, agreement or a prediction?
2. Identify the outcome variable and its type (numerical, ordinal, nominal).
3. Identify the grouping variable or the time points: how many categories are there?
4. Check whether the data are independent or paired.
5. For a numerical variable: check the distribution in each group (or of the differences) and the homogeneity of variances; decide between the parametric, Welch or nonparametric version.
6. For qualitative variables: calculate the expected frequencies and decide between chi-square, Yates, Fisher or McNemar.
7. For three or more groups: a global test, then post hoc comparisons with correction.
8. Decide in advance which effect size you will report (difference in means with 95% CI, Cohen's d, r_rb, RR, OR, φ, V).

8.3 Summary Table of the Tests

Table 8.1 brings together in one place the situations, the tests and where to find them in the software used in the practical sessions (JASP 0.98.1 [47], MedCalc 24.0 [50], Epi Info 7.2.7 [51]). Menu names may differ slightly between versions; in Practical Sessions 6, 7 and 8 the steps are described in detail.

Table 8.1. Summary of common statistical tests and the software menus.

Situation	Numerical variable, normal distribution (parametric)	Ordinal or skewed numerical variable (nonparametric)	Nominal variable (categorical)	JASP	MedCalc	Epi Info
1 sample against a reference value	one-sample t-test	one-sample Wilcoxon	binomial test; chi-square goodness-of-fit test	T-Tests > One Sample T-Test; Frequencies > Binomial Test, Multinomial Test	One sample t-test; Signed rank sum test	StatCalc: Binomial

Situation	Numerical variable, normal distribution (parametric)	Ordinal or skewed numerical variable (nonparametric)	Nominal variable (categorical)	JASP	MedCalc	Epi Info
2 independent groups	unpaired t-test (Student; Welch for unequal variances)	Mann-Whitney U	2 x 2 chi-square; Fisher when 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 with a grouping variable; TABLES; StatCalc: Tables 2 x 2
2 paired measurements	paired t-test	Wilcoxon (signed-rank); sign test	McNemar	T-Tests > Paired Samples T-Test (Student, Wilcoxon signed-rank)	Paired samples t-test; Wilcoxon test (paired samples); McNemar test	StatCalc: matched pairs (McNemar)
≥ 3 independent groups	one-way ANOVA + Tukey (Welch for unequal variances)	Kruskal-Wallis + Dunn	r x c chi-square	ANOVA > ANOVA (Post Hoc Tests; Nonparametrics: Kruskal-Wallis, Dunn)	One-way analysis of variance; Kruskal-Wallis test	MEANS (ANOVA and Kruskal-Wallis); TABLES
≥ 3 repeated measurements	repeated-measures ANOVA	Friedman + Wilcoxon with correction	Cochran's Q	ANOVA > Repeated Measures ANOVA (Nonparametrics: Friedman)	Repeated measures analysis of variance; Friedman test; Cochran's Q test	no dedicated procedure
proportion across ordered categories	-	Mann-Whitney or Kruskal-Wallis on the ordinal variable	chi-square test for trend	Frequencies > Contingency Tables	Chi-squared test for trend	StatCalc: Chi square for trend
association of two numerical or ordinal variables	Pearson correlation	Spearman, Kendall correlation	-	Regression > Correlation	Correlation; Rank correlation	Classic Analysis: REGRESS
association of two nominal variables	-	-	chi-square, Fisher; φ, Cramér's V; RR, OR with 95% CI	Frequencies > Contingency Tables (Phi and Cramér's V; Log odds ratio)	Chi-squared test; Relative risk; Odds ratio	TABLES; StatCalc: Tables 2 x 2
agreement between two methods or raters	Bland-Altman; intraclass correlation coefficient	weighted kappa	Cohen's kappa	the Reliability module	Bland-Altman plot; Intraclass correlation; Inter-rater agreement (kappa)	-
prediction of a	simple or	-	-	Regression >	Multiple	Classic

Situation	Numerical variable, normal distribution (parametric)	Ordinal or skewed numerical variable (nonparametric)	Nominal variable (categorical)	JASP	MedCalc	Epi Info
numerical outcome	multiple linear regression			Linear Regression	regression	Analysis: REGRESS
prediction of a binary outcome	logistic regression; ROC curve	-	-	Regression > Logistic Regression	Logistic regression; ROC curve analysis	Classic Analysis: LOGISTIC
time to event	Kaplan-Meier, log-rank, Cox regression	-	-	the Survival module	Kaplan-Meier survival analysis; Cox proportional-hazards regression	Classic Analysis: Kaplan-Meier, Cox

A student who has mastered the first five rows of the table can solve almost all the problems in the practical exam. The rows for agreement, prediction and survival are covered in detail in Chapters 9-13.

8.4 The Unit of Analysis and Independence of Observations

All the tests in Table 8.1 assume that the observations are independent (or, for paired tests, that the pairs are independent of each other). The assumption seems trivial, but it is often violated, especially when one patient contributes several values.

The **unit of analysis** is the entity on which randomization or sampling was carried out and about which we want to draw a conclusion: usually the patient. Problems arise when we analyze something else:

- **eyes, teeth, joints, lesions**: if a patient contributes two eyes, the values of the two eyes are correlated. In our colleagues' study of ocular biometry, 170 eyes came from 102 patients [81]. In dentistry, the surfaces of the same tooth and the teeth of the same patient are clustered (cluster effect) [104]; treating 600 teeth from 30 patients as 600 independent observations artificially inflates n and produces p values that are too small.
- **repeated visits**: a patient with 5 hospital admissions in one year does not count as 5 patients. If we analyze admissions, methods for repeated data must be used.
- **clusters**: in mg2026, the 160 patients come from 12 family practices. Patients of the same physician may resemble each other (the same counseling style, the same area). If randomization had been done by practice (all patients of a practice in the same arm), the unit of analysis would have become the practice, and the analysis would have had to take this grouping into account. In our scenario, randomization was individual, so the usual tests are acceptable, but a sensitivity analysis could include the practice as a factor.
- **events instead of people**: 30 episodes of hypoglycemia may come from 30 patients or from 3 patients. A chi-square test on episodes would violate independence.

The practical rule: before any test, ask yourself "how many independent units do I actually have?". The answer is the real n of the analysis.

8.5 The Test Is Specified Before Seeing the Data

A statistical test keeps its type I error at 5% only if it was chosen independently of the result. If we try several tests on the same data (t, Welch, Mann-Whitney, with and without extreme values, with and without transformation) and report the one that "comes out", the real probability of a false positive result rises well above 5%, even if each test, taken separately, is correct [70,74]. The phenomenon is known in the literature as the "garden of forking paths": each seemingly innocent decision opens a new chance of obtaining $p < 0.05$ by chance.

The mg2026 dataset offers a concrete example, shown in Figure 8.2. One of the secondary variables is the decrease in systolic blood pressure at 6 months. The means are 2.79 mmHg (SD 8.96) in arm A, 4.53 (SD 9.63) in arm B and 6.96 (SD 8.18) in arm C. The distributions are compatible with the normal in each arm (SW p between 0.07 and 0.43), and the variances are homogeneous (Levene $p = 0.67$). The appropriate test, specified in the analysis plan, is classic ANOVA.

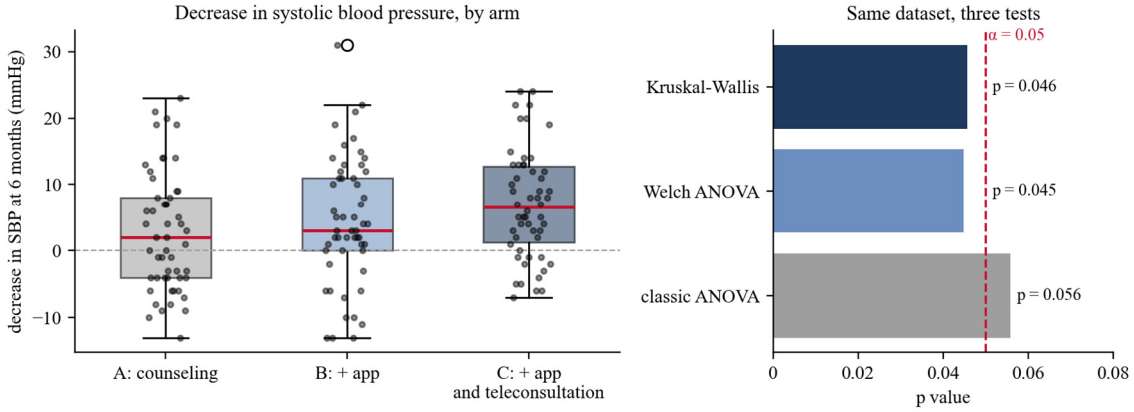


Figure 8.2. Left: decrease in systolic blood pressure at 6 months, by arm (boxes with the median and quartiles, individual points). Right: the p values of three tests applied to the same data: classic ANOVA 0.056, Welch's ANOVA 0.045, Kruskal-Wallis 0.046. The red line marks the 0.05 threshold. Synthetic data (mg2026 dataset).

Classic ANOVA gives $F(2, 157) = 2.94$, $p = 0.056$. The Welch version gives $p = 0.045$, and Kruskal-Wallis gives $p = 0.046$. A researcher who chooses the test according to the result will report "$p < 0.05$" and write that teleconsultation lowers blood pressure. A researcher who follows the plan will report $p = 0.056$, $\eta^2 = 0.036$, and will say that the data suggest a larger decrease in arm C (C-A difference of 4.17 mmHg, Tukey 95% CI 0.08-8.26), but that the global test did not reach the prespecified threshold. The second wording is the correct one. The difference between 0.045 and 0.056 is clinically negligible; the problem lies not in the figures but in the honesty of the procedure.

The solutions are simple and are part of good practice: the primary test, the primary variable and the threshold are written in the protocol before the study begins; additional analyses are explicitly labeled "sensitivity analyses" or "exploratory"; and if an assumption turns out to be violated, the fallback rule ("if Levene $p < 0.05$, we use Welch") is also written in advance. Reporting guidelines for clinical trials require that changes from the original plan be declared [56].

8.6 Common Mistakes in Choosing and Applying the Test

8.6.1 A Test for Independent Groups Applied to Paired Data

In Section 8.2.4 we saw the effect on blood glucose in mg2026: $t = 5.03$ instead of $t = 11.14$. The same mistake occurs with binary variables: an ordinary chi-square test on a "time point (0 or 6

months) x blood glucose ≥ 100 (yes/no)" table with 320 observations, instead of McNemar's test on 160 patients. The warning sign when reading an article: the same patients measured twice and the word "independent" or Pearson's chi-square test in the methods section.

8.6.2 Multiple t-Tests Instead of ANOVA

With three arms, the temptation is to compare A with B, A with C and B with C using three t-tests. Each comparison has a 5% risk of a false positive result, and the overall risk reaches 14.3%. The correct procedure: ANOVA (or Kruskal-Wallis), followed by post hoc comparisons with correction (Tukey, Bonferroni, Holm, Dunn). The legitimate exception is a primary comparison specified in advance (for example C versus A as the primary objective), which can be tested directly.

8.6.3 Chi-Square with Small Expected Frequencies

In Chapter 7, for a table with 18 patients and expected frequencies of 3.1 and 3.9, the chi-square test gave $p = 0.040$, and the correct test (Fisher) $p = 0.066$. Software sometimes displays a discreet warning ("x cells have an expected frequency below 5"), which many people ignore. Checking the expected frequencies is part of the solution and is written down in the a-g format, in the justification of the method.

8.6.4 Choosing the Test According to the p Value

The example in Section 8.5 is typical. Common variants of the same mistake: switching to a one-sided test after the two-sided test gave $p = 0.08$; removing "extreme values" only until the result becomes significant; swapping the primary variable for a secondary one that came out significant; testing 20 subgroups and reporting the one with $p < 0.05$.

8.6.5 The Mean of Ordinal Scores Treated as a Measurement

Satisfaction in mg2026 has means of 2.96 (A), 3.62 (B) and 3.98 (C). ANOVA on these scores gives $F = 17.32$, $p < 0.001$, and Kruskal-Wallis $H = 32.16$, $p < 0.001$: the conclusion does not change. The problem here is not the p value but the interpretation: "mean satisfaction increased by 1.02 points" assumes that the difference between "neutral" and "satisfied" is equal to that between "satisfied" and "very satisfied", which the scale does not guarantee. Correct reporting uses medians and the distribution across categories (Figure 7.2 in Chapter 7). In small samples or with distributions showing a ceiling effect, the mean of ordinal scores can also lead to wrong statistical conclusions.

8.6.6 Other Mistakes We Often Encounter

Testing normality on the whole sample instead of in each group (or on the differences, for paired data) is a common technical mistake: two normal groups with different means together form a bimodal distribution, and the normality test "fails" for no good reason. Another mistake, specific to randomized trials, is statistical testing of baseline differences between arms (Table 1 with p values). If randomization was done correctly, any baseline difference is, by definition, due to chance; the CONSORT guideline recommends describing them, not testing them [56]. In mg2026, the level of physical activity at enrollment has $p = 0.08$ between arms (Kruskal-Wallis), and smoking status $p = 0.07$ (chi-square): both are fluctuations of randomization, which are discussed in terms of their clinical size, not their p value. The last common mistake is ignoring multiple testing: an article that reports 7 odds ratios for 7 different disorders, each at the 0.05 threshold, has a probability of about 30% of obtaining at least one false positive result even if none of the associations were real.

8.7 Reporting a Result in the Standardized Format (a-g)

In the assignments and the practical exam of the discipline, each problem is solved in the same format, with seven components. The format is not a bureaucratic formality: it enforces the correct order of thinking (first the question, then the hypotheses, then the method and only at the end the calculation) and makes the solution verifiable by someone else.

- **(a) Research hypothesis**: what the researcher believes, in clinical language. "Patients in arm C have a larger decrease in blood glucose than those in arm A."
- **(b) Statistical hypotheses**, in words and mathematically: H0 and H1, with the population parameter (μ, Me, p) and stating whether the test is two-sided.
- **(c) The chosen method and its justification**: the type of variables, the design (independent or paired, the number of groups), the check of the assumptions (normality by group, Levene, expected frequencies) and the alternative that would have been used if an assumption had not been met.
- **(d) The solution procedure and the software**: the software and the menu (for example JASP, T-Tests > Independent Samples T-Test, with the options ticked), or the hand calculation step by step.
- **(e) The result, with a graph**: descriptive statistics by group, the test statistic, the degrees of freedom, the p value, the effect size with 95% CI and the appropriate graph (box plot or plot of means with CI for numerical variables, bar chart for proportions).
- **(f) Statistical conclusion**: "$p < 0.05$, so we reject H0" or "$p \geq 0.05$, so we do not reject H0", with the wording "there is / there is no statistically significant difference between...".
- **(g) Research conclusion**, in words: what the result means for the clinical question, with the effect size and the limitations of the design.

Table 8.2 gives a complete example, using the main question of the mg2026 study.

Table 8.2. Model solution in the a-g format: decrease in blood glucose at 6 months by arm (mg2026 dataset).

Component	Content
(a) Research hypothesis	The decrease in fasting plasma glucose at 6 months differs according to the intervention received (counseling, counseling + app, counseling + app + teleconsultation).
(b) Statistical hypotheses	H0: the mean decrease in blood glucose is the same in the three arms ($\mu A = \mu B = \mu C$). H1: at least one mean differs from the others.
(c) Method and justification	The outcome variable (decrease in blood glucose, mg/dL) is numerical; the grouping variable (arm) is nominal with 3 categories; the groups are independent. The distributions are approximately normal in each arm and the variances are homogeneous (Levene $p = 0.163$), so one-way ANOVA, followed by the Tukey test. Alternative if the assumptions are violated: Kruskal-Wallis with Dunn.
(d) Solution and software	JASP 0.98.1, ANOVA > ANOVA; Dependent Variable: decrease in blood glucose; Fixed Factors: brat; Assumption Checks: Homogeneity tests; Post Hoc Tests: Tukey; Estimates of effect size: η^2.
(e) Result	A: $3.0 \pm 8.8 (n\,53)$; B: $11.1 \pm 11.2 (n\,53)$; C: $16.3 \pm 10.6 (n\,54)$; $F(2, 157) = 22.93$, $p < 0.001$, $\eta^2 = 0.226$. Tukey: B-A 8.1 (95% CI 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; $p = 0.024$). Graph: box plot of the decrease in blood glucose by arm, as in Chapter 6.
(f) Statistical conclusion	$p < 0.001 < 0.05$, so we reject H0: there are statistically significant differences between arms; all three pairwise

Component	Content
	comparisons are significant.
(g) Research conclusion	Adding the mobile app and, in addition, monthly teleconsultation led to a larger decrease in fasting plasma glucose at 6 months (on average 8 and 13 mg/dL more, respectively, than counseling alone). The effect is large ($\eta^2 = 0.226$). The data are simulated, and in a real study the clinical relevance of a decrease of this size would also need to be checked.

8.8 Worked Problems

The problems below use the mg2026 dataset (the file mg2026_curat.csv on Moodle) and follow the style of the exam questions and of the discipline's handouts "Teste statistice - recapitulare" (Statistical tests: review) and "Probleme BPOC" (COPD problems). For each problem, the question, the type of variables, the design, the chosen test with its alternative, the hypotheses, the result and the conclusions are given. All figures were calculated on the dataset; you can reproduce them in JASP or MedCalc. The significance level is $\alpha = 0.05$, and the tests are two-sided.

Worked Problem 8.1. Fasting Plasma Glucose in Urban and Rural Patients

The question. Does fasting plasma glucose at enrollment differ between patients from urban and rural areas?

Variables and design. Outcome: blood glucose at enrollment (numerical, mg/dL). Grouping: area of residence (nominal, 2 categories). Two independent groups.

Chosen test and alternative. Shapiro-Wilk by group: urban p = 0.076, rural p = 0.054; large groups (95 and 65). Levene p = 0.117. Unpaired t-test (we report the Welch version, which does not assume equal variances); alternative: Mann-Whitney U. In JASP: T-Tests > Independent Samples T-Test, with Student, Welch, Mann-Whitney and Assumption checks ticked.

Hypotheses. H0: μurban = μrural; H1: μurban ≠ μrural.

Result. Urban 105.82 ± 15.30 mg/dL(n 95), rural 104.88 ± 16.52 mg/dL(n 65). Welch t = -0.37, df 130.5, p = 0.715; difference rural - urban -0.94 mg/dL (95% CI -6.05 to 4.16); Cohen's d -0.06. Mann-Whitney p = 0.81.

Statistical conclusion. $p = 0.715 > 0.05$, we do not reject H0: there is no statistically significant difference.

Clinical conclusion. Fasting plasma glucose was practically the same in the two settings. The confidence interval shows that a real difference larger than about 6 mg/dL in either direction is unlikely, which is more informative than a simple "nonsignificant".

Worked Problem 8.2. HDL Cholesterol in Women and Men

The question. Does HDL cholesterol differ between women and men?

Variables and design. Outcome: HDL (numerical, mg/dL). Grouping: sex (2 categories). Independent groups.

Chosen test and alternative. SW: women p = 0.41, men p = 0.48, normal distributions. Levene p = 0.034, unequal variances, hence the Welch t-test. Alternative: Mann-Whitney.

Hypotheses. H0: μF = μM; H1: μF ≠ μM.

Result. Women 53.92 ± 11.70 mg/dL(n 86), men 44.41 ± 9.66 mg/dL(n 74). Welch t = 5.63, df 157.7, $p < 0.001$; difference F - M 9.51 mg/dL (95% CI 6.18-12.85); d = 0.88. Mann-Whitney $p < 0.001$.

Statistical conclusion. $p < 0.001$, we reject H0: the difference is statistically significant.

Clinical conclusion. Women have an HDL on average about 9.5 mg/dL higher than men, a large effect, in agreement with the known physiological differences. Practical consequence: the cutoffs for "low" HDL differ by sex.

Worked Problem 8.3. BMI in Women and Men, a Skewed Variable

The question. Does BMI at enrollment differ between women and men?

Variables and design. Outcome: BMI (numerical, kg/m²), with positive skewness (1.42 in the whole sample). Grouping: sex. Independent groups.

Chosen test and alternative. SW: women $p = 0.0001$, men $p < 0.001$; the distributions are skewed. We choose the Mann-Whitney U test, and we report the Welch t-test as a sensitivity analysis (the groups are large, so it is also robust).

Hypotheses. H0: the BMI distributions are identical in women and men (MeF = MeM); H1: the distributions differ.

Result. Women: median 28.25 kg/m² (IQR 26.02-31.25), men: 28.25 (IQR 26.32-31.50). Mann-Whitney $U = 3020.5$, $p = 0.58$; $r_{rb} = 0.05$. Welch $p = 0.68$.

Statistical conclusion. $p = 0.58$, we do not reject H0.

Clinical conclusion. BMI does not differ between the sexes in this sample; both tests lead to the same conclusion. When the groups are large and the tests agree, the choice between them does not change the interpretation; what matters is that the description (median and IQR) is consistent with the chosen test.

Worked Problem 8.4. The FINDRISC Score in Patients With and Without Diabetes

The question. Do patients with OGTT-confirmed diabetes have a higher FINDRISC score than those without diabetes?

Variables and design. Outcome: FINDRISC (score 0-26, treated as numerical). Grouping: diabetes (yes/no). Independent groups, very unequal in size (29 and 131).

Chosen test and alternative. SW by group: $p = 0.17$ and 0.33; Levene $p = 0.67$. Unpaired t-test; alternative: Mann-Whitney.

Hypotheses. H0: $\mu_{\text{diabetes}} = \mu_{\text{no diabetes}}$; H1: $\mu_{\text{diabetes}} \neq \mu_{\text{no diabetes}}$.

Result. Without diabetes 9.40 ± 3.70 (median 10), with diabetes 14.76 ± 3.96 (median 14). Student $t = -6.97$, $p < 0.001$ (Welch $t = -6.67$, df 39.5, $p < 0.001$); difference 5.36 points (95% CI 3.74-6.99); $d = 1.43$. Mann-Whitney $p < 0.001$, $r_{rb} = 0.71$.

Statistical conclusion. We reject H0.

Clinical conclusion. The FINDRISC score is on average more than 5 points higher in patients with diabetes, a very large effect. If the question had been "how well does FINDRISC identify patients with diabetes?", the aim would have been prediction, and the appropriate method would have been the ROC curve (AUC 0.852 in mg2026), not the t-test. The same dataset, two questions, two different methods.

Worked Problem 8.5. Systolic Blood Pressure in Patients With and Without Diabetes, When the Tests Disagree

The question. Does SBP at enrollment differ between patients with and without diabetes?

Variables and design. Outcome: SBP (numerical, mmHg). Grouping: diabetes (yes/no). Independent groups (29 and 131).

Chosen test and alternative. SW: p = 0.19 and 0.20; Levene p = 0.53. The test specified in advance is the unpaired t-test; the alternative, Mann-Whitney.

Hypotheses. H0: μdiabetes = μno diabetes; H1: μdiabetes ≠ μno diabetes.

Result. Without diabetes 129.96 ± 13.86 mmHg (median 131), with diabetes 136.41 ± 12.92 mmHg (median 132). Student t = -2.30, p = 0.023 (Welch t = -2.40, p = 0.021); difference 6.45 mmHg (95% CI 1.04-11.87); d = 0.47. Mann-Whitney p = 0.065.

Statistical conclusion. According to the test chosen in advance (unpaired t-test), p = 0.023 < 0.05, we reject H0.

Clinical conclusion. Patients with diabetes had an SBP on average about 6.5 mmHg higher, a small to medium effect, imprecisely estimated (CI from 1 to 12 mmHg). Why does Mann-Whitney give p = 0.065? The medians are very close (131 and 132), and the difference in means comes mainly from the upper tail of the distribution in the patients with diabetes. The tests answer slightly different questions. We are not allowed to choose the test according to the result; if the protocol specified the t-test and its assumptions are met, we report the t-test and mention the discrepancy as a sensitivity analysis.

Worked Problem 8.6. Weight at Enrollment and at 6 Months, in the Whole Sample

The question. Did the patients' weight change between enrollment and 6 months?

Variables and design. Outcome: weight (numerical, kg), measured twice in the same 160 patients. Paired data.

Chosen test and alternative. For paired data, the distribution of the differences is checked: SW p = 0.86, normal. Paired t-test; alternative: Wilcoxon. In JASP: T-Tests > Paired Samples T-Test, the pair greutate0 - greutate6.

Hypotheses. H0: mean of the differences μd = 0; H1: μd ≠ 0.

Result. Mean weight 83.05 kg at enrollment and 80.74 kg at 6 months. Mean difference 2.31 kg (SD 2.66), 95% CI 1.90-2.73; t = 11.00, df 159, p < 0.001; dz = 0.87. Wilcoxon p < 0.001.

Statistical conclusion. We reject H0.

Clinical conclusion. The patients lost on average 2.3 kg in 6 months. The paired test on the whole sample shows that a change occurred, but not what caused it: without comparing the arms, we cannot separate the effect of the interventions from the effect of taking part in the study. The question of effectiveness is answered by comparing weight loss between arms (ANOVA, F = 18.25, p < 0.001), that is, problem L5 in Practical Session 8.

Worked Problem 8.7. Systolic Blood Pressure at Enrollment and at 6 Months

The question. Did SBP decrease between enrollment and 6 months?

Variables and design. SBP, numerical, the same patients at two time points: paired data.

Chosen test and alternative. The differences have SW p = 0.13. Paired t-test; alternative: Wilcoxon.

Hypotheses. H0: μd = 0; H1: μd ≠ 0.

Result. Mean SBP 131.13 mmHg at enrollment and 126.36 mmHg at 6 months; mean difference 4.78 mmHg (SD 9.05), 95% CI 3.36-6.19; t = 6.68, df 159, p < 0.001; dz = 0.53. Wilcoxon p < 0.001.

Statistical conclusion. We reject H0.

Clinical conclusion. SBP decreased on average by about 4.8 mmHg. At the population level, a decrease of this size is considered relevant for cardiovascular risk, but here too the lack of an

untreated comparison group limits the causal interpretation (regression to the mean and the effect of repeated measurement may contribute).

Worked Problem 8.8. Decrease in Blood Glucose Across the Three Arms

The question. Does the decrease in blood glucose at 6 months differ between arms A, B and C?

Variables and design. Numerical outcome, nominal grouping with 3 categories, independent groups.

Chosen test and alternative. One-way ANOVA with Tukey (Levene p = 0.163); alternative: Kruskal-Wallis with Dunn.

Hypotheses. H0: μA = μB = μC; H1: at least one mean differs.

Result. The complete solution is in Table 8.2: F(2, 157) = 22.93, p < 0.001, $\eta^2 = 0.226$; Tukey significant for all three pairs. Kruskal-Wallis H = 36.63, p < 0.001.

Statistical conclusion. We reject H0.

Clinical conclusion. The interventions with the app produced larger decreases in blood glucose, and the effect increases with the intensity of the intervention.

Worked Problem 8.9. Systolic Blood Pressure by Smoking Status

The question. Does SBP at enrollment differ between nonsmokers, former smokers and smokers?

Variables and design. Outcome: SBP (numerical). Grouping: smoking (nominal, 3 categories). Independent groups.

Chosen test and alternative. SW by group: 0.37, 0.86, 0.23; Levene p = 0.114. ANOVA with Tukey; alternative: Kruskal-Wallis.

Hypotheses. H0: μnonsmokers = μformer smokers = μsmokers; H1: at least one mean differs.

Result. Former smokers $127.55 \pm 14.39 (n\,33)$, smokers $137.04 \pm 14.63 (n\,53)$, nonsmokers $128.50 \pm 11.76 (n\,74)$. F(2, 157) = 7.85, p < 0.001 (p = 0.0006), $\eta^2 = 0.091$. Tukey: smokers - former smokers 9.49 mmHg (95% CI 2.50-16.48; p = 0.005); smokers - nonsmokers 8.54 mmHg (95% CI 2.87-14.21; p = 0.001); nonsmokers - former smokers 0.95 mmHg (p = 0.94). Kruskal-Wallis H = 12.77, p = 0.002.

Statistical conclusion. We reject H0; smokers differ from both other groups.

Clinical conclusion. Current smokers have an SBP about 9 mmHg higher than nonsmokers and former smokers, and former smokers do not differ from nonsmokers. The result agrees with the multiple regression in Chapter 9, in which smoking remains associated with SBP (+6.74 mmHg compared with former smokers) after adjustment for BMI, age and sex. Being a cross-sectional analysis, it does not by itself establish causality.

Worked Problem 8.10. Decrease in Systolic Blood Pressure by Arm, a Borderline Result

The question. Does the decrease in SBP at 6 months differ between the three arms?

Variables and design. Numerical outcome, three independent groups.

Chosen test and alternative. SW by arm 0.07-0.43; Levene p = 0.67. The analysis plan specifies classic ANOVA; the alternative in case of unequal variances: Welch; in case of skewness: Kruskal-Wallis.

Hypotheses. H0: μA = μB = μC; H1: at least one mean differs.

Result. A 2.79 ± 8.96; B 4.53 ± 9.63; C 6.96 ± 8.18 mmHg. F(2, 157) = 2.94, p = 0.056, η^2 = 0.036 (Figure 8.2).

Statistical conclusion. p = 0.056 > 0.05, we do not reject H0.

Clinical conclusion. The data suggest a larger decrease in SBP in arm C, but the prespecified global test does not reach the significance threshold. The fact that Welch (p = 0.045) and Kruskal-Wallis (p = 0.046) fall below 0.05 does not change the conclusion, because these were not the planned tests, and the assumptions of classic ANOVA are met. The result is a signal worth investigating as the primary objective in a future study. Post hoc comparisons are not interpreted as evidence after a nonsignificant global test.

Worked Problem 8.11. Satisfaction by Educational Level

The question. Does satisfaction at 6 months differ by educational level?

Variables and design. Outcome: satisfaction (ordinal 1-5). Grouping: education (ordinal, 4 levels). Independent groups.

Chosen test and alternative. Ordinal outcome, 4 groups: Kruskal-Wallis. Because education is also ordered, a complementary question, "does satisfaction increase with education?", is tested more efficiently with Spearman correlation.

Hypotheses. H0: the distribution of satisfaction is the same across the 4 educational levels; H1: at least one level differs.

Result. The medians: lower secondary school 3 (n 20), high school or vocational school 4 (n 57), university 3 (n 64), postgraduate 4 (n 19). Kruskal-Wallis H = 4.12, df 3, p = 0.25. Spearman rho = 0.125, p = 0.115.

Statistical conclusion. We do not reject H0 (neither for a difference nor for a trend).

Clinical conclusion. Satisfaction does not differ significantly by educational level. The absence of a difference is a useful result for implementation: the program seems to be equally well accepted by patients with different levels of education, although the extreme groups (20 and 19 patients) are small.

Worked Problem 8.12. Physical Activity at Enrollment by Arm: Does the Test Make Sense?

The question. Does self-rated physical activity (1-5) at enrollment differ between arms?

Variables and design. Ordinal outcome, three independent groups, measured before the intervention, in a randomized trial.

Chosen test and alternative. Technically, Kruskal-Wallis (alternative: chi-square on the 3 x 5 table, but one cell has very small frequencies, because only 3 patients have level 5).

Hypotheses. H0: the distribution of activity is the same in the three arms; H1: at least one arm differs.

Result. Median 3 in all arms (means of the codes 2.98, 2.58 and 2.61); Kruskal-Wallis H = 5.01, df 2, p = 0.082.

Statistical conclusion. We do not reject H0.

Clinical conclusion. More important than the result is the fact that the question itself is debatable. In a correctly randomized trial, baseline differences are due to chance by definition, and the p value says nothing about the quality of randomization. What matters is whether a difference, however small, is clinically large enough to influence the outcome; if so, it is adjusted for in the analysis (for example through regression), regardless of the p value [56].

Worked Problem 8.13. Blood Glucose at 0, 3 and 6 Months in Arm C

The question. Did fasting plasma glucose change over time in the patients in arm C?

Variables and design. Numerical outcome, the same 54 patients at 3 time points: repeated measurements.

Chosen test and alternative. In arm C, blood glucose at enrollment has a skewed distribution (skewness 1.05; SW p = 0.003), the 0-3 month differences are not normal (SW p = 0.027), and the sphericity assumption of repeated-measures ANOVA is violated (Mauchly p = 0.002). We choose the Friedman test with Wilcoxon post hoc tests and Bonferroni correction; the alternative is repeated-measures ANOVA with the Greenhouse-Geisser correction.

Hypotheses. H0: the distribution of blood glucose is the same at 0, 3 and 6 months; H1: at least one time point differs.

Result. The medians: 100.5 / 89.0 / 85.0 mg/dL. Friedman $\chi^2 = 64.63$, df 2, $p < 0.001$. Post hoc (Bonferroni): 0 versus 3 months $p < 0.001$; 3 versus 6 months $p = 0.011$; 0 versus 6 months $p < 0.001$.

Statistical conclusion. We reject H0; all three comparisons remain significant after correction.

Clinical conclusion. In arm C, blood glucose decreased in the first 3 months and continued to decrease up to 6 months, with a total decrease in the median of 15.5 mg/dL. The comparison with arms A and B shows that the decrease is not explained merely by the passage of time (see Chapter 7, Section 7.6.3).

Worked Problem 8.14. Smoking Status by Sex

The question. Is smoking status associated with sex?

Variables and design. Two nominal variables (sex: 2 categories; smoking: 3 categories), independent subjects: a 2 x 3 table.

Chosen test and alternative. Pearson's chi-square; the smallest expected frequency is 15.26, so the assumption is met. Alternative for small E: Fisher or merging categories.

Hypotheses. H0: smoking and sex are independent; H1: there is an association.

Result. Women: former smokers 16 (18.6%), smokers 30 (34.9%), nonsmokers 40 (46.5%). Men: 17 (23.0%), 23 (31.1%), 34 (45.9%). $\chi^2 = 0.54$, df 2, $p = 0.76$; Cramér's V = 0.06.

Statistical conclusion. We do not reject H0.

Clinical conclusion. The distribution of smoking status is similar in women and men in this sample. In the general population of Romania, smoking prevalence usually differs between the sexes; a sample of patients at cardiometabolic risk recruited in family practices does not necessarily reflect the general population.

Worked Problem 8.15. Diabetes by Sex

The question. Does the prevalence of OGTT-confirmed diabetes differ between women and men?

Variables and design. Two dichotomous variables, independent subjects: a 2 x 2 table.

Chosen test and alternative. The expected frequencies: the smallest is 13.4, so Pearson's chi-square; Fisher as a check.

Hypotheses. H0: $pF = pM$; H1: $pF \neq pM$.

Result. Women 15/86 (17.4%), men 14/74 (18.9%). $\chi^2 = 0.06$, df 1, $p = 0.81$ (Yates $p = 0.97$; Fisher $p = 0.84$).

Statistical conclusion. We do not reject H0.

Clinical conclusion. The prevalence of previously undetected diabetes was practically the same in women and men. In the logistic regression model for diabetes (obesity, age and sex as predictors), sex remains without effect after adjustment as well (OR 0.97, 95% CI 0.40-2.38).

Worked Problem 8.16. Age of at Least 50 Years and Diabetes: Which Version Do We Report?

The question. Is age ≥ 50 years associated with OGTT-confirmed diabetes?

Variables and design. Age dichotomized at 50 years (a cutoff set in advance, clinically meaningful for screening) and diabetes (yes/no); a 2 x 2 table, independent subjects.

Chosen test and alternative. The smallest expected frequency is 6.2 (34 x 29/160): the condition E ≥ 5 is met, but only just. We report chi-square, and Fisher's test as a check.

Hypotheses. H0: p≥50 = p<50; H1: p≥50 ≠ p<50.

Result. Under 50 years: 2/34 (5.9%); 50 years or over: 27/126 (21.4%). $\chi^2 = 4.36$, df 1, p = 0.037; Yates $\chi^2 = 3.38$, p = 0.066; Fisher p = 0.044; φ = 0.17.

Statistical conclusion. With Pearson's test and with Fisher's test, p < 0.05: we reject H0. The more conservative Yates correction gives p = 0.066.

Clinical conclusion. Diabetes was more common after age 50, but the result depends on only 2 cases in the younger group and is fragile. Age treated as a continuous variable in logistic regression (OR 1.11 per year, 95% CI 1.05-1.18) gives a more stable picture than dichotomization. The lesson: dichotomizing a numerical variable loses information.

Worked Problem 8.17. Weight Loss ≥ 5% in Arm C Versus Arm A

The question. Does teleconsultation, added to the app, increase the proportion of patients who lose at least 5% of their weight, compared with counseling alone?

Variables and design. Dichotomous outcome (loss ≥ 5% yes/no), two independent randomized arms (the primary comparison specified in the protocol).

Chosen test and alternative. 2 x 2 chi-square (the smallest expected frequency is above 10) and risk measures: RR, risk difference, NNT.

Hypotheses. H0: pC = pA; H1: pC ≠ pA.

Result. C: 16/54 (29.6%); A: 6/53 (11.3%). $\chi^2 = 5.49$, df 1, p = 0.019 (Yates 4.43, p = 0.035). RR = 2.62 (95% CI 1.11-6.17); absolute difference 18.3 percentage points (95% CI 3.4-33.2); NNT 5.5, rounded up to 6 (95% CI 3.0-29.1).

Statistical conclusion. We reject H0.

Clinical conclusion. For every 6 patients treated with the app and teleconsultation instead of counseling alone, one more patient achieves a weight loss of at least 5%. The confidence interval for the NNT is wide (from 3 to 29), so the size of the benefit is imprecisely estimated.

Worked Problem 8.18. Obesity at Enrollment and at 6 Months: McNemar

The question. Did the proportion of obese patients (BMI ≥ 30) change between enrollment and 6 months?

Variables and design. A dichotomous variable measured twice on the same patients: paired data. BMI at 6 months was calculated from weight at 6 months and height.

Chosen test and alternative. McNemar's test (with continuity correction or the exact version, because the number of discordant pairs is small).

Hypotheses. H0: the proportion of obese patients is the same at the two time points; H1: the proportion differs.

Result. Obese at both time points 47; obese only at enrollment 7; obese only at 6 months 4; non-obese at both 102. The proportion: 54/160 (33.8%) at enrollment, 51/160 (31.9%) at 6 months. χ^2McNemar = (7 - 4)²/11 = 0.82, p = 0.37 (with correction 0.36, p = 0.55; exact p = 0.55).

Statistical conclusion. We do not reject H0.

Clinical conclusion. Although the patients lost on average 2.3 kg (Problem 8.6), the classification as obese changed in very few patients: a modest weight loss does not easily move a patient across the 30 kg/m² threshold. The comparison with Problem 8.6 shows once again how much information is lost through dichotomization.

Worked Problem 8.19. BMI and Systolic Blood Pressure: Association

The question. Is BMI associated with SBP at enrollment?

Variables and design. Two numerical variables measured on the same patients; the aim is association.

Chosen test and alternative. Pearson correlation, with a check of the scatter plot; the alternative, because of the skewness of BMI: Spearman.

Hypotheses. H0: $\rho = 0$ (there is no linear correlation in the population); H1: $\rho \neq 0$.

Result. r = 0.405, p < 0.001; Spearman rho = 0.372, p < 0.001; $R^2 = 0.164$.

Statistical conclusion. We reject H0.

Clinical conclusion. SBP increases with BMI, with a moderate correlation. BMI explains about 16% of the variability in SBP; the remaining 84% is due to other factors (age, smoking, genetics, salt, medication). Correlation alone does not show causality. Details in Chapter 9.

Worked Problem 8.20. Predicting Systolic Blood Pressure from BMI

The question. What SBP do we expect a patient with a BMI of 30 kg/m² to have?

Variables and design. The aim is to predict a numerical variable from another numerical variable.

Chosen test and alternative. Simple linear regression, with a check of the residuals; for several predictors, multiple regression.

Hypotheses. H0: slope $\beta = 0$; H1: $\beta \neq 0$.

Result. SBP = 89.47 + 1.435 x BMI; 95% CI for the slope 0.925-1.944; p < 0.001; standard error of the estimate 12.73 mmHg. Prediction at a BMI of 30: 132.5 mmHg; at a BMI of 35: 139.7 mmHg.

Statistical conclusion. We reject H0: the slope is significantly different from 0.

Clinical conclusion. Each additional unit of BMI is associated, on average, with an SBP about 1.4 mmHg higher. For an individual patient, however, the prediction is imprecise: with a standard error of the estimate of 12.7 mmHg, the real SBP of a patient with a BMI of 30 can easily lie between 108 and 157 mmHg. The model describes the average trend; it does not replace measuring blood pressure.

Worked Problem 8.21. Satisfaction in Arm C Versus Arm B

The question. Does teleconsultation, added to the app, bring greater satisfaction?

Variables and design. Ordinal outcome (1-5), two independent groups.

Chosen test and alternative. Mann-Whitney U. There is no point in a t-test on a score with 5 values.

Hypotheses. H0: the distribution of satisfaction is the same in arms C and B; H1: the distributions differ.

Result. Median 4 (IQR 3-5) in C and 4 (IQR 3-4) in B; U = 1698, p = 0.079; r_rb = 0.19.

Statistical conclusion. We do not reject H0.

Clinical conclusion. Satisfaction tends to be slightly higher in arm C (31% very satisfied, versus 17% in B), but the difference is not statistically significant, and the estimated effect is small. For the decision to offer teleconsultation, the main criterion should be the clinical outcome and the cost, not satisfaction.

Worked Problem 8.22. Educational Level by Area of Residence

The question. Does educational level differ between patients from urban and rural areas?

Variables and design. Education is ordinal (4 levels), area of residence is nominal (2 categories); independent subjects.

Chosen test and alternative. Two approaches are acceptable: chi-square on the 2 x 4 table (ignores the order of the levels) or Mann-Whitney on education (uses the order and is more powerful if the difference is an overall shift).

Hypotheses. H0: the distribution of educational level is the same in urban and rural areas; H1: the distributions differ.

Result. Rural: 7 (10.8%), 22 (33.8%), 30 (46.2%), 6 (9.2%); urban: 13 (13.7%), 35 (36.8%), 34 (35.8%), 13 (13.7%). The smallest expected frequency 7.7. $\chi^2 = 2.04$, df 3, p = 0.56; Cramér's V = 0.11. Mann-Whitney p = 0.69.

Statistical conclusion. We do not reject H0 with either approach.

Clinical conclusion. Educational level is similar in the two settings in this sample. This result would be surprising in a population survey; here, it probably reflects the way the patients were recruited in the practices (a convenience sample), not the demographic reality.

8.9 Recognizing the Wrong Test: Ten "Methods" Descriptions

The excerpts below imitate wording found in articles and in graduation theses. They are fictitious, constructed for practice. For each one, first read the description and try to find the problem yourself, then compare with the explanation.

Description 1. "HbA1c was measured at enrollment and after 6 months of treatment in 45 patients. The values were compared with Student's t-test for independent samples (p = 0.09)." The problem: the same patients, two time points, hence paired data. The test for independent groups ignores the correlation between the two measurements and loses power. Correct: the paired t-test (if the differences are approximately normal) or the Wilcoxon test for paired data.

Description 2. "Length of hospital stay (median 6 days, IQR 4-15) was compared between the two wards (n = 18 and n = 22) using Student's t-test." The problem: the authors themselves describe the variable by the median and a very asymmetric IQR (from 4 to 15 around a median of 6), and the groups are small. The description and the test do not match. Correct: the Mann-Whitney U test, reporting the medians by group.

Description 3. "Systolic blood pressure was compared between the four BMI classes (normal weight, overweight, class I obesity, class II-III obesity) using pairwise t-tests; a p value < 0.05 was considered significant." The problem: four groups mean six comparisons, and the overall type

I error risk exceeds 26%. Correct: one-way ANOVA (or Kruskal-Wallis), followed by post hoc comparisons with correction; if the interest is in a monotonic increase of SBP with BMI class, a test for trend or a regression with continuous BMI.

Description 4. "Postoperative complications occurred in 2 of 15 patients who underwent laparoscopic surgery and in 7 of 14 patients who underwent open surgery (Pearson's chi-square, p = 0.033)." The problem: the expected frequencies in the "complications" column are 4.66 and 4.34, both below 5. Correct: Fisher's exact test, which gives p = 0.050 here, that is, exactly at the limit and, strictly speaking, nonsignificant. The conclusion "laparoscopic surgery significantly reduces complications" is not supported by the data.

Description 5. "Total cholesterol was compared between nonsmokers (n = 40), former smokers (n = 25) and smokers (n = 35) using the Friedman test." The problem: the Friedman test is for repeated measurements on the same subjects, not for independent groups. Correct: one-way ANOVA or, if the distributions are skewed, Kruskal-Wallis.

Description 6. "The prevalence of uncontrolled hypertension at 0, 6 and 12 months in the same 80 patients was compared using a chi-square test on the 3 x 2 table (time point x blood pressure control)." The problem: the table contains 240 "observations" that are actually 80 patients counted three times. Correct: Cochran's Q test for a repeatedly measured binary variable (or McNemar for pairs of time points, with correction).

Description 7. "Agreement between the automatic and the manual blood pressure monitor was excellent (Pearson correlation r = 0.93, p < 0.001)." The problem: correlation measures association, not agreement. A device that systematically reads 10 mmHg higher can have r = 0.99. Correct: the Bland-Altman method (mean of the differences and 95% limits of agreement) and, possibly, the intraclass correlation coefficient [102].

Description 8. "We analyzed 412 teeth from 52 children, comparing the frequency of caries in sealed and unsealed teeth using the chi-square test." The problem: the unit of analysis. The teeth of the same child are not independent (the same hygiene, the same diet, the same saliva), so the real n is much smaller than 412, and the p value is too optimistic [104]. Correct: methods that account for clustering (for example models with a patient effect) or analysis at the patient level.

Description 9. "The Shapiro-Wilk test applied to the whole sample (n = 300, both groups combined) gave p = 0.03, so we used the Mann-Whitney test." The problem: normality is checked in each group, not in the combined groups (if the means differ, their mixture is not normal even when each group is). Moreover, with 150 subjects per group, the t-test is robust to moderate departures. The decision must be based on the shape of the distribution in each group and specified in advance.

Description 10. "In a prospective cohort followed for up to 5 years, with follow-up durations that varied from one patient to another, mortality was compared between those with and without diabetes using chi-square, and the association was expressed as an odds ratio." The problem: when the duration of follow-up varies, the simple proportion of deaths ignores how long each patient was followed and what happened to those lost to follow-up. Correct: survival analysis (Kaplan-Meier curves, the log-rank test, Cox regression with hazard ratios and 95% CI).

8.10 Critical Reading of the Methods Section: Articles by Colleagues in the Discipline

8.10.1 How We Read the Methods Section

When we read an article, we apply the algorithm in Figure 8.1 "in reverse": we start from the declared design and variables and check whether the reported test is the one we would have chosen ourselves. Six questions are enough for a first reading:

1. What is the design (cross-sectional, cohort, case-control, clinical trial, before-after, method comparison)?
2. What is the primary outcome variable, and what type is it?
3. Are the data independent or paired? What is the unit of analysis?
4. Does the named test (or the one suggested by the way the results are reported) match the type of variable and the design?
5. Does the description (mean ± SD or median and IQR) match the test?
6. Is an effect size with a confidence interval reported, and is the effect measure compatible with the design (OR, RR, HR)?

Abstracts have limited space, and many details appear only in the full text. The exercise below works exclusively with the information in the abstract, as recorded in the bibliographic catalog of this book; the critical observations are points for discussion, not verdicts on the articles.

8.10.2 The Reading Table

Table 8.3. Match between design, outcome variable and statistical method in articles by colleagues from our discipline (based on the abstracts).

Article	Design and n	Outcome variable	Test or measure reported	Does it match?	Comment
Padurean et al. 2026 [82]	cross-sectional, 3 groups: 73, 86, 86	self-esteem score (numeric)	ANOVA with Tukey; Cohen's d; η^2; Pearson; multiple regression	yes	$F = 299.00$, $p < 0.001$, $\eta^2 = 0.712$; $d = 2.515$ for hearing aid users versus non-users; model reporting, with effect sizes
Munteanu et al. 2020 [94]	controlled pilot, repeated measurements over one year, n = 93	adherence and control scores (ordinal); exacerbations (binary)	Mann-Whitney; Friedman with Bonferroni; 2 x 2 chi-square	yes	$\chi^2(1) = 13.707$, $p < 0.001$ for 10.30% versus 46.30%; an RR with 95% CI and an effect size for the Mann-Whitney test would be useful
Gherbon et al. 2024 [59]	retrospective before-after, 1 year, n = 200	blood glucose, HbA1c, weight, SBP (numeric, skewed)	medians (IQR) before and after; test not named	probably yes	paired design: Wilcoxon is the suggested test; the medians are called "means" in the abstract; the median of the individual differences is missing

Article	Design and n	Outcome variable	Test or measure reported	Does it match?	Comment
Buda et al. 2019 [86]	before-after one year of perindopril, with a control group; n not stated in the abstract	serum thrombospondin-1 (numeric)	comparison of T0 versus T1, $p < 0.001$	partly	8061.9 ± 3684.80 versus 15380 ± 5887 ng/mL: an SD that is large relative to the mean suggests skewness (Wilcoxon or log transformation); the comparison with the controls (9221.03 ± 6510.21) requires a test for independent groups; n should be stated
Gherbon et al. 2021 [85]	placebo-controlled study, 20 + 20, 3 months	BMI, lipids, blood glucose, DBP (numeric)	between-group comparisons at 3 months and within-group before-after comparisons	partly	the abstract gives p values from both types of comparison; only the between-group comparison (ideally on the changes or by ANCOVA) answers the question about the effect; with 20 patients per group, checking the distribution matters
Danila et al. 2024 [105]	stated as cross-sectional, matching 168 + 168	severity, outcome (binary); inflammatory indices	ROC (AUC 0.624 and 0.752); hazard ratio 2.62 (CI 1.29-5.04)	not entirely	an HR implies follow-up over time (cohort design); percentages with three decimals (70.312%) suggest false precision
Ion et al. 2024 [106]	prospective cohort, n = 141	severe COVID-19 (binary)	ROC at two time points; logistic regression; HR	partly	logistic regression produces an OR, not an HR (SOFA HR 3.07, CI 2.29-4.12); it should be specified whether Cox regression with time to event was used

Article	Design and n	Outcome variable	Test or measure reported	Does it match?	Comment
Cioca et al. 2024 [8]	retrospective cohort, 3 groups: 200, 62, 224	severity, mortality (binary); length of hospital stay (numeric)	proportions, mean ± SD, OR	partly	OR 2.24 (ICU) and 2.46 (ventilation) without CI and without p; length of stay 11.6 ± 7.0 versus 9.1 ± 5.8 days is skewed (median, Mann-Whitney); 3 groups first require an overall test
Stoichescu-Hogea et al. 2021 [90]	3 groups: 62, 46, 40	global longitudinal strain (numeric)	comparison of 3 means with pairwise comparisons; Spearman; ROC	yes, with one remark	different SDs (1.4% in controls versus 3.0% in hypertensive patients): Welch's ANOVA or Games-Howell post hoc tests would have been worth considering
Popa et al. 2025 [100]	retrospective cohort, 26 + 106	symptoms, treatments (binary)	Fisher; OR	yes	small group and rare events, so Fisher is correct; CIs for the ORs are missing (for example OR = 3.97 for headache)
Piroş et al. 2026 [6]	cross-sectional, n = 693	prediabetes (binary)	chi-square across 3 clusters (p = 0.003); trend; logistic regression	yes	ordered categories, trend reported (p-trend < 0.001); the clusters are built from the same data
Moleriu et al. 2026 [4]	cross-sectional, n = 921	excess weight (binary)	chi-square, Cramér's V; logistic regression (aOR)	yes	reports the strength of association (V) alongside p; cross-sectional design, hence association, not causation
Miulescu et al. 2026 [91]	single-arm pilot, n = 21, 3 time points	POEM and SCORAD scores (numeric)	repeated-measures ANOVA	yes, with limitations	POEM 14.67 ± 5.22 to 7.10 ± 3.82; with 21 children, Friedman would be a useful sensitivity analysis; without a control group,

Article	Design and n	Outcome variable	Test or measure reported	Does it match?	Comment
					the effect cannot be attributed to the therapy
Popa et al. 2025 [61]	diagnostic, 69 with stroke and 16 without	GFAP, UCH-L1 (numeric)	comparison of means; diagnostic cut-off	partly	GFAP 142.91 ± 102.19 versus 37.76 ± 19.92 pg/mL; an SD close to the mean indicates skewness: median (IQR) and Mann-Whitney
Christodorescu et al. 2026 [23]	prospective cohort, 12 months, n = 278	primary event over time	Kaplan-Meier with log-rank; multivariable Cox	yes	log-rank p = 0.027, but after adjustment p = 0.240: an example of confounding
Marc et al. 2023 [39]	repeated cross-sectional, 1401 versus 1698	HCV, HBV prevalence (binary); age (numeric)	comparisons of proportions and means	yes	HCV prevalence 16.9% versus 10.5%, p < 0.0001; age 59.7 versus 61 years, p = 0.005: statistically significant, but clinically negligible

A few general conclusions emerge from the table. Most articles use tests that match the design. The problems that occur most often concern not the choice of the main test but reporting: the test is not named, confidence intervals for the ORs are missing, skewed variables are described by mean ± SD, and within-group comparisons are presented as evidence of an effect. These are exactly the points a reviewer checks and the ones you will run into when writing your own graduation thesis.

Case study: model reporting for three groups

Pădurean and colleagues, with Dr. Frandeș and Dr. Muntean, compared self-esteem in 245 people: 73 with normal hearing, 86 hearing aid users and 86 people with hearing loss who do not wear hearing aids [82]. The outcome, a numeric score, is compared across three independent groups, so one-way ANOVA is the natural choice, followed by Tukey's test for the pairwise comparisons. The authors report the overall test ($F = 299.00$, $p < 0.001$), the overall effect size ($\eta^2 = 0.712$) and the effect size for the clinically relevant comparison: hearing aid users had scores of 35.41 ± 5.32 versus 22.99 ± 4.53 in non-users ($p < 0.001$, Cohen's $d = 2.515$). The associations with hearing handicap and tinnitus scores are given by Pearson correlation ($r = -0.573$ and $r = -0.443$), and the multiple regression model explains about 65% of the variability. It is a complete example of the format used in this chapter: an appropriate test, an overall test followed by post hoc tests, effect sizes. The limitation of the design is that of any cross-sectional study: we do not know whether hearing aids raise self-esteem or whether people with higher self-esteem are more willing to wear hearing aids.

8.10.3 When the Design and the Effect Measure Do Not Match

Two of the articles in the table raise the same question: is the effect measure compatible with the stated design? The odds ratio, the relative risk and the hazard ratio are not synonyms. The OR is obtained from 2 x 2 tables or from logistic regression and can be calculated in any design; the RR requires a design in which incidence can be estimated (cohort, clinical trial); the HR requires "time-to-event" data, that is, follow-up of patients over time, and is usually obtained by Cox regression.

Case study: a hazard ratio in a cross-sectional study and in a logistic regression

Danila and colleagues, with Dr. Cioca, evaluated inflammatory indices derived from the complete blood count in 336 patients with COVID-19 (168 with diabetes, 168 without diabetes, matched for sex, BMI and severity) [105]. The study is stated to be cross-sectional. The diagnostic accuracy part is consistent with this design: for dNLR, at the cut-off of 2.685, sensitivity was 70.312%, specificity 65.978%, AUC = 0.624, and for ALRI AUC = 0.752. The abstract, however, also reports a hazard ratio for SII at 7 days, 2.62 (CI 1.29-5.04, $p < 0.001$). A hazard ratio presupposes patients followed over time and event times, which does not fit the "cross-sectional" label; either the design is in fact a retrospective cohort, or the measure reported is a different one (for example an OR). Note also that a CI from 1.29 to 5.04 is rather wide for such a small p value, a detail a reader would check in the full text. Ion and colleagues, also with Dr. Cioca, prospectively followed 141 patients with diabetes and COVID-19 and compared clinical scores using ROC curves (for example SOFA AUC 0.771 at admission and 0.873 on day five) [106]. The methods mention logistic regression, but the results are given as hazard ratios (SOFA HR 3.07, 95% CI 2.29-4.12; APACHE II HR 2.96, 95% CI 2.21-3.96). Logistic regression produces odds ratios; hazard ratios come from Cox regression. Both observations are exercises in careful reading and do not call into question the clinical value of the scores, but they illustrate a useful rule: the name of the effect measure must correspond to the statistical model and to the design.

8.10.4 A Within-Group Comparison Does Not Replace a Between-Group Comparison

A very common interpretive error is to conclude that a treatment works because "a significant decrease from baseline was obtained in the treated group". A before-after test within a group shows only that a change has occurred. The same change may also appear in the control group, through regression to the mean, the placebo effect, seasonality or mere participation in the study. In mg2026, weight decreased significantly even in the arm with simple counseling (paired t-test, $p = 0.043$). Only the comparison between arms (ANOVA, $F = 18.25$, $p < 0.001$) shows the effect of the interventions.

Case study: Aloe ferox and perindopril, within-group and between-group comparisons

Gherbon and colleagues, with Dr. Frandeș, tested in a placebo-controlled study, reported according to the CONSORT guideline, the effect of 2 capsules/day of Aloe ferox (460 mg) in 20 people with obesity, compared with 20 matched people who received diet and placebo [85]. The methods describe two types of comparison: between groups at 3 months and before-after within groups. For the same variables, the abstract gives two series of p values: for example, for BMI $P = 0.03$ and $P = 0.002$, for LDLc $P = 0.01$ and $P = 0.039$, for fasting plasma glucose $P = 0.018$ and $P < 0.001$, to which are added total cholesterol ($P = 0.032$) and diastolic blood pressure ($P = 0.006$). The reader has to check in the full text which value answers which question, because they do not carry the same weight. A decrease from baseline in the treated group does not by itself prove the effect of the supplement: a decrease in BMI may also occur on diet with placebo. The comparison that answers the clinical question is the between-group

one, ideally on the changes from baseline to 3 months or by ANCOVA with the baseline value as covariate. With 20 patients per group, the choice between the t-test and Mann-Whitney must be justified by the shape of the distributions. Buda and colleagues, with Dr. Tudor, compared serum thrombospondin-1 in hypertensive patients with endothelial dysfunction before and after one year of perindopril: 8061.9 ± 3684.80 versus 15380 ± 5887 ng/mL, $p < 0.001$ [86]. The data are paired, so the appropriate test is the paired t-test or Wilcoxon, applied to the individual differences. A standard deviation of almost half the mean suggests a skewed distribution, so the nonparametric option or log transformation should be considered. The control group without hypertension (9221.03 ± 6510.21 ng/mL) raises a second question, with a different test: the comparison with the controls is made with a test for independent groups. The abstract does not state the number of patients, information without which no p value can be interpreted.

8.11 Proposed Exercises

For each of situations 1-6, specify: the type of outcome variable, the design (independent or paired, number of groups), the chosen test and its alternative, and the statistical hypotheses.

1. A family physician wants to find out whether the number of practice visits per year (0, 1, 2,..., with many patients with 0-2 visits and a few with more than 15) differs between patients with and without diabetes.
2. Oxygen saturation was measured in 25 patients with COPD at rest and after the 6-minute walk test.
3. The proportion of patients vaccinated against influenza is compared across three counties (450, 380 and 520 patients).
4. A sleep quality questionnaire (score 0-21) was administered to the same group of 30 residents after a night shift, after a day off and after a week of vacation.
5. Two radiologists independently classified 120 chest radiographs as "pneumonia" or "no pneumonia". What measure do you use to assess their agreement?
6. The aim is to estimate the risk of developing diabetes as a function of the FINDRISC score, age and sex, in a cohort followed for 10 years.
7. In one study, the authors compared HbA1c across 3 types of treatment using three t-tests and found $p = 0.03$, $p = 0.20$ and $p = 0.04$. Calculate the overall risk of type I error for three independent comparisons at $\alpha = 0.05$ and the Bonferroni-corrected threshold. Which comparisons remain significant after correction?
8. A 2 x 2 table has 12 and 48 patients in the rows, and 10 and 50 in the columns. Calculate the smallest expected frequency and decide whether you apply chi-square or Fisher.
9. On the mg2026 dataset, solve completely, in the a-g format, the question: "Does total cholesterol differ between men and women?" (in JASP, T-Tests > Independent Samples T-Test).
10. On the mg2026 dataset, solve in the a-g format the question: "Is diabetes confirmed by OGTT associated with area of residence?"
11. Read the abstract of a recent article in the specialty that interests you (for example via PubMed) and fill in for it a row of the kind shown in Table 8.3.

Common Pitfalls

- Choosing the test according to the type of the grouping variable, forgetting the type of the outcome variable (for example chi-square for "blood glucose by sex", just because sex is nominal).

- Confusing paired data with independent groups, especially when the same patients appear in two columns of the spreadsheet.
- Checking normality on the total sample instead of in each group (or on the differences, for paired data).
- Switching automatically to a nonparametric test when Levene's $p < 0.05$: the natural solution is the Welch version.
- Reporting the test that "came out" instead of the one specified in advance.
- Testing baseline differences between the arms of a randomized trial and interpreting them as a "failure of randomization".
- Interpreting a before-after comparison within a group as evidence of efficacy.
- Using correlation to demonstrate agreement between two methods.
- Counting eyes, teeth or visits as if they were independent patients.

Key Points

- The order of the decision: the aim (difference, association, agreement, prediction), the type of outcome variable, the number of groups or time points, independent or paired, then the distribution and the sample size.
- For each situation there is a parametric option, a nonparametric one and a categorical one (Figure 8.1 and Table 8.1).
- The unit of analysis is the patient (or the randomized unit); multiple observations from the same patient are not independent.
- The test, the primary variable and the fallback rule are specified in the protocol, before seeing the data.
- The a-g format: research hypothesis, statistical hypotheses, method and justification, solution and software, result with a graph, statistical conclusion, research conclusion.
- A complete result includes the test statistic, the degrees of freedom, the exact p value and an effect size with a confidence interval; the effect measure must match the design (OR, RR, HR).

Self-Assessment Questions

1. (SC) A study measures blood pressure in 40 patients before and after 8 weeks of treatment; the differences are normally distributed. The appropriate test is:
    a) the independent-samples t-test
    b) the paired t-test
    c) the Mann-Whitney test
    d) the chi-square test
    e) one-way ANOVA
2. (SC) Pain scores (0-10) are compared across three analgesic regimens, with 12 patients per regimen. The appropriate test is:
    a) the Friedman test
    b) the Wilcoxon test for paired data
    c) the Kruskal-Wallis test
    d) three Mann-Whitney tests without correction
    e) the McNemar test
3. (SC) To assess agreement between the portable glucometer and the laboratory analyzer on the same 80 samples, the most appropriate method is:
    a) Pearson correlation
    b) the independent-samples t-test

c) Bland-Altman analysis

d) the chi-square test

e) logistic regression

4. (SC) The overall risk of at least one type I error when 3 independent comparisons are made, each at $\alpha = 0.05$, is approximately:

a) 5%

b) 10%

c) 14%

d) 15%

e) 20%

5. (SC) In a cohort followed for 5 years, with different follow-up durations, the effect of a factor on mortality is most correctly expressed by:

a) an odds ratio from a 2 x 2 table

b) a hazard ratio from a Cox regression

c) the Spearman correlation coefficient

d) Cramér's V

e) the Kruskal-Wallis H statistic

6. (MC) The data are paired in the following situations:

a) blood glucose of the same patient at 0 and at 6 months

b) visual acuity in the right eye and in the left eye of the same patients

c) HbA1c in men and in women

d) cases and controls individually matched for age and sex

e) SBP in smokers and in non-smokers

7. (MC) Regarding the choice between the t-test and the nonparametric option, the following are correct:

a) normality is checked in each group, not on the total sample

b) for the paired t-test, the normality of the differences is checked

c) a Shapiro-Wilk test with $p < 0.05$ at $n = 300$ always requires a nonparametric test

d) unequal variances (Levene $p < 0.05$) are usually handled with the Welch version

e) the test is chosen after seeing which option gives $p < 0.05$

8. (MC) An article describes a randomized trial with three arms and reports $p < 0.05$ for the decrease in blood glucose within each arm. The following are correct:

a) the result shows that each intervention produces a change from baseline

b) the result does not show that one intervention is superior to the others

c) the appropriate comparison is between arms (for example ANOVA on the changes)

d) the within-arm p values prove the efficacy of all interventions

e) regression to the mean may contribute to the observed decreases

9. (MC) The unit of analysis is violated (the observations are not independent) if:

a) 400 teeth from 40 patients are analyzed as 400 independent observations

b) 170 eyes from 102 patients are analyzed without taking the patient into account

c) each patient contributes a single value

d) hospital admissions are counted instead of patients, and some patients have several admissions

e) patients are individually randomized and each appears only once in the analysis

10. (MC) The standardized a-g format includes:

a) the statistical hypotheses in words and in mathematical form

b) the justification for the choice of test

c) the software and the method of solution

d) only the p value, without descriptive statistics

e) the research conclusion stated in words

Chapter 9. Analyzing Relationships between Variables: Correlation and Linear Regression for Numeric and Ordinal Variables

Objectives

By the end of this chapter, the student will be able to:

- recognize a bivariate analysis problem and name the predictor (X) and the response variable (Y);
- construct and read a scatter plot, distinguishing between independent, functionally dependent and correlated variables;
- calculate by hand the covariance, the Pearson correlation coefficient and the Spearman coefficient on a small dataset;
- interpret the sign, the strength (according to Colton's rule) and the statistical significance of a correlation coefficient, together with its confidence interval;
- explain what the coefficient of determination R^2 means and answer the question "what proportion of the variation is due to other factors";
- estimate the regression line by the least squares method, interpret the slope and the intercept and make a prediction, recognizing the danger of extrapolation;
- check the assumptions of linear regression by residual analysis;
- read a multiple linear or logistic regression table from a medical article;
- argue why a correlation, however strong, does not by itself prove a causal relationship and does not measure the agreement between two methods.

9.1 Bivariate Analysis: Two Variables Measured on the Same Patient

So far we have worked almost always with one numeric variable and one or more groups: blood pressure in women and in men, blood glucose in the three arms of the study, blood glucose at 0 and at 6 months. In this chapter we change the question. For each patient we have two numeric measurements, for example body mass index (BMI) and systolic blood pressure (SBP), and we want to know whether they vary together. Do patients with a higher BMI also have, on average, a higher blood pressure? If so, how close is the relationship, and by how much does blood pressure increase for each unit of BMI?

An analysis that follows the behavior of two variables simultaneously is called bivariate analysis. By convention, the variable that we believe influences the other one, or that we use to predict the other one, is denoted by X and is called the predictor (explanatory variable, independent variable). The variable we want to explain or predict is denoted by Y and is called the response variable (dependent variable, outcome variable). In our example, BMI is X and SBP is Y. The choice is not always obvious: between fasting plasma glucose and glycated hemoglobin (HbA1c) there is no "natural predictor", since both reflect the same process, and the choice of X depends on the clinical question.

Two methods answer two different questions. Correlation measures the direction and strength of the association between two numeric variables, treated symmetrically: both are considered random, and the correlation coefficient between X and Y is the same as that between Y and X. Regression builds a mathematical model that describes how Y changes, on average, when X changes, and allows Y to be predicted for a given value of X. In regression the roles are not

symmetric: the line of Y as a function of X is not the same as the line of X as a function of Y. The classic textbooks of our discipline present the two methods together, because they use the same sums of squares and each one checks the other [18,42,53].

9.1.1 The Scatter Plot

Every bivariate analysis begins with a graph, not with a number. The scatter plot has the variable X on the horizontal axis and the variable Y on the vertical axis; each patient is a point. When looking at such a graph, the student should ask four questions: what shape the cloud of points has (elongated along a straight line, curved, shapeless), in which direction it goes (increasing or decreasing), how tightly the points are gathered around an imaginary line, and whether there are isolated points, far from the rest.

Figure 9.1 shows the three patterns described in the lectures of our discipline. In panel A the points are scattered uniformly, without any direction: knowing X tells us nothing about Y, and the variables are independent. In panel B all the points lie exactly on a curve: Y is completely determined by X, and we have a functional dependence (of the type of the relationship between the side and the area of a square). In biology and medicine, such a perfect relationship almost never occurs. Panel C is the usual situation: the points gather around a straight line, but with visible scatter. Y tends to increase with X, yet for the same value of X patients have different values of Y. We say that the variables are correlated: the relationship is statistical (stochastic), not functional.

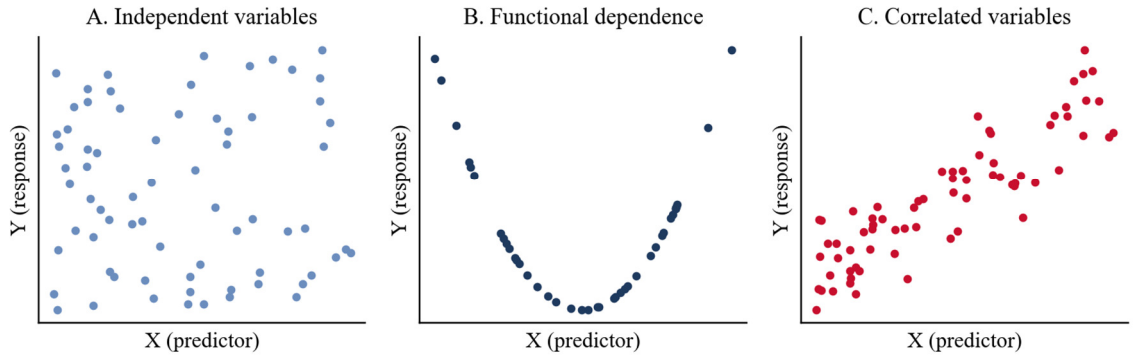


Figure 9.1. Patterns of relationship between two numeric variables. A: independent variables, the cloud of points has no direction. B: functional dependence, each value of X determines exactly one value of Y. C: correlated variables, Y increases on average with X, but with scatter around the trend. Simulated data.

9.1.2 What the Graph Can and Cannot Tell Us

The graph shows things that the coefficients hide. A curved relationship (for example, mortality as a function of BMI, increased both in underweight and in obese people, with the shape of the letter U) may have a Pearson coefficient close to zero, even though the relationship between the variables is strong. A single patient with extreme values can create an apparent correlation or hide one. Two distinct subgroups (men and women, for example, in the height-weight relationship) can together produce an elongated cloud, even if within each subgroup the relationship is weak. This is why, in the practical sessions, we do not accept any reported correlation coefficient without the corresponding scatter plot.

On the mg2026 teaching dataset (160 adults at cardiometabolic risk, Chapter 1) we will follow several pairs of variables throughout the chapter. We announce them now so that you can compare the results: BMI and SBP (weak but clearly significant correlation), age and SBP (weak correlation), fasting plasma glucose and HbA1c (moderate, almost strong correlation), HDL cholesterol and triglycerides (no correlation), self-rated physical activity and BMI (ordinal variable, no correlation).

9.2 Covariance: The First Step Toward a Measure of Association

Suppose that for each patient we calculate the deviation of X from the mean of X and the deviation of Y from the mean of Y. If the patient has both values above the mean, the product of the deviations is positive. If both values are below the mean, the product is still positive (minus times minus). If one is above and one below the mean, the product is negative. By summing the products over all patients, we obtain a number that is large and positive when the variables increase together, large and negative when one increases while the other decreases, and close to zero when there is no common trend.

Dividing this sum by n - 1, we obtain the sample covariance:

$$\text{cov}(X,\ Y) = \frac{\sum(x_i - m_x)\,(y_i - m_y)}{n-1}$$

where mx and my are the means of the two variables. From now on we denote the sum of the products of the deviations by Sxy, and the sums of the squared deviations by Sxx and Syy:

$$S_{\text{xy}} = \sum(x_i - m_x)\,(y_i - m_y);\ S_{\text{xx}} = \sum(x_i - m_x)^2;\ S_{\text{yy}} = \sum(y_i - m_y)^2$$

Note that Sxx / (n - 1) is precisely the variance of X from Chapter 2. Covariance generalizes variance: the variance is the covariance of a variable with itself.

Covariance has a serious drawback for interpretation: it depends on the units of measurement. If we express weight in grams instead of kilograms, the covariance between height and weight becomes 1000 times larger, although the biological relationship is the same. Looking at a covariance of 43, we cannot say whether the association is strong or weak. The solution is standardization: we divide the covariance by the product of the standard deviations and obtain a unitless number, the correlation coefficient.

9.3 The Pearson Correlation Coefficient

9.3.1 The Formula and Hand Calculation

The Pearson linear correlation coefficient is denoted by r for the sample and by ρ (rho) for the population:

$$\text{r} = \text{cov}\frac{X, Y}{s_x \times s_y} = \frac{S_{\text{xy}}}{\sqrt{S_{\text{xx}} \times S_{\text{yy}}}}$$

The two forms are equivalent, because the n - 1 factors cancel out. For hand calculation, the second form is more convenient.

Let us take a small example, with figures chosen so that the calculation can be followed without a calculator. Eight patients from a family practice have the BMI and SBP values in Table 9.1.

Table 9.1. Hand calculation of the Pearson coefficient for 8 patients $\left(X = \text{BMI},\ \frac{\text{kg}}{m^2};\ Y = \text{SBP},\ \text{mmHg}\right)$.

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
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

Patient	xi	yi	xi - mx	yi - my	(xi - mx)(yi - my)	$(x_i - m_x)^2$	$(y_i - m_y)^2$
8	37	148	8	17	136	64	289
Total	232	1048	0	0	$S_{xy} = 302$	$S_{xx} = 146$	$S_{yy} = 684$

The steps are as follows. The means: $m_x = \frac{232}{8} = \frac{29\ \text{kg}}{m^2}$ and $m_y = \frac{1048}{8} = 131$ mmHg. The deviations are calculated row by row, and their sum must be zero in each column; if it is not zero, you have made a subtraction error. Then the deviations are multiplied and squared. The covariance is 302 / 7 = 43.14 kg/m² x mmHg (a unit of measurement that is hard to imagine, which confirms the discussion above). The standard deviations are $s_x = \sqrt{\frac{146}{7}} = \frac{4.57\ \text{kg}}{m^2}$ and $s_y = \sqrt{\frac{684}{7}} = 9.89$ mmHg. Finally:

$$r = \frac{302}{\sqrt{146 \times 684}} = \frac{302}{\sqrt{99\,864}} = \frac{302}{316.01} = 0.956$$

All the products are positive or zero: every patient with a BMI above the mean also has an SBP above or at the mean. The coefficient is close to 1, a strong positive correlation. On real data you will not obtain such values very often; the example was constructed for clarity. We will continue with the same eight pairs in regression (Section 9.7) and in Worked Problem 9.1.

9.3.2 Properties of the Pearson Coefficient

The coefficient r takes values between -1 and +1. The sign indicates the direction of the association: r > 0 means a direct (positive) correlation, the two variables increase together; r < 0 means an inverse (negative) correlation, when one increases the other tends to decrease. The value r = +1 corresponds to a perfect direct correlation (all points on an increasing straight line), r = -1 to a perfect inverse correlation, and r = 0 to the absence of a linear relationship.

Several properties are important for the exam. The coefficient has no unit of measurement and does not change if we transform one of the variables linearly (weight in grams or in kilograms, temperature in degrees Celsius or Fahrenheit give the same r). It is symmetric: r(X, Y) = r(Y, X). It measures only the linear relationship; a strong curved relationship may have a small r. It is sensitive to extreme values, because it is based on means and on squared deviations. And, very importantly, it depends on the range of values in the sample: if we study the BMI-SBP relationship only in patients with a BMI between 27 and 29, r will be much smaller than in the complete sample, because we have cut off precisely the variation that makes the relationship visible (the phenomenon is called range restriction).

9.3.3 How Strong Is a Correlation?

Figure 9.2 shows what the clouds of points look like for a few values of r. It is worth memorizing visually: an r of 0.5 looks much more "scattered" than students usually expect, while an r of 0.95 appears as a narrow band.

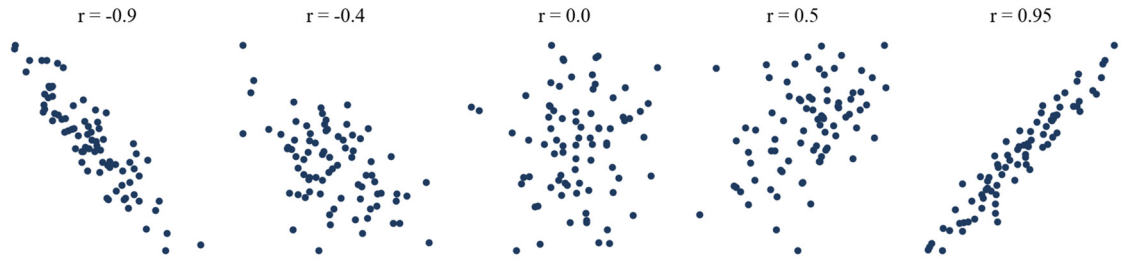


Figure 9.2. Simulated clouds of points for r = -0.9; -0.4; 0; 0.5 and 0.95 (80 points each). The strength of the correlation is read from how tightly the points are gathered around a straight line, not from its slope.

To turn the number into words, the lectures of our discipline use the rule of thumb proposed by Colton in 1974. The absolute value of r is interpreted as follows:

Table 9.2. Colton's rule of thumb for interpreting the correlation coefficient.

Value of r	Interpretation
$-0.25 < r < 0.25$	no correlation (negligible correlation)
$0.25 \leq r < 0.50$ or $-0.50 < r \leq -0.25$	weak correlation
$0.50 \leq r < 0.75$ or $-0.75 < r \leq -0.50$	moderate correlation
$r \geq 0.75$ or $r \leq -0.75$	strong correlation
$r = 1$ or $r = -1$	perfect correlation

In mg2026, the BMI-SBP correlation ($r = 0.405$) is weak, the age-SBP correlation ($r = 0.256$) is weak, lying right at the lower limit, and the glucose-HbA1c correlation ($r = 0.738$) is moderate, very close to the 0.75 threshold. HDL and triglycerides ($r = 0.076$) are not correlated.

Colton's rule is not the only one. In psychology and the behavioral sciences, Cohen's scale is frequently used; it considers $r = 0.1$ a small effect, $r = 0.3$ a medium effect and $r = 0.5$ a large effect [76]. Other textbooks use thresholds in steps of 0.2 (very weak, weak, moderate, strong, very strong). On Cohen's scale, the $r = 0.405$ between BMI and SBP would be a "medium to large" effect, not a "weak" correlation. There is no right scale and no wrong one: there are conventions. What matters is that you state which scale you use and that you do not change the scale depending on which result suits you. The exam of our discipline uses Colton's rule.

9.3.4 Conditions for Applying the Pearson Coefficient

The Pearson coefficient can be calculated for any pair of numeric columns, but its interpretation and the significance test require three conditions. First, the relationship between the variables must be linear, which is checked on the scatter plot. Second, for the test and for the confidence interval, the pairs must come from a bivariate normal distribution; in practice, we check that each variable has an approximately normal distribution (histogram, Q-Q plot, Shapiro-Wilk test, Chapter 3). Third, there must be no extreme values that dominate the calculation. To these is added the general condition of independence of observations: each patient appears only once, not several times with repeated measurements.

Figure 9.3 shows why the third condition matters so much. In panel A, 25 simulated points have no relationship with one another ($r = -0.16$). In panel B a single point has been added, far from all the others, and r has jumped to 0.64. A student who reports only the number would say that there is a moderate correlation; the graph shows that the "correlation" is produced by a single patient.

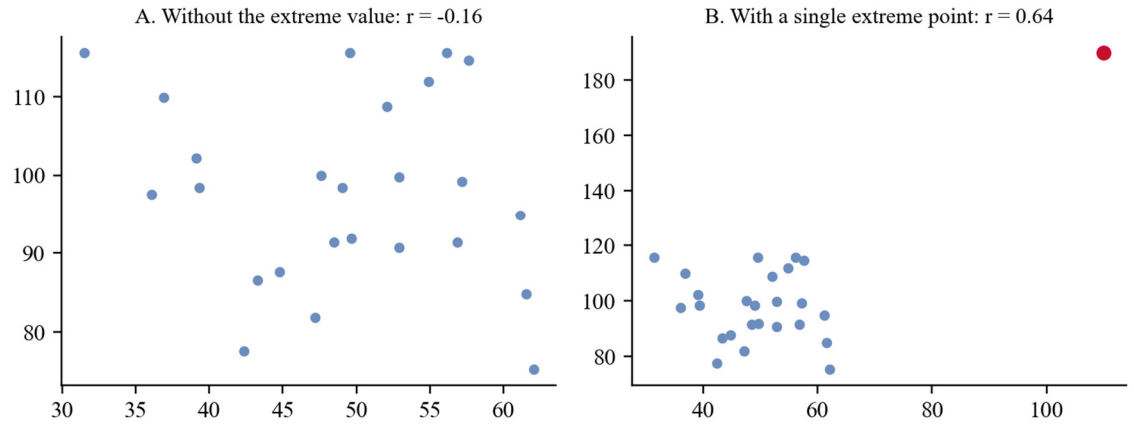


Figure 9.3. The influence of a single extreme value on the Pearson coefficient. A: 25 points with no relationship, r = -0.16. B: the same points plus a single extreme point (red), r = 0.64. Simulated data.

What do we do with an extreme value? First we check that it is not a data entry error (a height of 1.72 entered in meters instead of 172 cm, as we saw in the file mg2026_brut.csv). If the value is real, we report the analysis with and without it and prefer a rank-based coefficient (Spearman), which is much less sensitive. Deleting inconvenient patients without justification is not an option.

In mg2026, BMI does not satisfy normality (skewness 1.42; Shapiro-Wilk $p < 0.001$), because of a few patients with severe obesity (the maximum is 48.3 kg/m²). SBP is normally distributed ($p = 0.303$). Therefore, for the BMI-SBP pair we will report both Pearson and Spearman, and we will see that the two lead to similar conclusions.

9.3.5 Testing the Significance of the Correlation Coefficient

An r calculated on a sample is never exactly zero, even if in the population the variables have no relationship at all. The inferential question is: could the observed value arise from sampling variation alone, if in the population $\rho = 0$? The hypotheses are:

H0: $\rho = 0$ (in the population, the two variables are not linearly correlated) H1: $\rho \neq 0$ (in the population, the variables are correlated)

Under H0, the test statistic follows a Student distribution with n - 2 degrees of freedom:

$$t = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}}$$

We lose two degrees of freedom because we have estimated two means. For the example with eight patients: $t = \frac{0.956 \times \sqrt{6}}{\sqrt{1-0.913}} = \frac{0.956 \times 2.449}{0.294} = 7.95$. The critical value for 6 degrees of freedom and two-sided $\alpha = 0.05$ is 2.447; since $7.95 > 2.447$, H0 is rejected ($p = 0.0002$).

In mg2026, for BMI-SBP: $t = \frac{0.405 \times \sqrt{158}}{\sqrt{1-0.164}} = \frac{0.405 \times 12.57}{0.914} = 5.56$, with 158 degrees of freedom, $p < 0.001$. For age-SBP: $t = 3.34$, $p = 0.001$. For HDL-triglycerides: $t = 0.96$, $p = 0.34$, H0 is not rejected.

The formula hides a lesson you will often meet in articles: significance depends heavily on n. Table 9.3 shows the smallest r that becomes significant at $\alpha = 0.05$ for different sample sizes.

Table 9.3. Minimum value of |r| that is significant at $\alpha = 0.05$ (two-sided test), as a function of n.

n	10	20	30	50	100	160	867

n	10	20	30	50	100	160	867
minimum significant r	0,632	0,444	0,361	0,279	0,197	0,155	0,067

With 10 patients, a moderate correlation of 0.60 is not significant. With 867 patients, a correlation of 0.07, which Colton would call "no correlation", is significant. Statistical significance answers only the question "is ρ different from zero?", not the question "is the relationship important?". We will see a published example in Section 9.5.

9.3.6 The Confidence Interval for r: The Fisher Transformation

The sampling distribution of r is not symmetric when ρ is far from zero: r cannot exceed 1, so the values "crowd" toward the end. The statistician R. A. Fisher proposed a transformation that makes the distribution approximately normal:

$z = 0.5 \times \ln\left[\frac{1+r}{1-r}\right]$, with standard error $\mathrm{SE}(z) = \frac{1}{\sqrt{n-3}}$

The interval is calculated on the z scale, $z \pm 1.96 \times \mathrm{SE}(z)$, and then the limits are transformed back into r with the inverse formula $r = \frac{e^{2z}-1}{e^{2z}+1}$. For BMI-SBP in mg2026: $z = 0.5 \times \ln\left(\frac{1.405}{0.595}\right) = 0.429$; $\mathrm{SE} = \frac{1}{\sqrt{157}} = 0.0798$; the interval on the z scale is 0.429 ± 0.156, that is, 0.273 to 0.586; transformed back, the 95% CI for ρ is 0.266-0.527.

The interval tells us more than the p value. We now know that, very probably, the correlation in the population is at least weak (above 0.25) and at most moderate (below 0.53). For glucose-HbA1c, the 95% CI is 0.659-0.802: the correlation is at least moderate and may be strong. For HDL-triglycerides, the 95% CI is -0.080 to 0.229 and contains zero, in agreement with p = 0.34. For the sample of 8 patients, although r = 0.956, the interval is wide: 0.769-0.992, because n - 3 = 5 is very small.

9.3.7 The Coefficient of Determination R^2

By squaring r we obtain the coefficient of determination, R^2 (or r^2). It shows what proportion of the variation in Y is "explained", in a statistical sense, by the linear relationship with X. It takes values between 0 and 1 (or between 0% and 100%) and no longer keeps the sign: a correlation of -0.6 and one of +0.6 have the same $R^2 = 0.36$. This is why we say that R^2 informs us about the strength of the correlation, not about its direction.

For BMI and SBP in mg2026, $R^2 = 0.405^2 = 0.164$. About 16.4% of the variability in systolic blood pressure between patients can be attributed to differences in BMI; the remaining 83.6% is due to other factors (age, sex, smoking, dietary salt, genetics, medication, variation from one measurement to another). This is exactly the question asked in the practical session of our discipline on height and weight: "what proportion of the variation in weight is caused by factors other than the variation in height?". The answer is 1 - R^2. In mg2026, height and weight have r = 0.652, so $R^2 = 0.424$, and $1 - R^2 = 0.576$: about 57.6% of the variation in weight is due to factors other than height.

For blood glucose and HbA1c, $R^2 = 0.545$: slightly more than half of the variation in HbA1c is linearly associated with the variation in fasting plasma glucose. For age and SBP, $R^2 = 0.066$, that is, only 6.6%. A "weak" r of 0.26 means, translated into R^2, that age explains very little of the differences in blood pressure among our patients. Students are often surprised at how small a moderate r becomes after squaring: r = 0.5 explains only a quarter of the variation.

The last point concerns language. "Explained" does not mean "caused". The expression used in the classic problem statements ("the variation in weight caused by other factors") must be read

statistically: the proportion of the variation that is not linearly related to height. The fact that height "explains" 42% of the variation in weight does not mean that 42% of someone's kilograms come from their height.

Case study: correlations in liver elastography (the Timișoara ultrasonography group)

Colleagues from our discipline who work with the gastroenterology team have published several studies in which a new ultrasonography method is compared with a reference method, and correlation is the first result reported. In a diagnostic accuracy study of 64 people (37 with cirrhosis, 27 healthy volunteers), the parameter obtained by contrast-enhanced ultrasound (CEUS-PAT) had a correlation of r = 0.81 with vibration-controlled transient elastography (VCTE) [2]. In a prospective study of 165 patients, time-harmonic elastography had, against the same reference, r = 0.82, p < 0.0001 [3]. For steatosis, the UGAP parameter had r = 0.73, p < 0.0001 with CAP, in 179 consecutive subjects [58], and in a 2026 study of 169 patients, the ultrasound fat fraction (USFF) had r = 0.81, p < 0.001 with CAP, while two other parameters had r = 0.66 and r = 0.83 [65].

By Colton's rule, values of 0.81, 0.82 and 0.83 indicate a strong correlation, and 0.73 and 0.66 a moderate correlation. The abstract of the last study nevertheless calls a correlation of 0.83 "moderate", which illustrates precisely the problem of different scales discussed above: authors may use a different threshold, and the reader must judge the number, not the adjective. The second lesson is more important. A correlation of 0.82 between two methods of measuring liver stiffness does not mean that they give the same values and can be used interchangeably. Correlation shows that patients are ranked similarly; agreement is assessed differently (Section 9.6), and the 2020 study reported, for reproducibility, intraclass correlation coefficients of 0.92-0.97 [3]. What else would be useful in such an abstract: the confidence interval for r and the scatter plot, which would show whether the relationship remains linear over the whole range of fibrosis.

9.4 Correlation for Ordinal or Skewed Data: The Spearman and Kendall Coefficients

9.4.1 The Idea: We Work with Ranks, Not Values

The Pearson coefficient assumes numeric variables, a linear relationship and an approximately normal distribution. Many medical variables do not meet these conditions. Scores on Likert scales, disease stages, NYHA class, educational level or self-rated physical activity (1-5) in mg2026 are ordinal: we know the order of the categories, but the distance between "2" and "3" is not necessarily the same as that between "4" and "5". Other variables are numeric but highly skewed (triglycerides, BMI in mg2026) or have extreme values.

Spearman's rank correlation coefficient (denoted rs or Spearman's ρ) solves the problem with a simple device: it replaces each value with its rank (its position in ascending order) and calculates the Pearson coefficient on the ranks. The result measures how monotonic the relationship is: if Y increases whenever X increases, even nonlinearly, the Spearman coefficient is +1. Like all the rank-based methods in Chapter 7, it is robust to extreme values: the largest BMI receives rank 160, regardless of whether it is 40 or 48.3.

9.4.2 Hand Calculation

When there are no ties (tied values), Spearman's coefficient can be calculated with a short formula:

$$r_s = 1 - \frac{6 \times \sum {d_i}^2}{n \times (n^2 - 1)}$$

where di is the difference between the rank of X and the rank of Y for patient i. Let us take eight patients in whom the FINDRISC score and fasting plasma glucose were measured.

Table 9.4. Calculation of the Spearman coefficient for the FINDRISC score and fasting plasma glucose in 8 patients.

Patient	FINDRISC	Blood glucose (mg/dL)	Rank X	Rank 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

$$r_s = 1 - \frac{6 \times 8}{8 \times (64 - 1)} = 1 - \frac{48}{504} = 1 - 0.095 = 0.905$$

The correlation is positive and strong: patients with a higher FINDRISC score tend to have higher blood glucose. The order is not perfectly preserved (pairs 1-2, 3-4, 5-6, 7-8 swap places), but the deviations are of a single rank. For small n, the significance of rs is read from special tables or from software (here p = 0.002); for n above roughly 10-20, the same t formula as for Pearson can be used.

9.4.3 Ties (Tied Values)

Ordinal variables almost always have tied values. Patients with the same value receive the mean of the ranks they would have occupied: if two patients have activity level 2 and would occupy positions 2 and 3, both receive rank 2.5. With many ties, the short formula becomes inaccurate and Pearson must be calculated on the ranks (which all software does automatically). Worked Problem 9.2 shows an example in which the difference between the two calculations is large: -0.85 with the short formula versus -0.91 calculated correctly.

9.4.4 Kendall's Tau Coefficient

An alternative, less used in clinical articles but preferred by some statisticians for small samples and for many ties, is Kendall's tau (τ) coefficient. Its idea is to count pairs: for any two patients we ask whether their order by X is the same as their order by Y (concordant pair) or the reverse (discordant pair). Tau is the difference between the proportion of concordant pairs and that of discordant pairs. In the example in Table 9.4 there are 28 possible pairs of patients; only the four neighboring pairs that swap places are discordant, so τ = (24 - 4) / 28 = 0.71. Values of τ are generally smaller than those of Spearman's coefficient for the same data (in mg2026, for BMI-SBP, τ = 0.258 versus rs = 0.372), so Colton's thresholds do not apply directly to τ. The τ-b version, reported by JASP and other software, corrects for ties.

9.4.5 What the Rank Coefficients Show in mg2026

Self-rated physical activity (1-5) is ordinal, so its relationship with BMI is assessed with Spearman's coefficient: rs = -0.012, p = 0.88. The sign is the expected one (more activity, slightly

lower BMI), but the value is practically zero: in our sample, how much patients say they move tells us nothing about their BMI. The possible explanations have to do with the quality of the measurement (self-rating is imprecise) and with the fact that the sample includes only overweight people. Between educational level and satisfaction at 6 months, rs = 0.125, p = 0.115: neither a correlation according to Colton, nor significance.

For BMI and SBP, where BMI is skewed, Spearman gives rs = 0.372 ($p < 0.001$), very close to Pearson's r = 0.405. When the two coefficients are close, extreme values do not distort the result much. When they differ greatly, the graph must be examined carefully.

Case study: rank coefficients in validation and medical education studies

Four articles by our colleagues show how rank correlations are reported correctly. In a diagnostic accuracy study of 157 patients with chronic hepatitis B or C, with liver biopsy as the reference standard, stiffness measured by two methods was correlated with the fibrosis stage (an ordinal variable, F0-F4) using the Spearman coefficient: $\rho = 0.826$, $p < 0.001$, with a 95% CI for rho of 0.769-0.870 [107]. It is one of the few abstracts that give the confidence interval, and in Worked Problem 9.4 we will check by hand that the interval corresponds to the Fisher transformation. In a validation study of a questionnaire on the acceptability of mobile technology in diabetes, in 103 patients, the authors reported both Spearman coefficients ($\rho = -0.429$; $\rho = -0.322$; $\rho = 0.466$; all with $P < 0.001$) and a Kendall coefficient ($\tau = 0.51$; $P < 0.001$) [54]; note that tau can be large even though it is generally smaller than rho.

A study using an anonymous questionnaire in 121 medical students used the median and interquartile range for scores on a 1-10 scale and Spearman correlation [16]. Two correlations from the abstract deserve to be placed side by side: $R = 0.869$, $p < 0.001$, and $R = 0.107$, $p = 0.246$, the latter between measured and self-assessed knowledge. The second correlation is practically zero and not significant: what students think they know says little about what they actually know. For a biostatistics course, the result is an invitation to modesty, and statistically it shows that a nonsignificant result is also informative, provided the sample is not tiny. Finally, a study on COVID-19 surveillance data from Timiș County found Spearman correlations of rho = 0.852 and 0.821 ($p < 0.001$) between incidence among healthcare workers and the trend in the community [40]. Here the unit of analysis is not the person but the time period: two time series that rise and fall together. The correlation is strong, but it does not tell us where a particular healthcare worker was infected; we discuss this pitfall in the next section.

9.5 Correlation Does Not Mean Causation

9.5.1 Four Explanations for a Correlation

When two variables are correlated, there are at least four possible explanations, and direct causation is only one of them. The first is chance: with enough pairs of variables tested, some will turn out significantly correlated purely through sampling variation (with $\alpha = 0.05$, one in 20 pairs with no real relationship). The second is reverse causation: Y may influence X, not the other way around. Patients with coronary heart disease engage in less physical exertion because they have angina, not only the other way around. The third is confounding: a third variable (a confounder) influences both variables and creates a correlation between them without a direct link. Ice cream sales and the number of drownings are correlated because both increase in summer. In mg2026, age is correlated both with SBP and with the risk of diabetes; a crude correlation between SBP and blood glucose should be checked after adjusting for age. The fourth explanation is causation proper, which must be argued, not assumed.

To weigh causal arguments in observational studies, the epidemiologist Austin Bradford Hill proposed in 1965 a set of criteria: strength of association, consistency across studies, specificity,

temporality (the cause precedes the effect), biological gradient (dose-response relationship), plausibility, coherence with existing knowledge, experimental evidence and analogy [108]. Only temporality is mandatory. The criteria are discussed at length in Chapter 10; here we note that a correlation coefficient, however large, covers at most the strength of association and, sometimes, the gradient.

9.5.2 The Ecological Fallacy

Correlations calculated on aggregate data (countries, counties, months) can be completely different from those at the individual level. If countries with higher fat consumption have higher coronary mortality, it does not automatically follow that people who eat more fat die more often of coronary heart disease; perhaps other factors also differ in those countries (income, smoking, access to health services). The mistaken inference from the group to the individual is called the ecological fallacy. The correlation between COVID-19 incidence among healthcare workers and that in the community, discussed above, is an ecological correlation of this type: useful for epidemiological surveillance, but without clarifying the source of infection of each employee [40].

9.5.3 Regression to the Mean

A statistical phenomenon that has fooled many generations of physicians is regression to the mean. If we select patients because they had an extreme value at one measurement (for example, SBP above 160 mmHg at a single visit), at the next measurement their values will be, on average, closer to the population mean, even without any treatment. The reason: every measurement also contains a random component (stress, coffee, a cuff applied in a hurry), and those selected for an extreme value had, on average, an unfavorable random component, which does not repeat itself. The lower the correlation between the two measurements, the greater the regression to the mean. The phenomenon explains why many treatments appear to work in studies without a control group and why randomization, with a control group measured in the same way (Chapter 13), is the only reliable protection. The term "regression" itself comes from here: in the 19th century, Francis Galton observed that the sons of very tall parents are, on average, less tall than their parents.

9.5.4 Statistical Significance and Clinical Importance

The last pitfall in this section concerns sample size. We saw in Table 9.3 that, with hundreds of patients, very small correlations become significant. A small p shows that ρ is probably different from zero, but not that the relationship matters clinically. Always report r (or R^2) together with p and, ideally, with the confidence interval.

Case study: a significant but negligible correlation

A retrospective study of 867 patients with colon cancer, operated on in a university hospital in Timişoara over four years (two before and two during the COVID-19 pandemic), examined the changes in surgical care [55]. The main results are comparisons of proportions between periods: a decrease in elective procedures ($p < 0.001$), an increase in emergency procedures ($p < 0.001$) and an increase in stage IV cases in the second year of the pandemic ($p = 0.045$). The abstract also reports an inverse correlation between patient age and duration of surgery: $r = -0.088$, $p = 0.01$.

Let us check. With $n = 867$, $t = \frac{0.088 \times \sqrt{865}}{\sqrt{1-0.0077}} = \frac{0.088 \times 29.41}{0.996} = 2.60$, and for 865 degrees of freedom $p = 0.0095$, that is, approximately 0.01, in agreement with the abstract. The correlation is therefore statistically significant. However, according to Colton, $|r| = 0.088 < 0.25$ means "no correlation", and $R^2 = 0.0077$: age explains less than 1% of the variation in the duration of surgery. What the result says: in this series, older patients did not have longer operations, but, very slightly, shorter ones. What it does not say: that age is a useful predictor

> of duration. The result is a very good teaching example of the difference between "significant" and "important"; a careful reader would also want to see the scatter plot and, possibly, an analysis adjusted for the type of procedure (emergency or elective), which may be a confounder.

9.6 Correlation and Agreement: The Bland-Altman Method

A frequent error in the medical literature, pointed out as early as 1986 by Bland and Altman, is the use of the correlation coefficient to show that two measurement methods "match" [102]. Suppose that a new blood pressure monitor systematically reads 10 mmHg higher than the reference device. The points will lie perfectly on a straight line parallel to the line of identity, r will be almost 1, but the methods are not interchangeable: a patient with 135 mmHg on the reference device would be classified as hypertensive on the new one. Correlation measures how well patients are ranked, not how close the values are.

Mg2026 contains a natural example. SBP at baseline and SBP at 6 months have r = 0.835 (strong correlation, 95% CI 0.781-0.877), but the mean decreased from 131.1 to 126.4 mmHg. Patients keep their place in the ranking, but the values are not the same; a high correlation does not exclude a systematic difference.

The Bland-Altman method, which assesses agreement, has three steps. For each patient, the difference between the two methods and their mean are calculated. The mean of the differences, called the bias or systematic error, and the standard deviation of the differences are calculated. The limits of agreement are calculated: the mean of the differences $\pm 1.96 \times$ SD of the differences. Approximately 95% of the individual differences lie between these limits. The Bland-Altman plot has the mean of the two methods on the horizontal axis and the difference on the vertical axis, with three horizontal lines (the bias and the two limits).

A short numerical example: eight patients measured with a portable glucometer and in the laboratory have the differences (glucometer minus laboratory) 4, -2, 6, 3, 8, 1, 5 and -1 mg/dL. The mean of the differences is 3.0 mg/dL, and the SD of the differences is 3.46 mg/dL. The limits of agreement are $3.0 \pm 1.96 \times 3.46$, that is, -3.8 to 9.8 mg/dL. If a difference of up to 10 mg/dL is clinically acceptable, the device can be used for monitoring; if not, it cannot. The decision is clinical, not statistical, and the correlation coefficient plays no part in it. For categorical variables, agreement is measured with the kappa coefficient [103], discussed together with diagnostic tests.

9.7 Simple Linear Regression

9.7.1 The Model

Correlation tells us how tightly the points gather around a straight line. Regression finds the line. The simple linear regression model writes each value of Y as a part explained by X plus an error:

$$y = a + b\,x + \varepsilon$$

In this notation, a is the intercept, that is, the predicted value of Y when X = 0; b is the slope (the regression coefficient), that is, the mean change in Y for a one-unit increase in X; ε is the residual error, the part of Y that the model does not explain. The sign of b is always the same as the sign of r. The line itself, $\hat{y} = a + b\,x$, gives the mean predicted value of Y for a given value of X; a particular patient will lie above or below it, at a distance equal to their residual, $e_i = y_i - \hat{y}_i$.

9.7.2 The Least Squares Method

Countless straight lines can be drawn through the cloud of points. Which is "the best"? The classic answer, given by Gauss and Legendre at the beginning of the 19th century, is the line for which the sum of the squared residuals is minimal:

$$\text{SSE} = \sum e_i{}^2 = \sum (y_i - a - b\,x_i)^2 = \text{minimum}$$

Squaring ensures that positive and negative residuals do not cancel out and penalizes large deviations more heavily. Solving the minimization problem yields the formulas:

$$b = \frac{S_{\text{xy}}}{S_{\text{xx}}} = r \times \left(\frac{s_y}{s_x}\right) \quad a = m_y - b \times m_x$$

The first form of the slope is used for hand calculation, the second when we have only summary statistics (means, standard deviations and r), a frequent situation when reading articles. The second formula implies that the line always passes through the point of the means (mx, my).

For the eight patients in Table 9.1: $b = \frac{302}{146} = 2.068$ mmHg for each kg/m², and a = 131 - 2.068 x 29 = 71.01 mmHg. The line is:

$$\text{SBP predicted} = 71.01 + 2.068 \times \text{BMI}$$

At a BMI of 30 kg/m², the predicted SBP is 71.01 + 2.068 x 30 = 133.1 mmHg.

9.7.3 Interpreting the Slope and the Intercept

The slope is the clinical result of the regression. "For each additional kg/m², SBP is, on average, higher by 2.07 mmHg" is a sentence any clinician understands. Pay attention to units: if X had been expressed in other units (for example, weight in grams), the slope would have changed accordingly, although the relationship would have remained the same. Pay attention to the meaning as well: this is a difference between patients, observed cross-sectionally, not a promise that a patient who loses 1 kg/m² will lower their blood pressure by 2 mmHg. That second statement would be a causal claim, which only an intervention study can support.

The intercept often has an absurd interpretation. In our example, 71 mmHg would be the blood pressure "of a patient with a BMI of zero", who does not exist. The intercept is a mathematical parameter needed to position the line, but it has clinical meaning only if X = 0 lies within the range of the data (for example, a zero dose of a drug). When you want an interpretable intercept, you can center the variable X (subtract the mean from each value); the intercept then becomes the predicted value for a patient with X equal to the mean.

9.7.4 The mg2026 Example: BMI and SBP

For the 160 patients in mg2026, the regression of SBP on BMI gives the results in Table 9.5, and Figure 9.4 shows the cloud of points, the line and the confidence band of the line.

Table 9.5. Simple linear regression of SBP (mmHg) on BMI (kg/m²), mg2026, n = 160.

Coefficient	Estimate	Standard error	t	p	95% CI
Intercept (a)	89,47	7,56	11,84	< 0,001	74,54-104,40
BMI (b)	1,435	0,258	5,56	< 0,001	0,925-1,944

$R^2 = 0.164$; standard error of the estimate (SEE) = 12.73 mmHg; F(1, 158) = 30.93, p < 0.001.

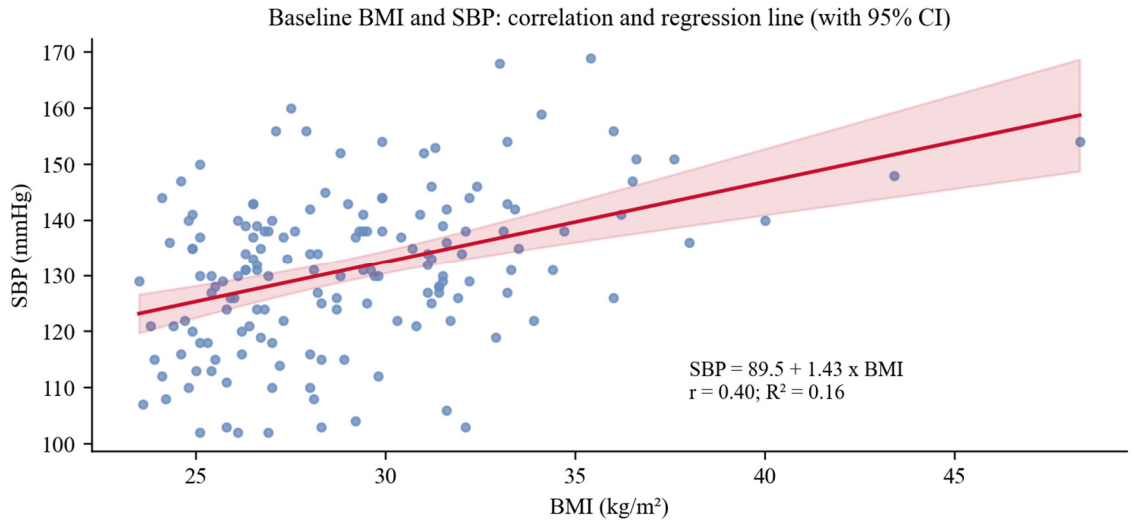


Figure 9.4. Body mass index and systolic blood pressure at baseline in the mg2026 dataset (n = 160): each point is a patient, the red line is the regression line SBP = 89.47 + 1.435 x BMI, and the shaded band is the 95% confidence interval for the predicted mean value. Synthetic data (mg2026 dataset).

The equation is SBP = 89.47 + 1.435 x BMI. For each additional kg/m², SBP is on average higher by 1.4 mmHg; the confidence interval for the slope (0.925-1.944) does not contain zero, so the slope is significantly different from zero. It is no coincidence that the t for the slope (5.56) is exactly the t from the test for r: in simple regression, the test of H0: b = 0 and the test of H0: ρ = 0 are the same test, and $F = t^2 (5.56^2 = 30.93)$. The predictions are: at a BMI of 30, SBP = 89.47 + 1.435 x 30 = 132.5 mmHg; at a BMI of 35, SBP = 139.7 mmHg. The difference between them, 7.2 mmHg, is 5 x 1.435.

As a check, the slope can also be obtained from the summary statistics: $b = \frac{r \times s_y}{s_x} = \frac{0.405 \times 13.88}{3.91} = 1.44$, and a = 131.13 - 1.435 x 29.04 = 89.46 (the differences in the second decimal come from rounding).

For age, the equation is SBP = 108.3 + 0.406 x age ($R^2 = 0.066$): for every additional 10 years, SBP is on average higher by about 4 mmHg. For blood glucose and HbA1c: HbA1c = 3.38 + 0.0209 x glucose, that is, 0.21 percentage points of HbA1c for every 10 mg/dL of blood glucose.

9.7.5 The Standard Error of the Estimate and R^2

Two numbers describe how well the line fits. The coefficient of determination R^2 is a relative measure: the proportion of the variation in Y explained by the model. It can be written as

$$R^2 = 1 - \frac{\text{SSE}}{S_{yy}}$$

that is, 1 minus the ratio between the variation left unexplained after drawing the line and the total variation of Y around its mean. For the eight patients, SSE = 59.3 and Syy = 684, so $R^2 = 1 - \frac{59.3}{684} = 0.913$, exactly $r^2 = 0.956^2$.

The standard error of the estimate (SEE, also called the residual standard deviation) is an absolute measure, expressed in the units of Y:

$$\text{SEE} = \sqrt{\left[\frac{\text{SSE}}{n-2}\right]}$$

It shows how far from the line a patient typically lies. For the eight patients, SEE $= \sqrt{\frac{59.3}{6}} =$ 3.14 mmHg. For mg2026, SEE = 12.73 mmHg, very close to the standard deviation of SBP (13.88 mmHg). The comparison is instructive: if we did not know the BMI, the best estimate of SBP would be the mean, with a typical scatter of 13.9 mmHg; knowing the BMI, the scatter drops only to 12.7 mmHg. BMI explains a real but small part of the variation. As R^2 increases, SEE decreases, but the two say different things: R^2 answers the question "how much does the model explain?", and SEE the question "how precise is an individual prediction?". For a clinician who wants to estimate a patient's blood pressure, the second question is often the more important one.

9.7.6 The Confidence Interval for the Slope

The slope estimated on the sample has its own standard error:

$$\mathrm{SE}(b) = \frac{\mathrm{SEE}}{\sqrt{S_{\mathrm{xx}}}}$$

and the 95% confidence interval is $b \pm t(0.975;\ n-2) \times \mathrm{SE}(b)$. For the eight patients, $\mathrm{SE}(b) = \frac{3.14}{\sqrt{146}} = 0.260$, and t for 6 degrees of freedom is 2.447; $\mathrm{CI}\,95\% = 2.068 \pm 2.447 \times 0.260 = 2.068 \pm 0.637$, that is, 1.43-2.71 mmHg per kg/m². For mg2026, SE(b) = 0.258, and with t ≈ 1.975 for 158 degrees of freedom, $\mathrm{CI}\,95\% = 1.435 \pm 0.510 = 0.925 - 1.944$. The interval is much more useful than p: it tells us that, in the population from which the sample comes, the difference in SBP associated with one kg/m² plausibly lies between about 0.9 and 1.9 mmHg.

9.7.7 Residuals and Checking the Assumptions

Linear regression rests on four assumptions, easy to remember by their initials (LINE): the relationship is linear (L), the observations are independent (I), the residuals are normally distributed (N) and they have the same variance over the whole range of X (E, for equal variance, homoscedasticity). Note that normality is required of the residuals, not of the variable X. X may even be skewed, as BMI is in mg2026.

The assumptions are checked on the residuals. Figure 9.5 shows the two standard plots for the SBP-BMI model. In panel A, the residuals are plotted against the predicted values: if the model is adequate, the points form a horizontal band, without curvature and without a funnel shape, centered on zero. A curve would indicate a nonlinear relationship; a funnel (scatter that increases with the predicted values) would indicate heteroscedasticity. In panel B, the histogram of the residuals should have approximately the shape of a symmetric bell; in mg2026, the Shapiro-Wilk test on the residuals gives p = 0.878, so there is no reason to reject normality.

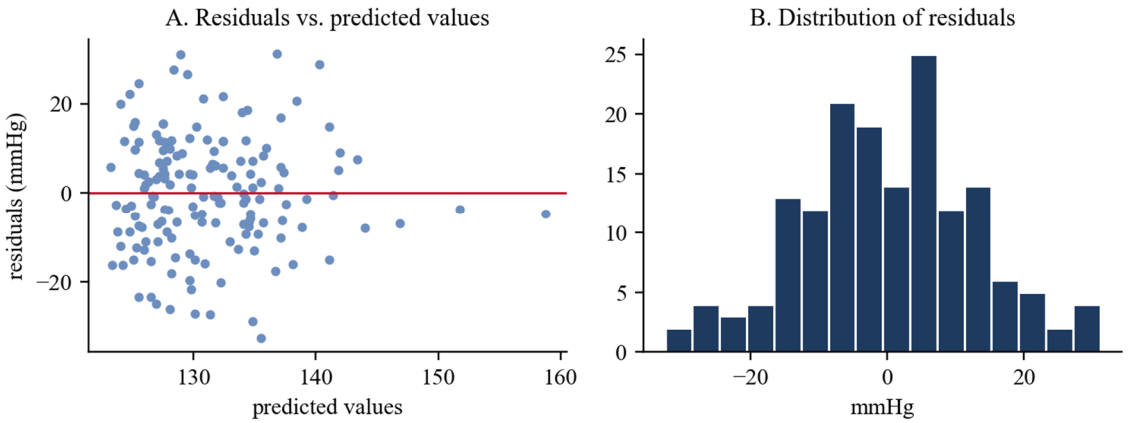


Figure 9.5. Residual analysis for the regression SBP = 89.47 + 1.435 x BMI (mg2026). A: the residuals plotted against the predicted values form a horizontal band, without curvature or a funnel shape. B: the histogram of the residuals is approximately symmetric (Shapiro-Wilk p = 0.878). Synthetic data (mg2026 dataset).

Independence cannot be seen on these plots; it follows from the study design: each patient must appear only once. If we put SBP at 0 and at 6 months for the same patients into the same regression, as if they were 320 independent observations, we would violate this assumption and obtain p values that are too small. For time-ordered data, software reports the Durbin-Watson statistic (values close to 2 indicate no autocorrelation of the residuals; in mg2026, in the multiple model in Section 9.8, the value is 1.97).

9.7.8 Prediction and the Danger of Extrapolation

Once the line has been estimated, prediction seems simple: we substitute X and obtain Y. Two subtleties, however, separate a good student from an average one.

The first concerns the type of uncertainty. The value of 132.5 mmHg predicted for a BMI of 30 is an estimate of the mean blood pressure of all patients with a BMI of 30. For this mean, the 95% confidence interval is narrow: 130.5-134.6 mmHg. For a single new patient with a BMI of 30, the uncertainty is much greater, because individual variation is added (SEE = 12.7 mmHg): the 95% prediction interval is 107.3-157.7 mmHg. In other words, the model knows fairly well what the mean is, but it cannot say whether a particular patient with a BMI of 30 will be normotensive or hypertensive. For a BMI of 35, the confidence interval of the mean is 136.1-143.3 mmHg, and the prediction interval 114.3-165.1 mmHg. The confidence band in Figure 9.4 widens toward the ends, because we have fewer data there.

The second subtlety is extrapolation. The line was estimated on patients with a BMI between 23.5 and 48.3 kg/m², and we have no guarantee that the relationship remains linear outside this range. The practical session of our discipline on height and weight deliberately asks what weight a person 1.78 m tall and one 2.35 m tall should have. In mg2026, the line is weight = -100.40 + 1.087 x height (cm). At 178 cm, the predicted weight is -100.40 + 1.087 x 178 = 93.1 kg, a legitimate prediction, because 178 cm lies within the range of the data (145-192 cm). At 235 cm, the formula gives 155.0 kg. The number is just as easy to calculate, but it has no support in the data: nobody in the sample exceeds 192 cm, and people over 2.30 m often have conditions (gigantism, acromegaly) that change the relationship completely. Figure 9.6 makes the problem visible: the red dotted line is extrapolation.

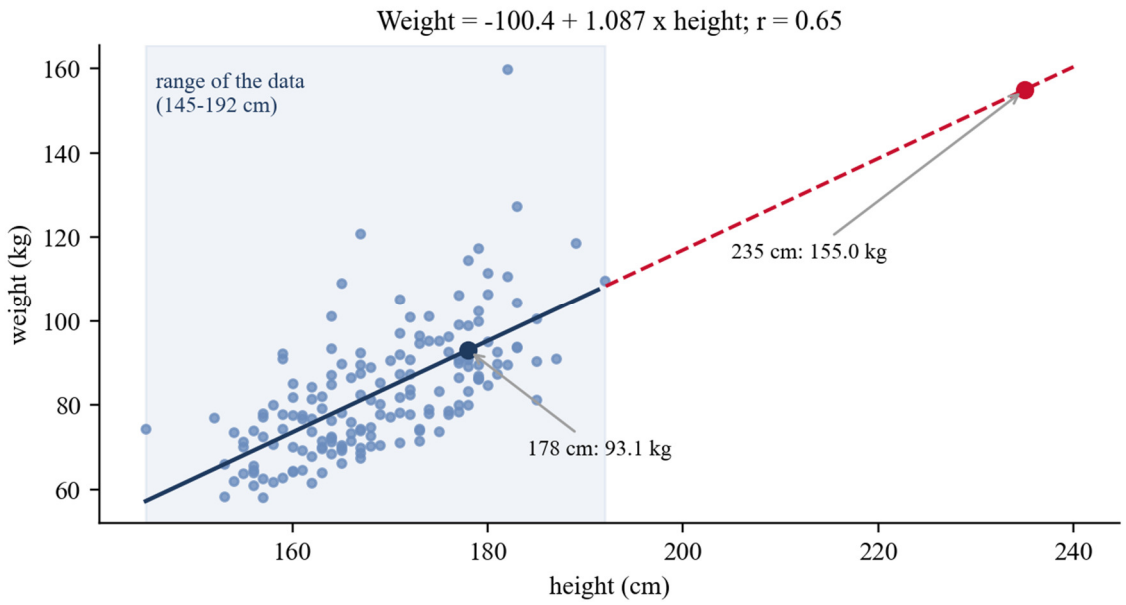


Figure 9.6. Height and weight in the mg2026 dataset (n = 160), with the regression line weight = -100.4 + 1.087 x height (r = 0.65). The shaded area is the range of the data (145-192 cm). The prediction for 178 cm (93.1 kg) is an interpolation; the one for 235 cm (155.0 kg) is an extrapolation without support in the data. Synthetic data (mg2026 dataset).

An even clearer example is BMI-SBP: if we enter a BMI of 60 kg/m², the formula gives 175.6 mmHg. It may be true, it may not; our data tell us nothing about patients with extreme obesity. In the exam, the question "what value does the model predict for..." must almost always be accompanied by a comment on the range within which the model is valid.

9.8 Multiple Linear Regression: An Introduction

9.8.1 Why We Need Several Predictors

SBP does not depend only on BMI. Age, sex, smoking, salt intake and many other factors contribute. Moreover, the predictors are related to one another: older patients may also have a higher BMI, and part of the "crude" association between BMI and SBP might in fact come from age. Multiple linear regression estimates the effect of each predictor while holding constant the other predictors in the model:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + \varepsilon$$

Each coefficient bj is read as follows: "the mean change in Y for a one-unit increase in xj, in patients with the same values of the other variables in the model". The expression "adjusted for" in articles means exactly this. The calculation is done only with software, but every physician who reads the literature should know how to read the results.

9.8.2 Categorical Variables in Regression

Sex and smoking are not numeric. They enter the model through indicator variables (dummy variables), coded 0/1. For sex a single variable is enough (1 = man, 0 = woman); its coefficient is the mean difference between men and women, at the same values of the other predictors. Smoking has three categories, so it needs two indicator variables; the category that receives no indicator becomes the reference category, with which the others are compared. In the model below, the reference is "former smoker". The choice of reference does not change the model, only the comparisons displayed. Usually the most numerous or the most clinically natural category is chosen (here it could have been the non-smokers).

9.8.3 The mg2026 Example: SBP as a Function of BMI, Age, Sex and Smoking

Table 9.6. Multiple linear regression of SBP at baseline (mmHg), mg2026, n = 160.

Predictor	Coefficient (B)	Standard error	95% CI	p
Intercept	64,51	9,39	45,96-83,06	< 0,001
BMI $\left(\frac{\text{kg}}{m^2}\right)$	1,389	0,240	0,915-1,862	< 0,001
Age (years)	0,449	0,108	0,235-0,662	0,0001
Male sex (versus female)	-1,481	1,878	-5.19 to 2.23	0,43
Smoker (versus former smoker)	6,738	2,636	1,53-11,95	0,012
Non-smoker (versus former smoker)	-1,119	2,487	-6.03 to 3.79	0,65

$R^2 = 0.308$; R^2 adjusted $= 0.286$; F(5, 154) = 13.72, p < 0.001.

Let us read the table row by row. At the same age, the same sex and the same smoking status, each additional kg/m² is associated with an SBP higher by 1.39 mmHg, very close to the crude slope of 1.435. Age does not "absorb" the effect of BMI, which means that, in mg2026, age is not an important confounder of the BMI-SBP relationship. Each year of age adds, at the same BMI, 0.45 mmHg (4.5 mmHg per decade). Men have an SBP 1.5 mmHg lower than women, but the interval (-5.2 to 2.2) contains zero and p = 0.43: we have no evidence for a difference between the sexes, in agreement with the t-test in Chapter 6. Current smokers have an SBP 6.7 mmHg higher than former smokers (CI 1.5-11.9; p = 0.012), and non-smokers do not differ significantly from former smokers.

With an estimated model we can also make predictions for a specific patient. A 60-year-old male smoker with a BMI of 32 has a predicted SBP = 64.51 + 1.389 x 32 + 0.449 x 60 - 1.481 + 6.738 = 141.2 mmHg. A female former smoker of the same age and with the same BMI has 64.51 + 1.389 x 32 + 0.449 x 60 = 135.9 mmHg.

9.8.4 Adjusted R^2

R^2 increases every time we add a predictor, even one with no relationship to Y (for example, the last digit of the personal identification number), because any variable "captures" a small part of the random variation. Adjusted R^2 penalizes the number of predictors:

$$R^2 \text{ adjusted } = 1 - \frac{(1 - R^2) \times (n - 1)}{n - k - 1}$$

where k is the number of predictors (here 5, because smoking counts twice). For mg2026: $1 - \frac{(1-0.308)\times 159}{154} = 1 - 0.692 \times 1.0325 = 0.286$. The small difference between R^2 and adjusted R^2 shows that the model does not have too many predictors relative to the sample size. A frequently cited rule of thumb requires at least 10-15 observations for each predictor; with 160 patients and 5 predictors we are far from the limit.

The multiple model explains 30.8% of the variation in SBP, almost twice as much as BMI alone (16.4%). Nevertheless, almost 70% remains unexplained. In clinical research, R^2 values of 0.2-0.4 for biological variables are usual; this does not mean that the model is "bad", but that blood pressure has many causes that we have not measured.

9.8.5 Unstandardized and Standardized Coefficients

Articles sometimes report the unstandardized coefficient B (in units of Y per unit of X, as in Table 9.6), and sometimes the standardized coefficient β (beta), obtained after transforming all the variables into z scores. Beta is read as "by how many standard deviations Y changes for one standard deviation of X" and allows predictors measured in different units to be compared. Unfortunately, the notation is not uniform: some authors use β to denote the unstandardized coefficient itself. When you read an article, check in the methods section or from the units (a β expressed in cm or in kg/m² is unstandardized).

9.8.6 Collinearity

If two predictors are very strongly correlated with each other, the model cannot separate their effects, and the coefficients become unstable and have large standard errors. In mg2026, BMI and waist circumference have r = 0.885: both measure adiposity. Entering them simultaneously in a model for SBP would produce coefficients that are hard to interpret. Software reports the variance inflation factor (VIF); values above 5-10 signal collinearity.

9.8.7 How to Read a Regression Table in an Article

For a regression table in an article, a useful reading order is the following. Identify the dependent variable and the type of regression (linear for a numeric Y, logistic for a binary Y, Cox for time to event). Check how many observations actually entered the model (patients with missing data are usually excluded) and how many predictors there are. Read the coefficient of each predictor of interest together with its unit of measurement and its confidence interval; a coefficient without a unit cannot be interpreted. Look for the reference categories of the categorical variables. Look at R^2 (or adjusted R^2) to judge how much the model explains. Finally, ask yourself which variables are missing: an omitted confounder can make a coefficient appear significant. Reporting guidelines for observational studies explicitly require a description of the adjustment variables and of the way they were chosen [25].

Case study: multiple regression in our colleagues' articles

A cross-sectional study of 220 adults with chronic dermatological diseases analyzed treatment adherence (MARS-5 score, 21.50 ± 4.19) as a function of satisfaction with pharmaceutical services and social support [72]. The abstract reports an unstandardized coefficient B = 0.076 (95% CI 0.017-0.135; p = 0.011) and an increase in explained variance from 13.3% to 28.0% after other variables were added to the model; one association reached borderline significance (p = 0.053) after adjustment. The lesson: R^2 increases as relevant predictors are added, and a coefficient can be attenuated when we introduce a variable that mediates or confounds the relationship. Interpreting B requires the scale of the predictor: a coefficient of 0.076 seems small, but if the predictor is a score with a wide range (the PSPSQ-2 satisfaction score is 62.92 ± 9.57), the difference between patients at the two ends of the scale may become relevant. Without a unit, the coefficient says nothing; with a unit, it becomes a clinical quantity.

A multivariate cross-sectional study of 245 people compared self-esteem in users of smart hearing aids, in non-users and in people with normal hearing [82]. In addition to ANOVA ($F = 299.00$, $p < 0.001$, $\eta^2 = 0.712$), the authors report Pearson correlations: self-esteem correlates negatively with hearing handicap ($r = -0.573$, $p < 0.001$) and with tinnitus ($r = -0.443$, $p < 0.001$), and the two scales correlate positively with each other ($r = 0.729$, $p < 0.001$). The multiple regression model, with a coefficient of about 11.9 for wearing a hearing aid, explains about 65% of the variation in self-esteem. Such a high R^2 is unusual in clinical research and is

explained by the very large difference between the groups. The correlation of 0.729 between two potential predictors also signals the risk of collinearity, if both enter the same model.

In cardiology, a case-control study of 60 patients with ischemic heart failure and 60 age-matched controls reported higher osteoprotegerin values in the patients (4.7 ± 0.25 versus 1.3 ± 0.67 ng/mL) and lower vitamin D values (20.49 ± 7.31 versus 37.09 ± 4.59 ng/mL), followed by very strong Pearson correlations of the markers studied with end-diastolic volume (r = 0.862), with ejection fraction (r = -0.832) and with pulse wave velocity (r = 0.833), all with $p < 0.001$, a moderate correlation with vitamin D (r = -0.636) and a weak one with the glomerular filtration rate (r = -0.488) [87]. The authors went on with a multiple regression in which vitamin D was the dependent variable. Note that, because the study includes both patients and healthy people, part of the strong correlations may come from the difference between the two groups: within each group, the relationship could be much weaker (the situation of the two subgroups in Section 9.1.2).

A retrospective cohort study of 186 patients after acute myocardial infarction used multiple linear regression for ventricular remodeling as a function of the number of classes of guideline-directed medical therapy (GDMT) [109]. The results are given as β = 0.120 cm (p = 0.004) and β = 0.204 cm ($p < 0.001$) for each additional class. The unit (cm) shows that these are unstandardized coefficients, interpreted directly as the difference in diameter for each additional treatment class. Only 18.8% of the patients received all four therapeutic pillars, which limits precision at the upper end of the scale.

Two studies by the nutrition group illustrate linear regression in nutritional epidemiology. In a prospective cohort of 921 students followed during their first year at university, the mean change in BMI was -0.45 ± 1.07 kg/m² (95% CI -0.52 to -0.38), and a chrononutrition behavior was associated with an adjusted difference of $\beta = +\frac{0.22\ \text{kg}}{m^2} (p = 0.004)$ [5]. In a cross-sectional study of 693 students, dietary patterns were associated with HbA1c: β = 0.147 ($p < 0.001$) for one pattern and β = -0.109 ($p < 0.001$) for another, in the opposite direction [6]. Both abstracts report coefficients with the p value, without a confidence interval; for a reader, the interval would have shown how precisely the effect is estimated.

Finally, regression can be used to build a clinical score. In a study of 124 patients with compensated cirrhosis, the authors combined spleen stiffness (SS), liver stiffness (LS) and spleen size (SSZ) into a score obtained by multiple regression: 0.053 x SS + 0.054 x LS + 0.059 x SSZ - 1.84, with a cut-off > 0.34 and an area under the ROC curve of 0.93, larger than that of each parameter taken separately (0.84; 0.86; 0.71) [110]. The formula has exactly the structure b0 + b1 x1 + b2 x2 + b3 x3 from Section 9.8.1. Its validity in patients other than those on whom it was built must be checked separately, in a validation cohort.

9.9 Logistic Regression: A Bridge to Chapter 10

9.9.1 When Y Has Only Two Values

What do we do when the outcome variable is binary: diabetes yes/no, death yes/no, weight loss of at least 5% yes/no? Linear regression does not fit: it could predict negative "probabilities" or ones greater than 1, and the residuals cannot be normal. Logistic regression models not the probability p directly, but the logarithm of the odds (log odds or logit):

$$\ln\left[\frac{p}{1-p}\right] = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

The odds are the ratio between the probability of having the event and the probability of not having it; a probability of 20% means odds of 0.2 / 0.8 = 0.25, that is, "1 to 4". Chapter 10

explains this notion at length. From a practical point of view, what you need to remember now is that the exponential of a coefficient, e^b, is an odds ratio (OR): the factor by which the odds of the event are multiplied when the predictor increases by one unit, with the other predictors held constant.

9.9.2 The mg2026 Example: Diabetes Confirmed by OGTT

Table 9.7. Logistic regression of diabetes confirmed by OGTT (29 cases out of 160), mg2026.

Predictor	Adjusted OR	95% CI	p
Obesity (BMI ≥ 30), versus BMI < 30	6,01	2,35-15,39	0,0002
Age, per year	1,11	1,05-1,18	0,0003
Male sex, versus female	0,97	0,40-2,38	0,95

Reading: at the same age and the same sex, the odds of having diabetes are about 6 times higher in obese patients than in non-obese patients; the wide interval (2.35-15.39) reflects the small number of cases (29). Each year of age multiplies the odds by 1.11, that is, an increase of 11% per year, at the same weight status and the same sex. For 10 years, the effects multiply rather than add: 1.11^10 ≈ 2.8, so a patient 10 years older has almost three times higher odds. Sex does not matter (OR 0.97, an interval that contains 1, p = 0.95). Note the rule that will recur in Chapter 10: for ratios (OR, RR), the "no effect" value is 1, not 0, and an interval that contains 1 corresponds to a nonsignificant result.

In Chapter 10 we will calculate by hand, from the 2 x 2 table, the crude OR between obesity and diabetes: 4.32 (95% CI 1.86-10.03). The OR adjusted for age and sex (6.01) is larger than the crude one. The difference does not indicate an error: adjusting for a factor strongly related to the outcome (age) can change the OR, even if that factor is weakly related to the exposure. For now it is enough to remember that the crude OR and the adjusted OR answer different questions and that a good article reports both.

9.9.3 The Same Question, Two Models

The prospective study of 921 students cited above shows how the choice between linear and logistic regression depends on how we treat the outcome variable [5]. The change in BMI, as a continuous variable, was analyzed by linear regression $\left(\beta = +\frac{0.22\ \text{kg}}{m^2}\right)$. The same change, dichotomized as "clinically significant weight gain" (present in 7.2% of the students), was analyzed by logistic regression. The first approach uses all the information; the second answers a more direct clinical question (who reaches a relevant gain?), but loses information by cutting the continuous variable. A cross-sectional study of 2976 students in Timiș County used logistic regression to identify factors associated with heavy smoking [22]; the predictors identified in this way are associated factors, not proven causes, a discussion we continue in Chapter 10.

9.10 Worked Problems

Worked Problem 9.1. Pearson and Regression, by Hand, on Eight Patients

Statement. For the eight patients in Table 9.1 (BMI and SBP): (a) calculate the Pearson coefficient and interpret it according to Colton; (b) test its significance at α = 0.05 and calculate the 95% CI for ρ; (c) calculate R^2; (d) estimate the regression line of SBP on BMI; (e) calculate the SEE and the 95% CI for the slope; (f) predict SBP for a BMI of 30 and comment on a prediction for a BMI of 45.

Solution. (a) From Table 9.1: Sxy = 302, Sxx = 146, Syy = 684. $r = \frac{302}{\sqrt{146 \times 684}} = 0.956$. According to Colton, |r| ≥ 0.75: strong direct correlation.

(b) Hypotheses: H0: ρ = 0; H1: ρ ≠ 0. $t = \frac{0.956\times\sqrt{6}}{\sqrt{1-0.9133}} = \frac{2.342}{0.2945} = 7.95$; df = 6; critical t = 2.447. Since 7.95 > 2.447, H0 is rejected (p = 0.0002). 95% CI by the Fisher transformation: $z = 0.5 \times \ln\left(\frac{1.956}{0.044}\right) = 1.893$; SE $= \frac{1}{\sqrt{5}} = 0.447$; $z \pm 0.877$ gives 1.016 and 2.770; transformed back: 0.769-0.992.

(c) $R^2 = 0.956^2 = 0.913$. About 91% of the variation in SBP in this small group is linearly associated with BMI.

(d) $b = \frac{302}{146} = 2.068$; a = 131 - 2.068 x 29 = 71.01. Equation: SBP = 71.01 + 2.068 x BMI.

(e) The predicted values are 118.6, 122.7, 124.8, 128.9, 131.0, 135.1, 139.3 and 147.5, and the residuals -0.6, 3.3, -3.8, 0.1, 4.0, -4.1, 0.7 and 0.5. The sum of the residuals is zero (a useful check). SSE = 59.3; SEE $= \sqrt{\frac{59.3}{6}} = 3.14$ mmHg. SE$(b) = \frac{3.14}{\sqrt{146}} = 0.260$; CI $95\% = 2.068 \pm 2.447 \times 0.260 = 1.43 - 2.71$.

(f) SBP$(30) = 71.01 + 2.068 \times 30 = 133.1$ mmHg. For a BMI of 45 the formula gives 164.1 mmHg, but the value 45 lies outside the observed range (23-37 kg/m²): it is an extrapolation, and with only eight patients we have no grounds to believe that the relationship remains linear there.

Clinical interpretation: in this group, patients with a higher BMI have higher blood pressure, by about 2 mmHg per kg/m². The result is significant, but the sample is very small, and the intervals for ρ (0.77-0.99) and for the slope (1.4-2.7) are wide.

Worked Problem 9.2. Spearman with Tied Values

Statement. Eight patients rated their physical activity on a scale from 1 (sedentary) to 5 (very active) and had their BMI measured: activity 1, 2, 2, 3, 3, 3, 4, 5; BMI 33.5, 31.0, 29.4, 30.2, 27.8, 28.6, 26.1, 24.9 kg/m². Calculate the Spearman coefficient and compare the result of the short formula with the exact one.

Solution. Activity is ordinal, so we choose Spearman. Ranks of activity: the value 1 receives rank 1; the two values of 2 would occupy positions 2 and 3, so they receive 2.5 each; the three values of 3 would occupy positions 4, 5 and 6, so they receive 5 each; 4 receives 7 and 5 receives 8. Ranks of BMI (in ascending order): 24.9 receives 1; 26.1 receives 2; 27.8 receives 3; 28.6 receives 4; 29.4 receives 5; 30.2 receives 6; 31.0 receives 7; 33.5 receives 8. By patient, in the order of the statement: the activity ranks are 1, 2.5, 2.5, 5, 5, 5, 7, 8, and the BMI ranks are 8, 7, 5, 6, 3, 4, 2, 1.

The differences di are -7, -4.5, -2.5, -1, 2, 1, 5, 7, and $\sum {d_i}^2 = 49 + 20.25 + 6.25 + 1 + 4 + 1 + 25 + 49 = 155.5$. With the short formula: $r_s = 1 - \frac{6\times155.5}{8\times63} = 1 - \frac{933}{504} = 1 - 1.851 = -0.851$.

The exact calculation (Pearson on ranks, as any software does) gives rs = -0.908, p = 0.002. The difference (-0.85 versus -0.91) arises because of the five tied activity values, which reduce the variation of the ranks; the short formula assumes that the ranks are 1, 2,..., n without repetitions. Conclusion: a strong inverse correlation in this small group; the more active patients have a lower BMI. In mg2026, the same analysis on 160 patients gave rs = -0.012, practically zero. The eight-patient example was constructed for calculation, and the contrast reminds us that a small sample can produce spectacular correlations.

Worked Problem 9.3. Height and Weight, in the Format of the Discipline's Assignments

Statement. On the mg2026 dataset, investigate the association between height and weight. Answer in the a-g format and specify: what proportion of the variation in weight is due to factors other than height? What weight does the model predict for a person 1.78 m tall and for one 2.35 m tall?

(a) Research hypothesis: taller people have, on average, a higher weight.

(b) Statistical hypotheses: H0: ρ = 0, height and weight are not linearly correlated in the population; H1: ρ ≠ 0, the two variables are correlated.

(c) Method: two numeric variables measured on the same patients, hence a bivariate correlation and regression analysis. Height is approximately normal (Shapiro-Wilk p = 0.078); weight is skewed (skewness 1.29; p < 0.001), so we report Pearson and, as a check, Spearman. The predictor X is height, the response Y is weight.

(d) Solution: JASP (Regression > Correlation and Regression > Linear Regression) or Google Sheets (=CORREL, =SLOPE, =INTERCEPT, =RSQ).

(e) Results: r = 0.652 (95% CI 0.552-0.733), t = 10.79, df 158, p < 0.001; Spearman rs = 0.693, p < 0.001. The line: weight = -100.40 + 1.087 x height (cm); 95% CI for the slope 0.888-1.286 kg/cm; $R^2 = 0.424$; SEE = 11.49 kg. The graph is Figure 9.6.

(f) Statistical conclusion: p < 0.001, H0 is rejected; the correlation is positive and moderate according to Colton (0.50 ≤ r < 0.75).

(g) Research conclusion: taller people weigh more, on average by 1.09 kg for each centimeter. $R^2 = 0.424$, so 42.4% of the variation in weight is associated with height, and 1 - 0.424 = 0.576, that is, 57.6%, is due to other factors (sex, body composition, diet, activity). For 178 cm, the predicted weight is -100.40 + 1.087 x 178 = 93.1 kg (95% prediction interval for one person: 70.2-115.9 kg). For 235 cm, the formula gives 155.0 kg, but this is an extrapolation: the greatest height in the sample is 192 cm. The correct answer to the second question is that the model cannot make a credible prediction.

Worked Problem 9.4. Checking Published Correlations

Statement. (a) A study of 867 patients reports r = -0.088 and p = 0.01 [55]. Check the p value. (b) A study of 121 students reports R = 0.107, p = 0.246 [16]. Check it. (c) A study of 157 patients reports ρ = 0.826 with a 95% CI of 0.769-0.870 [107]. Check the interval using the Fisher transformation.

Solution. $(a)\ t = \frac{0.088\times\sqrt{865}}{\sqrt{1-0.088^2}} = \frac{2.588}{0.996} = 2.60$ (in absolute value); with 865 degrees of freedom, p = 0.0095, that is, approximately 0.01. The result is confirmed; the correlation is significant but negligible ($R^2 = 0.8\%$).

(b) $t = \frac{0.107\times\sqrt{119}}{\sqrt{1-0.0114}} = \frac{1.167}{0.994} = 1.17$; with 119 degrees of freedom, p ≈ 0.24, very close to the reported 0.246 (the difference comes from the rounding of R). H0 is not rejected.

(c) $z = 0.5 \times \ln\left(\frac{1.826}{0.174}\right) = 0.5 \times \ln(10.494) = 1.175$. SE $= \frac{1}{\sqrt{157-3}} = \frac{1}{\sqrt{154}} = 0.0806$. $z \pm 1.96 \times 0.0806 = 1.175 \pm 0.158$, that is, 1.017 and 1.333. We transform back with $r = \frac{e^{2z}-1}{e^{2z}+1}$. For z = 1.017: $e^{2.034} = 7.64$, so $r = \frac{6.64}{8.64} = 0.769$. For z = 1.333: $e^{2.666} = 14.38$, so $r = \frac{13.38}{15.38} = 0.870$. The published interval is reproduced exactly. The lesson: when an article gives r and n, you can reconstruct the confidence interval yourselves.

Worked Problem 9.5. The Regression Line from Summary Statistics

Statement. An article reports for waist circumference (X) and SBP (Y): mx = 88.5 cm, sx = 12.1 cm, my = 131.1 mmHg, sy = 13.9 mmHg, r = 0.32, n = 160. (a) Write the regression line. (b) Predict SBP for a waist of 100 cm. (c) What proportion of the variation in SBP is associated with waist circumference? (d) Is the correlation significant?

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. SBP = 98.6 + 0.368 x waist.

(b) $\text{SBP}(100) = 98.6 + 0.368 \times 100 = 135.3$ mmHg. The value of 100 cm lies within the usual range of waist circumference in overweight adults, so the prediction is an interpolation.

(c) $R^2 = 0.32^2 = 0.102$; about 10% of the variation in SBP is linearly associated with waist circumference, and 90% is due to other factors.

(d) $t = \frac{0.32 \times \sqrt{158}}{\sqrt{1 - 0.1024}} = \frac{4.022}{0.947} = 4.25$; df 158; p < 0.001. The correlation is significant and weak according to Colton. In mg2026, the direct calculation gives r = 0.322 and SBP = 98.49 + 0.369 x waist, very close to the result obtained from the rounded statistics.

Worked Problem 9.6. Reading a Multiple Regression Table

Statement. Using Table 9.6, answer: (a) What is the mean difference in SBP between two patients of the same sex, with the same smoking status and the same BMI, one of whom is 50 years old and the other 70? (b) Is sex a significant predictor? (c) Why does the BMI coefficient (1.389) differ slightly from the slope in the simple regression (1.435)? (d) Interpret the adjusted R^2.

Solution. (a) 20 years $\frac{x\ 0.449\ \text{mmHg}}{\text{year}} = 9.0$ mmHg; the 70-year-old patient has, on average, an SBP about 9 mmHg higher. The interval for 20 years is obtained by multiplying the limits: 20 x 0.235 = 4.7 and 20 x 0.662 = 13.2 mmHg.

(b) No: the coefficient -1.481 has a 95% CI from -5.19 to 2.23, which contains 0, and p = 0.43. We have no evidence that men and women differ in SBP after adjustment for the other factors.

(c) In simple regression, the BMI slope also absorbs the part of the association that overlaps with other factors correlated with BMI (age, smoking). After adjustment, the coefficient refers to patients comparable in age, sex and smoking. The very small change (from 1.435 to 1.389) shows that these factors do not seriously confound the BMI-SBP relationship in mg2026.

(d) R^2 adjusted $= 0.286$: after the penalty for the 5 predictors, the model explains about 29% of the variation in SBP. The small difference from R^2 (0.308) shows that there is no excess of useless predictors.

9.11 Proposed Exercises

1. In six patients, age (years) and SBP (mmHg) were measured: (35, 118), (42, 121), (50, 130), (58, 128), (63, 140), (70, 145). Calculate r by hand, test it at α = 0.05, write the regression line of SBP on age and predict SBP at 60 years.
2. A study reports r = 0.30 between daily salt intake and SBP. Test the significance for n = 40 and for n = 200. What general conclusion do you draw about the relationship between n and significance? How do you classify the correlation according to Colton in both cases?

3. For r = 0.62 obtained on 50 patients, calculate the 95% CI for ρ using the Fisher transformation. Could the correlation in the population be "strong" according to Colton?
4. Seven patients have pain scores on the visual analog scale (VAS) of 2, 5, 3, 8, 6, 9, 4 and anxiety scores of 10, 14, 9, 20, 18, 17, 12. Calculate the Spearman coefficient by hand and interpret it.
5. In mg2026, fasting plasma glucose and HbA1c have r = 0.738. Calculate R^2 and 1 - R^2 and write one sentence for each, in the style of the discipline's assignments. Classify the correlation according to Colton.
6. Classify the following coefficients according to Colton's rule and state their direction: 0.81; -0.573; 0.256; -0.088; 0.729; -0.832; 0.107; 0.50.
7. The differences between an automatic blood pressure monitor and the reference device, in eight patients, are (mmHg): 4, -2, 6, 3, 8, 1, 5, -1. Calculate the mean bias and the Bland-Altman limits of agreement. Why would the correlation coefficient between the two devices not have answered the question "can they be used interchangeably?"
8. A logistic model for diabetes gives OR = 1.11 per year for age. Calculate the approximate OR for a difference of 5 years and of 10 years. Is it correct to say that diabetes is "1.11 times more probable" with each additional year? Explain the difference between odds and probability.
9. Using Table 9.6, calculate the predicted SBP for a 45-year-old female non-smoker with a BMI of 27 and for a 65-year-old male smoker with a BMI of 35. Which variable contributes most to the difference between them?
10. A regression of weight (kg) on height (cm) in 50 students gives weight = -95 + 0.95 x height, with $R^2 = 0.38$, the heights ranging from 155 to 195 cm. (a) Interpret the slope. (b) What proportion of the variation in weight is due to other factors? (c) Predict the weight for 1.70 m and discuss the prediction for 1.20 m.

Common Pitfalls

- Reporting a correlation coefficient without the scatter plot; a single extreme value can create or hide a correlation (Figure 9.3).
- Confusing significance with strength: r = -0.088 with p = 0.01 remains a negligible correlation.
- Interpreting correlation as evidence of causation, ignoring confounders, reverse causation and the ecological fallacy.
- Using r to demonstrate the agreement of two measurement methods, instead of the Bland-Altman method.
- Reading R^2 as the proportion "caused" by X; R^2 is the proportion of the variation linearly associated with X.
- Interpreting the intercept clinically when X = 0 makes no sense (zero BMI, zero height).
- Extrapolating the regression line outside the range of the data (weight at 2.35 m).
- Checking normality on the variable X instead of on the residuals; in regression, the assumption concerns the residuals.
- Using the short Spearman formula when there are many tied values.
- Writing "H0 is accepted" when p > 0.05; the correct wording is "H0 is not rejected".

Key Points

- Correlation measures the direction and strength of a linear (Pearson) or monotonic (Spearman) association; regression describes how Y changes on average with X and allows prediction.

- $r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}}$, between -1 and 1; the test uses $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ with n - 2 degrees of freedom; the CI is calculated by the Fisher transformation.
- Colton's rule: |r| < 0.25 no correlation; 0.25-0.50 weak; 0.50-0.75 moderate; ≥ 0.75 strong.
- R^2 = the proportion of the variation in Y explained by the model; 1 - R^2 = the proportion due to other factors.
- $b = \frac{S_{xy}}{S_{xx}} = \frac{r \times s_y}{s_x}$; a = my - b x mx; the line passes through the point of the means.
- The SEE (in the units of Y) shows the precision of an individual prediction; the prediction interval for a patient is much wider than the confidence interval for the mean.
- The assumptions of regression (LINE) are checked on the residuals: linearity, independence, normality, constant variance.
- In multiple regression, each coefficient is adjusted for the other predictors; in logistic regression, e^b is an OR, and the no-effect value is 1.
- In mg2026: BMI-SBP r = 0.405, rs = 0.372, $R^2 = 0.164$, SBP = 89.47 + 1.435 x BMI (CI for the slope 0.925-1.944), SEE 12.73 mmHg.

Self-Assessment Questions

1. (SC) For two normally distributed numeric variables measured on the same 50 patients, the most appropriate measure of the strength of the linear association is:
    a) the independent-samples t-test
    b) the Pearson correlation coefficient
    c) the chi-square test
    d) the kappa coefficient
    e) the odds ratio
2. (SC) A correlation coefficient of r = -0.62 indicates, according to Colton's rule:
    a) a strong inverse correlation
    b) no correlation, because r is negative
    c) a moderate inverse correlation
    d) a moderate direct correlation
    e) a weak inverse correlation
3. (SC) Between height and weight, r = 0.60 was obtained. The proportion of the variation in weight due to factors other than height is:
    a) 40%
    b) 36%
    c) 60%
    d) 64%
    e) 16%
4. (SC) The regression line SBP = 89.47 + 1.435 x BMI was obtained on patients with a BMI between 23.5 and 48.3. The correct statement is:
    a) a patient with a BMI of 0 has an SBP of 89.47 mmHg
    b) the predicted SBP for a BMI of 30 is 132.5 mmHg and represents the estimated mean for patients with a BMI of 30
    c) the slope shows that BMI causes SBP to increase by 1.435 mmHg in every patient
    d) the prediction for a BMI of 70 is as credible as the one for a BMI of 30

e) r is negative, because the intercept is positive

5. (SC) To check whether a portable glucometer can replace the laboratory determination, the appropriate method is:
   a) the Pearson coefficient between the two determinations
   b) logistic regression
   c) the Bland-Altman method, with the mean bias and the limits of agreement
   d) the Spearman coefficient
   e) the chi-square test of independence
6. (MC) The following statements about the Pearson correlation coefficient are true:
   a) it has no unit of measurement
   b) it is sensitive to extreme values
   c) a value close to zero excludes any relationship between the variables
   d) it changes if weight is expressed in grams instead of kilograms
   e) it has the same sign as the slope of the regression line
7. (MC) The Spearman correlation coefficient is preferred to the Pearson coefficient when:
   a) one of the variables is ordinal (for example, physical activity on the 1-5 scale)
   b) the distributions are highly skewed or there are extreme values
   c) both variables are normal and the relationship is linear
   d) the relationship is monotonic but nonlinear
   e) the variables are nominal with three categories
8. (MC) In a sample of 867 patients, r = -0.088, p = 0.01 was obtained between age and the duration of surgery. The following are correct:
   a) the correlation is statistically significant
   b) according to Colton, the correlation is negligible
   c) age explains about 8.8% of the variation in duration
   d) R^2 is below 1%
   e) the result proves that advanced age shortens the procedure
9. (MC) The assumptions of simple linear regression include:
   a) a linear relationship between X and Y
   b) independence of the observations
   c) normal distribution of the residuals
   d) normal distribution of the variable X
   e) constant variance of the residuals over the whole range of X
10. (MC) In a multiple regression of SBP, the coefficient for "smoker, versus former smoker" is 6.738 (95% CI 1.53-11.95; p = 0.012), and $R^2 = 0.308$, R^2 adjusted $= 0.286$. The following are correct:

a) at the same age, the same BMI and the same sex, smokers have an SBP on average about 6.7 mmHg higher than former smokers

b) the model explains about 31% of the variation in SBP

c) adjusted R^2 is smaller than R^2 because it penalizes the number of predictors

d) smoking causes an increase in SBP of exactly 6.738 mmHg in every patient

e) a confidence interval that does not contain 0 corresponds to a coefficient that is significant at $\alpha = 0.05$

Chapter 10. Statistical Applications in Epidemiology: Types of Observational Studies and Risk Analysis

Objectives

By the end of this chapter, the student will be able to:

- define the risk factor, classify it and describe the types of exposure (one-time, chronic, dose, duration);
- calculate and interpret prevalence, cumulative incidence and the incidence rate with person-time;
- recognize from the problem statement the type of study (experimental, cross-sectional, prospective or retrospective cohort, case-control, ecological) and choose the appropriate measure of association;
- construct the 2 x 2 contingency table with the notation a, b, c, d and calculate by hand the relative risk (RR) and the odds ratio (OR);
- calculate the 95% confidence intervals for the RR and OR by the logarithmic method and use them to decide on significance;
- explain why the OR exceeds the RR when the disease is common and why the two are close for rare diseases;
- calculate the attributable risk, the attributable fraction and the number needed to treat (NNT) or to harm (NNH);
- recognize a confounder, explain the idea of stratification (Mantel-Haenszel) and distinguish confounding from effect modification;
- apply the Bradford Hill criteria and the STROBE guideline in the critical reading of an observational study.

10.1 The Risk Factor

10.1.1 What We Call a Risk Factor

In the previous chapters we compared means and proportions between groups. Epidemiology puts the same tools at the service of a precise question: what increases the probability that a healthy person will develop a particular disease? A risk factor is a characteristic, a behavior, an exposure or a condition associated with a higher probability of the disease occurring. The key word is "probability": smoking does not guarantee COPD, and many patients with COPD have never smoked, but among smokers COPD occurs more often. The risk factor is a hypothetical cause, which the study tries to confirm or refute.

In family medicine we work with risk factors every day: when we calculate the SCORE2 score for cardiovascular risk, when we administer the FINDRISC questionnaire for diabetes or when we ask a patient how many cigarettes they smoke per day. Each of these tools synthesizes the results of dozens of epidemiological studies of the kind described in this chapter [64,111].

10.1.2 Classification of Risk Factors

The lectures of our discipline group risk factors by their nature. Environmental factors include air pollutants, toxins, radiation, infectious microorganisms and occupational exposures (dusts, solvents, noise). Behavioral factors relate to the person's habits: smoking, alcohol or drug use, diet, sedentary lifestyle, failure to follow occupational safety measures. Social factors include poverty, educational level, isolation and traumatic life events (the death of a family member,

divorce). Genetic factors are inherited: familial hypercholesterolemia, BRCA mutations in breast and ovarian cancer, a family history of diabetes.

A second classification, clinically useful, separates modifiable factors (smoking, obesity, hypertension, sedentary lifestyle) from non-modifiable ones (age, sex, genetics). The former are the target of prevention; the latter help identify the people who need closer follow-up. In the mg2026 dataset, obesity and smoking are modifiable, while age and sex are not.

10.1.3 Exposure to the Risk Factor

Exposure can be one-time (a single event, limited in time: an accident, an accidental needlestick injury with a contaminated needle, food poisoning at a wedding) or chronic (prolonged, the most common type in noncommunicable diseases). For chronic exposure, three dimensions are described: the current dose (how many cigarettes per day the patient smokes now), the cumulative dose (for example, the number of pack-years, that is, packs per day multiplied by the years of smoking) and the duration of exposure (for how many years). A patient who has smoked one pack a day for 30 years has 30 pack-years; one who smoked two packs a day for 15 years also has 30 pack-years. Measuring exposure correctly is one of the greatest difficulties of observational studies, because it often relies on the patient's memory.

10.1.4 Causal Factor and Risk Marker

Not every factor associated with the disease also causes it. A causal factor actually takes part in the mechanism of the disease: eliminating it reduces incidence (quitting smoking reduces the risk of lung cancer). A risk marker (predisposing factor, indicator) is associated with the disease and helps us identify people at high risk, but it is not necessarily part of the causal chain. Fingers stained yellow by nicotine are associated with lung cancer, but nobody would set out to prevent cancer by washing their hands. Elevated C-reactive protein is associated with coronary events, but lowering it artificially does not guarantee a lower risk. The difference has practical consequences: a marker is good for screening, a causal factor is good for prevention as well.

10.2 Measuring Disease Frequency

Before comparing exposed and unexposed people, we need to know how to measure how frequent a disease is. There are two families of measures: prevalence, which describes how many people are ill, and incidence, which describes how many become ill.

10.2.1 Prevalence

Prevalence is the proportion of people in a population who have the disease at a given time or during a given period:

Prevalence = number of people with the disease / total number of people in the population

Point prevalence refers to a moment (for example, the day of the examination), period prevalence to an interval (for example, the proportion of people who have had at least one episode of low back pain in the past year). Prevalence includes both new and existing cases. In the mg2026 dataset, at baseline, 29 of the 160 patients had diabetes confirmed by OGTT: the prevalence is 29 / 160 = 18.1%, with a 95% Wilson CI of 12.9-24.8% (Chapter 4). Prevalence is the measure a health care manager needs: how many beds, how many consultations, how many doses of medication are needed now.

10.2.2 Cumulative Incidence (Risk)

Cumulative incidence is the proportion of people initially free of the disease who develop the disease during a defined period:

Cumulative incidence = number of new cases in the period / number of people free of the disease at the beginning of the period

It is equivalent to the probability (risk) of developing the disease during that period. The period must always be specified: a risk of 8% "over 5 years" is something entirely different from 8% "per year". If, of 1000 healthy people followed for 5 years, 80 develop diabetes, the 5-year cumulative incidence is 80 / 1000 = 8%. Cumulative incidence requires all participants to be followed for the whole period; for those lost to follow-up we do not know whether they became ill or not.

10.2.3 The Incidence Rate and Person-Time

In real studies, participants enter the study at different times, some are lost to follow-up, others die of other causes. The solution is to count the actual observation time. A person followed for 3 years contributes 3 person-years; a person followed for 6 months contributes 0.5 person-years. The incidence rate (incidence density) is:

Incidence rate = number of new cases / sum of person-time of observation

Example: 200 patients with type 2 diabetes are followed for the onset of retinopathy. Because of staggered entry and losses, the total observation time adds up to 850 person-years, and during this time 34 new cases occur. The incidence rate is 34 / 850 = 0.040 cases per person-year, that is, 4.0 cases per 100 person-years. The rate is not a proportion: its unit is "per time", and its value can theoretically exceed 1 (a very common disease, with repeated episodes). The ratio of two incidence rates is called the incidence rate ratio (IRR) and is interpreted in the same way as the relative risk.

10.2.4 The Link between Prevalence and Incidence

For a disease in a steady state (approximately constant incidence and duration) and with low prevalence, there is a simple relationship:

$$P \approx I \times D$$

where P is the prevalence, I the incidence rate and D the mean duration of the disease. An incurable chronic disease with long survival (diabetes, hypertension) has a high prevalence even if the incidence is moderate: an incidence of 2 cases per 1000 person-years and a mean duration of 10 years give a prevalence of about 0.002 x 10 = 0.02, that is, 2%. A short acute disease (influenza, a week-long respiratory infection) can have a very high incidence and a low point prevalence. The relationship also explains an apparent paradox: a new treatment that prolongs survival without curing increases the prevalence of the disease, even though it is an advance. This is why comparing prevalences between regions or periods calls for caution.

10.3 Types of Epidemiological Studies

10.3.1 Experimental Studies and Observational Studies

The fundamental difference between types of studies is who decides the exposure. In an experimental (intervention) study, the researcher allocates the exposure, ideally by randomization: the researcher decides who receives the new treatment and who receives placebo or the standard treatment. Randomization distributes all known and unknown factors equally between the groups, on average, so the difference in outcome can be attributed to the intervention. Randomized clinical trials provide the strongest causal evidence, but they cannot be used for harmful risk factors: nobody can randomize people to smoke or to work in dust. The mg2026 dataset is such a simulated experimental study (arms A, B, C), and Chapter 13 discusses it in detail.

In an observational study, the exposure occurred independently of the researcher (the patient chose to smoke, worked in a factory, was born with a particular gene). The researcher only observes and measures. The main disadvantages are the limited precision of exposure measurement and the fact that exposed and unexposed people may also differ in other factors. Observational studies are classified according to when exposure and disease are measured. Figure 10.1 shows the three main designs schematically.

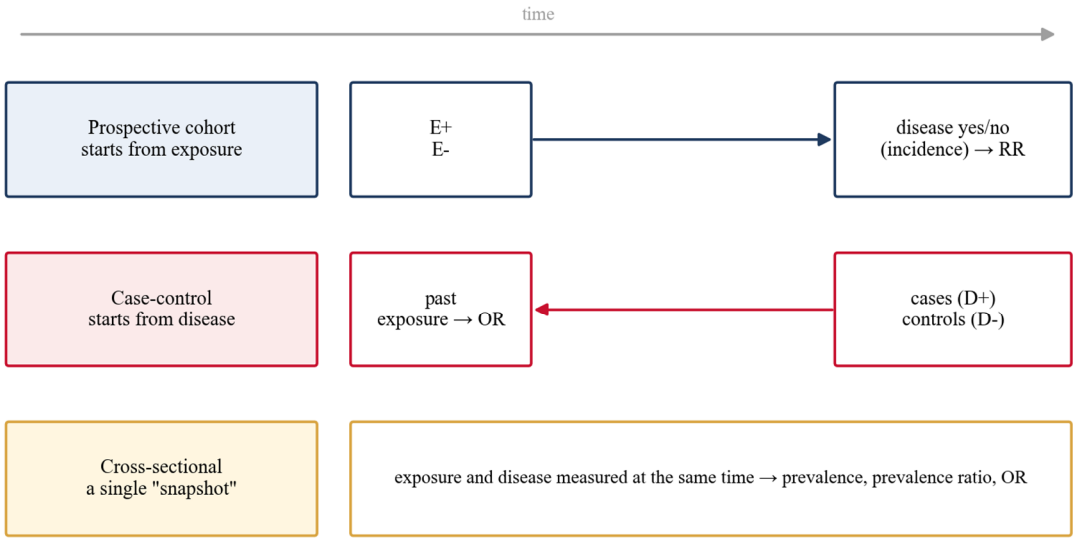


Figure 10.1. The three main observational designs. The cohort study starts from the exposure (E+ and E-) and follows the onset of disease over time, allowing incidence and relative risk to be calculated. The case-control study starts from the disease (cases and controls) and looks back for the exposure, allowing only the odds ratio to be calculated. The cross-sectional study measures exposure and disease at the same time and estimates the prevalence, the prevalence ratio and the OR.

10.3.2 The Cross-Sectional Study

The cross-sectional study is a "snapshot" of the population at a given moment: for each participant, the presence or absence of exposure and of disease are recorded at the same time. It is a prevalence study and the simplest research design. The baseline analysis in mg2026, in which we related obesity to diabetes confirmed by OGTT, is cross-sectional: both were measured at the first visit.

The natural measure of association is the prevalence ratio (PR): the prevalence of the disease in the exposed divided by the prevalence in the unexposed. The formula is identical to that of the relative risk, but the name differs, because we are not measuring new cases but existing ones. The OR is also frequently reported, especially when logistic regression is used.

The advantages are low cost and speed, the possibility of studying several factors and several diseases simultaneously, and usefulness for health services planning. The major disadvantage is that we cannot establish the temporal sequence: if obese people more often have diabetes, we do not know, from this study alone, whether obesity preceded diabetes. A cross-sectional survey of 2976 students in Timiș County, 62.5% of them female, identified factors associated with heavy smoking by logistic regression [22]; the design allows us to describe the associations, but not to establish which came first, the smoking or the associated behaviors. A second disadvantage has to do with selective survival: diseases with a rapid course (death or recovery) are underrepresented, because the patients concerned are no longer present at the moment of the "snapshot". The prevalence estimate is distorted for diseases with a rapid course.

10.3.3 The Cohort Study

In a cohort study, participants are selected initially free of the disease, classified by exposure (exposed E+ and unexposed E-) and followed over time for the onset of the disease [112]. The term comes from the cohorts of the Roman legions: a group that moves forward together. Since we start from healthy people and count new cases, we can directly calculate the incidence in the exposed and the unexposed, and therefore the relative risk.

A prospective cohort starts now and follows participants into the future (for example, 921 students followed during their first year at university [5]). A retrospective (historical) cohort uses past records: the researcher establishes from documents who was exposed 10 years ago and checks who has become ill since then. Both keep the "from exposure to disease" logic; they differ only in the moment when the researcher enters the scene. The lectures of our discipline, following an older tradition, state that only the OR is calculated in a retrospective cohort; in reality, if the records allow complete identification of the exposed and unexposed and of the new cases, the RR can also be calculated in a retrospective cohort. What matters is whether the proportions of ill people in each exposure group reflect the real risk.

The advantages of the cohort study are a clear temporal sequence (exposure precedes disease), direct calculation of incidence and RR, the possibility of studying several diseases caused by the same exposure, and protection against recall errors, because exposure is recorded before the disease occurs. The disadvantages are cost, duration (diseases with a long latency, such as cancer, require decades), losses to follow-up and inefficiency for rare diseases: to observe a few cases of a rare disease, tens of thousands of people would have to be followed.

10.3.4 The Case-Control Study

The case-control study reverses the logic: it starts from the disease [113]. The researcher selects a group of people with the disease (the cases, B+) and a group of people without the disease (the controls, B-), then looks into each person's past for exposure to the risk factor, through interviews, questionnaires or medical records.

Because the researcher decides how many cases and how many controls to include (for example, one case to two controls), the proportion of ill people in the sample does not reflect the risk in the population. Therefore, in a case-control study neither the incidence nor the relative risk can be calculated. Only the odds ratio is calculated, which has a remarkable property (Section 10.6): the OR of exposure in cases versus controls is equal to the OR of disease in the exposed versus the unexposed.

The choice of controls is the most delicate point. Controls must come from the same population as the cases: if a control had developed the disease, they would have been included in the study as a case. For cases from a hospital, controls are usually chosen from the same hospital, with other diseases unrelated to the exposure, or from the population served by the hospital. Controls chosen from a different population (for example, healthy students for elderly hospitalized cases) produce selection bias.

Matching is the choice of controls so that they resemble the cases in variables that would otherwise confound the result (age, sex). Matching can be individual (for each case, one or more controls of the same age and sex are found) or frequency matching (the distribution by age and sex is the same in the two groups). Matching eliminates confounding by the matched variables, but those variables can no longer be studied as risk factors, and the analysis must take the pairs into account.

Recall bias is specific to case-control studies: ill people, who have searched for explanations for their disease, remember exposures better or overestimate them, while healthy people minimize

them. A mother of a child with a congenital malformation will remember in detail every medication taken during pregnancy; a mother of a healthy child may forget. The result is an artificially inflated OR.

The advantages of the case-control study are low cost, speed, efficiency for rare diseases and for those with a long latency, and the possibility of studying several risk factors for the same disease simultaneously. The disadvantages are the impossibility of calculating incidence and RR, selection and recall bias, and the difficulty of establishing with certainty that the exposure preceded the disease.

An important variant is the nested case-control study, in which cases and controls are drawn from an already defined cohort. The combination keeps the advantages of the cohort (exposure was documented prospectively) and the efficiency of the case-control study.

10.3.5 The Ecological Study

In an ecological study, the unit of analysis is not the person but the group: the country, the county, the city, the month. The frequency of exposure is compared with the frequency of disease at the group level, for example mean salt intake per county with stroke mortality per county. Ecological studies are inexpensive, use already existing data and are useful for generating hypotheses and for evaluating public health interventions. Their major limitation is the ecological fallacy, discussed in Chapter 9: an association at the group level does not automatically transfer to the individual level, and numerous external variables can act as confounders. Interrupted time series analysis, in which the trend of a rate before and after an event is compared, is a modern form of ecological study.

10.3.6 Comparison of Designs

Table 10.1. Comparison of the main types of epidemiological studies.

Characteristic	Cross-sectional	Cohort	Case-control	Ecological	Experimental
Starting point	the population, at one moment	the exposure	the disease	groups	allocation by the researcher
Direction	simultaneous	forward (toward the future)	backward (toward the past)	usually simultaneous	forward
Measure of frequency	prevalence	incidence	cannot be calculated	group rates	incidence
Measure of association	PR, OR	RR, OR, IRR, HR	OR only	correlation, IRR	RR, risk difference, NNT
Temporal sequence	no	yes	uncertain	no	yes
Rare diseases	unsuitable	unsuitable	very suitable	possible	unsuitable
Cost and duration	low	high	low	very low	very high
Typical biases	selective survival	losses to follow-up	selection, recall	ecological fallacy	protocol deviations

The table should not be memorized as a list; it follows from the question "where does the study start?". If it starts from the exposure and counts new cases, we can calculate the risk. If it starts from the disease, we cannot.

> **Case study: five designs in our colleagues' articles**
>
> An observational follow-up study of 110 patients with high-grade cervical lesions treated by excision (LEEP), with HPV testing at 6, 12 and 18 months, is a cohort: the patients were classified by age and HPV type and followed for persistence of the infection [114]. The persistence rate fell from 40.9% at 6 months to 20% at 12 months and 11.8% at 18 months, and age, dichotomized at 36.5 years, was associated with persistence: RR = 2.75, 95% CI 1.34-

5.64, p = 0.0027. The interval does not contain 1, so the association is significant; its width (from a 1.3-fold to a 5.6-fold higher risk) reflects the 110 patients.

A cross-sectional study of 101 HIV-positive patients on antiretroviral therapy in western Romania analyzed the association between HIV disease severity and psychiatric disorders using odds ratios: ORs between 2.81 (p = 0.038) and 8.72 (p < 0.001) for different disorders [84]. The cross-sectional design justifies the OR, but it does not allow the direction to be established (does severe HIV disease lead to psychiatric disorders, or do psychiatric disorders lead to poorer adherence and thus to more severe disease?). That the abstract gives ORs with p but without confidence intervals, and that many disorders are tested simultaneously, are two points for critical reading: with seven comparisons, the probability that at least one turns out significant by chance increases.

A study of 656 patients with ST-segment elevation myocardial infarction treated by primary angioplasty compared the 96 patients with the no-reflow phenomenon with the rest of the cohort regarding short-term complications [115]. This is a case-control design built within a cohort. A retrospective study of 435 patients with acute calculous cholecystitis used matching for age and sex, with a 1:2 ratio (145 cases with the gangrenous form, 290 controls with the chronic form), and reported for an inflammatory index OR = 1.976, p < 0.001, in a logistic regression model [116]. Matching for age and sex means that these two variables can no longer be analyzed as risk factors, but neither can they confound the result. Finally, a cross-sectional clinical study with a nested case-control analysis examined 136 patients (52 cases with disorders of gut-brain interaction and 84 asymptomatic controls) and found a protective OR of 0.36 (95% CI 0.18-0.74; p = 0.009) [71]. An OR below 1, with an interval that does not contain 1, means that the exposure in question was rarer in cases than in controls.

10.4 The 2 x 2 Contingency Table

All the calculations in risk analysis start from the same table. The convention used in our discipline (and in most textbooks) puts exposure in the rows and disease in the columns, with the exposed in the first row and the ill in the first column.

Table 10.2. Notation of the 2 x 2 contingency table.

	Disease present (B+)	**Disease absent (B-)**	**Total**
Exposed (E+)	a	b	a + b
Unexposed (E-)	c	d	c + d
Total	a + c	b + d	$N = a + b + c + d$

Cell a contains the exposed ill people, b the exposed healthy people, c the unexposed ill people, d the unexposed healthy people. The row margins (a + b and c + d) are the numbers of exposed and unexposed people; the column margins (a + c and b + d) are the numbers of ill and healthy people. In the exam, the most frequent error in risk analysis is not the wrong formula but the wrongly constructed table: rows and columns swapped, or a total put in place of a cell. Always check that a + b + c + d = N and that every total in the problem statement can be found in the table.

In a cohort or experimental study, the row totals are fixed by the design (so many exposed, so many unexposed). In a case-control study, the column totals are fixed (so many cases, so many controls). In a cross-sectional study, only N is fixed. This difference determines which proportions make sense: in a case-control study, a / (a + b) is not the risk of disease in the exposed, because the number of ill people was chosen by the researcher.

10.5 The Relative Risk

10.5.1 The Formula and Its Interpretation

The risk of disease in the exposed is $R_1 = \frac{a}{a+b}$, and in the unexposed $R_0 = \frac{c}{c+d}$. The relative risk (RR, also called the risk ratio) is their ratio:

$$\text{RR} = \frac{R_1}{R_0} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

The RR shows how many times greater the risk of disease is in the exposed than in the unexposed. There are three situations for interpretation. RR > 1: the exposure is a risk factor (the exposed become ill more often). RR = 1: the exposure is not associated with the disease (an indifferent factor). RR < 1: the exposure is a protective factor (the exposed become ill less often); for example, RR = 0.6 means a 40% lower risk in the exposed. The last interpretation is important for treatments and vaccines, where the "exposure" is the intervention.

Of course, the final decision is not made on the point estimate alone, but on the confidence interval: an RR of 1.8 with a 95% CI of 0.9-3.6 does not show a significant risk factor, because the interval contains 1.

10.5.2 The Classic Example: Smoking and COPD

A study follows 1200 initially healthy people for 5 years, of whom 300 are smokers. At the end, 50 of the smokers and 70 of the non-smokers have COPD. The first step is the table. Smokers: a = 50 ill, b = 300 - 50 = 250 healthy. The non-smokers number 1200 - 300 = 900: c = 70 ill, d = 900 - 70 = 830 healthy.

Table 10.3. Smoking and COPD, a cohort study with 1200 participants.

	COPD yes	COPD no	Total
Smokers	50	250	300
Non-smokers	70	830	900
Total	120	1080	1200

R1 = 50 / 300 = 0.1667 (16.7% over 5 years); $R_0 = \frac{70}{900} = 0.0778 (7.8\%)$; $\text{RR} = \frac{0.1667}{0.0778} =$ 2.14. Smokers have a 2.14 times higher risk of COPD than non-smokers over the 5 years of follow-up.

An error we see year after year: some students divide 70 by 1200 (the total number of participants) instead of 900 (the number of non-smokers) and obtain R0 = 0.0583 and RR = 2.86. The risk in the unexposed is calculated only among the unexposed. An RR of 2.86 makes no sense for this table, and a problem solved incorrectly at the first step leads to all the other results being wrong.

10.6 The Odds Ratio

10.6.1 Odds

The odds of an event are the ratio between the probability that the event occurs and the probability that it does not occur:

$$\text{odds} = \frac{p}{1-p}$$

The term comes from betting: odds of "1 to 4" mean that, for every person who becomes ill, 4 do not, that is, $p = \frac{1}{5} = 20\%$ and odds $= \frac{0.20}{0.80} = 0.25$. For rare events, odds and probability are close (p = 2% gives odds of 0.0204); for common events, they differ greatly (p = 50% gives odds of 1; p = 80% gives odds of 4).

In the 2 x 2 table, the odds of disease in the exposed are a / b (ill to healthy, among the exposed), and in the unexposed c / d.

10.6.2 The Formula and Its Interpretation

The odds ratio (OR) is:

$$\text{OR} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \times d}{b \times c}$$

The second form, "the product of the main diagonal divided by the product of the secondary diagonal", is easy to remember. Interpretation follows the same rules as for the RR: OR > 1 risk factor, OR = 1 indifferent factor, OR < 1 protective factor. For the smoking-COPD example: $\text{OR} = \frac{50 \times 830}{250 \times 70} = \frac{41\,500}{17\,500} = 2.37$. The odds of developing COPD are 2.37 times higher in smokers.

The OR has a symmetry property that makes it indispensable in case-control studies. The odds of exposure in cases are a / c, and in controls b / d. Their ratio is $\frac{\frac{a}{c}}{\frac{b}{d}} = \frac{a \times d}{b \times c}$, exactly the same OR. Therefore, although in a case-control study we cannot calculate the risk of disease, we can calculate the OR of exposure, which is numerically equal to the OR of disease. The correct interpretation in a case-control study is: "the odds of having been exposed are OR times higher in cases than in controls", and, by symmetry, "the odds of disease are OR times higher in the exposed".

10.6.3 Why the OR Is Larger Than the RR When the Disease Is Common

In the smoking-COPD example, OR = 2.37 and RR = 2.14: close, but not equal. The lectures of our discipline note "usually OR > RR". The exact rule is this: when RR > 1, the OR is always farther from 1 than the RR; when RR < 1, the OR is smaller than the RR (again farther from 1). The difference depends on how common the disease is. Table 10.4 shows, for an RR fixed at 2, how the OR increases as the risk in the unexposed increases.

Table 10.4. The OR corresponding to an RR = 2, for different risks in the unexposed.

Risk in the unexposed (R0)	Risk in the exposed (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

The explanation can be seen from the formulas: $\text{OR} = \frac{\text{RR} \times (1 - R_0)}{1 - R_1}$. When both risks are small, the ratio (1 - R0) / (1 - R1) is close to 1, and OR ≈ RR. This is the "rare disease assumption": for diseases with a frequency below about 10%, the OR obtained in a case-control study is a good approximation of the RR. For common diseases, the OR exaggerates the effect if it is read as an RR. In the mg2026 dataset, the prevalence of diabetes in obese people is 33.3%: the OR (4.32) is

visibly larger than the PR (3.21). An article that reports an OR of 4.32 and translates it as "the risk is more than 4 times higher" is exaggerating.

10.7 Confidence Intervals and Statistical Testing

10.7.1 The Confidence Interval for the RR by the Logarithmic Method

The sampling distribution of the RR is skewed: the RR cannot be negative, but it can increase without limit. The natural logarithm of the RR, however, has an approximately normal distribution. The calculation has three steps: we move to the logarithmic scale, construct the usual interval (estimate $\pm 1.96 \times$ SE), and then return by exponentiation.

$$\text{SE}[\ln(\text{RR})] = \sqrt{\left[\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}\right]} \quad \text{CI 95\% for RR} = \exp[\ln(\text{RR}) \pm 1.96 \times \text{SE}]$$

For smoking-COPD: $\ln(2.1429) = 0.762$. $\text{SE} = \sqrt{\frac{1}{50} - \frac{1}{300} + \frac{1}{70} - \frac{1}{900}} = \sqrt{0.02 - 0.00333 + 0.01429 - 0.00111} = \sqrt{0.02984} = 0.1727$. Margin: 1.96 x 0.1727 = 0.339. On the logarithmic scale: 0.762 - 0.339 = 0.424 and 0.762 + 0.339 = 1.101. Exponentiating: $e^{0.424} = 1.53$ and $e^{1.101} = 3.01$. 95% CI for the RR: 1.53-3.01.

Note that the interval is not symmetric around 2.14 (2.14 - 1.53 = 0.61, but 3.01 - 2.14 = 0.87). The asymmetry is normal for ratios and is symmetric on the logarithmic scale; this is why forest plots use a logarithmic axis.

10.7.2 The Confidence Interval for the OR (Woolf Method)

$$\text{SE}[\ln(\text{OR})] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \quad \text{CI 95\% for OR} = \exp[\ln(\text{OR}) \pm 1.96 \times \text{SE}]$$

For smoking-COPD: $\ln(2.3714) = 0.8635$. $\text{SE} = \sqrt{\frac{1}{50} + \frac{1}{250} + \frac{1}{70} + \frac{1}{830}} = \sqrt{0.02 + 0.004 + 0.01429 + 0.00120} = \sqrt{0.03949} = 0.1987$. Margin: 1.96 x 0.1987 = 0.3895. Logarithmic scale: 0.474 and 1.253. Exponentiating: 1.61 and 3.50. 95% CI for the OR: 1.61-3.50.

The formula shows why small cells produce wide intervals: if one of the cells has the value 5, the term 1/5 = 0.2 dominates the sum. If a cell is 0, the OR and SE cannot be calculated; software then adds 0.5 to each cell (the Haldane correction) or uses exact methods.

10.7.3 The Chi-Square Test and Fisher's Test

The significance test for the association in the 2 x 2 table is the chi-square test of independence from Chapter 7. The hypotheses are stated in terms of the measure of association:

H0: RR = 1 (or OR = 1), the exposure is not associated with the disease H1: RR ≠ 1 (or OR ≠ 1), the exposure is associated with the disease

For smoking-COPD, the expected frequencies under H0 are: for cell a, $\frac{300 \times 120}{1200} = 30$; b: $\frac{300 \times 1080}{1200} = 270$; c: $\frac{900 \times 120}{1200} = 90$; d: $\frac{900 \times 1080}{1200} = 810$. All the observed-expected differences have the same absolute value, 20 (50 - 30 = 20; 250 - 270 = -20 etc.), which always happens in a 2 x 2 table. $\chi^2 = \frac{20^2}{30} + \frac{20^2}{270} + \frac{20^2}{90} + \frac{20^2}{810} = 13.33 + 1.48 + 4.44 + 0.49 = 19.75$, with 1 degree of freedom, $p < 0.001$. H0 is rejected.

When an expected frequency is below 5, Fisher's exact test is used. The Yates (continuity) correction reduces the value of χ^2 and makes the test more conservative; software reports it alongside the uncorrected test.

10.7.4 The Link between the Interval and the Test

A 95% CI for the RR or OR that does not contain 1 corresponds, almost always, to a test that is significant at α = 0.05; an interval that contains 1 corresponds to a nonsignificant result. The equivalence is not perfect, because the logarithmic interval and the chi-square test are calculated differently, and at the borderline they can give different conclusions (Worked Problem 10.5 shows such a case). Reporting the interval is preferable to reporting only the p value, because it shows both the size and the precision of the effect [70].

10.8 Measures of Impact: Attributable Risk, Attributable Fraction, NNT and NNH

10.8.1 Attributable Risk (Risk Difference)

The RR is a relative measure: it tells us how many times the risk increases, but not by how much in absolute terms. An RR of 2 can mean an increase from 1 to 2 cases per million (negligible for public health) or from 20% to 40% (enormous). The attributable risk (AR), also called the risk difference (RD), is:

$$\text{AR} = R_1 - R_0$$

For smoking-COPD: AR = 0.1667 - 0.0778 = 0.0889, that is, 8.9 percentage points over 5 years. Of 100 smokers, about 9 cases of COPD over 5 years are due to smoking (if the relationship is causal), on top of the 7-8 that would have occurred anyway. The 95% CI for the risk difference is calculated as for the difference of two proportions (Chapter 4): $\text{AR} \pm 1.96 \times \sqrt{\left[\frac{R_1(1-R_1)}{a+b} + \frac{R_0(1-R_0)}{c+d}\right]}$. For our example: $0.0889 \pm 1.96 \times \sqrt{\frac{0.1667 \times 0.8333}{300} + \frac{0.0778 \times 0.9222}{900}} = 0.0889 \pm 1.96 \times 0.0233 = 0.043 - 0.135$.

10.8.2 The Attributable Fraction

The attributable fraction in the exposed (AFe, also called the etiologic fraction in the exposed) shows what proportion of the cases occurring in the exposed is due to the exposure:

$$\text{AFe} = \frac{R_1 - R_0}{R_1} = \frac{\text{RR} - 1}{\text{RR}}$$

For smoking-COPD: AFe = (2.14 - 1) / 2.14 = 0.533. About 53% of the COPD cases occurring in smokers could be attributed to smoking.

The population attributable fraction (PAF) shows what proportion of all cases in the population could be avoided if the exposure disappeared. It depends both on the RR and on how widespread the exposure is (pe = the proportion of exposed people). Levin's formula is:

PAF = pe x (RR - 1) / [1 + pe x (RR - 1)]

For smoking-COPD, pe = 300 / 1200 = 0.25, so $\text{PAF} = \frac{0.25 \times 1.14}{1 + 0.25 \times 1.14} = \frac{0.286}{1.286} = 0.222$. About 22% of all COPD cases in this population would be avoidable by eliminating smoking. Equivalently, PAF = (Rtotal - R0) / Rtotal = (0.100 - 0.0778) / 0.100 = 0.222, where Rtotal = 120 / 1200 = 0.100 is the risk in the total population. The difference Rtotal - R0 = 2.2 percentage points is called the population attributable risk. These calculations represent what the course syllabus calls "estimating risk at the population level". They assume that the association is causal

and that there is no confounding; on observational data, they are optimistic estimates of the benefit of prevention.

10.8.3 The Number Needed to Treat (NNT) and the Number Needed to Harm (NNH)

The inverse of the absolute risk difference has a very intuitive clinical interpretation. In an intervention study with a favorable outcome, the NNT (number needed to treat) is the number of patients who need to be treated to obtain one additional favorable outcome compared with the comparison treatment:

$$\text{NNT} = \frac{1}{|R_1 - R_0|}$$

and it is always rounded up, to the next whole number (you cannot treat 5.5 patients). For a harmful factor, the same calculation is called the NNH (number needed to harm): how many people must be exposed for one additional case to occur. For smoking-COPD, $\text{NNH} = \frac{1}{0.0889} = 11.25$, that is, 12: for every 12 smokers followed for 5 years, one more case of COPD occurs than among non-smokers.

In the mg2026 study, at 6 months, 16 of 54 patients in arm C (counseling, app and teleconsultation) lost at least 5% of their weight, compared with 6 of 53 in arm A (standard counseling). The proportions are 29.6% and 11.3%; RR = 2.62 (95% CI 1.11-6.17); the absolute difference is 18.3 percentage points (95% CI 3.4-33.2); $\text{NNT} = \frac{1}{0.183} = 5.5$, that is, 6 (95% CI 3.0-29.1). Clinically: for every 6 patients who receive the full package instead of standard counseling, one additional patient achieves a relevant weight loss. The NNT interval is wide (from 3 to 29), because the risk difference is estimated imprecisely with only 107 patients; the upper limit (29) corresponds to the lower limit of the difference (3.4 percentage points). The OR for the same comparison is 3.30, again larger than the RR, because the outcome is not rare.

Figure 10.2 brings together in a forest plot the measures of association calculated in this chapter. Each square is the point estimate, and the horizontal line is the 95% confidence interval; the vertical dotted line marks the value 1 (no effect). An interval that crosses to the other side of the dotted line means a nonsignificant result.

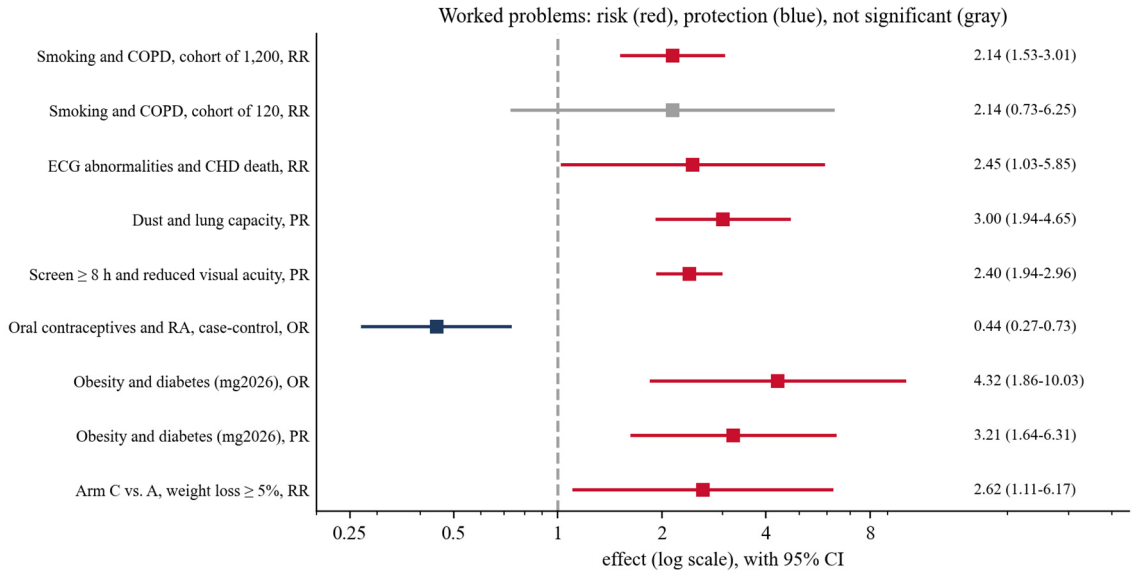


Figure 10.2. The measures of association from the examples and worked problems of the chapter, with the 95% CI calculated by the logarithmic method, on a logarithmic axis. Red: significant risk factor; blue: significant protective factor; gray: interval that contains 1 (nonsignificant). The same RR = 2.14 for smoking and COPD is significant with

1200 participants and nonsignificant with 120. The classic exercises of our discipline and the synthetic mg2026 dataset.

10.9 Confounding, Stratification and Effect Modification

10.9.1 The Confounder

A confounder is a third variable that simultaneously meets three conditions: it is associated with the exposure, it is a risk factor for the disease (independently of the exposure), and it does not lie on the causal pathway between exposure and disease. When it is present, the crude association between exposure and disease mixes the effect of the exposure with the effect of the confounder.

The classic teaching example is the apparent link between coffee consumption and a disease, confounded by smoking. Suppose that among 1000 people, smokers drink coffee much more often (400 of 500 smokers, versus 100 of 500 non-smokers) and that smoking increases the risk of the disease (20% in smokers, 5% in non-smokers), while coffee has no effect at all. Figure 10.3 shows what happens.

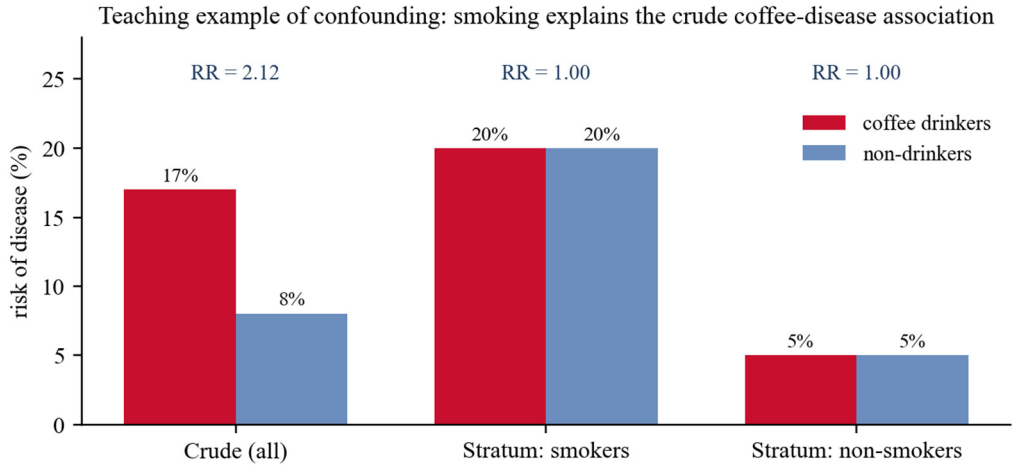


Figure 10.3. A teaching example of confounding (constructed data). In the crude analysis, coffee drinkers have a risk of 17%, versus 8% in the others (RR = 2.12). In each smoking stratum, the risk is identical in drinkers and non-drinkers (20% in smokers, 5% in non-smokers; RR = 1.00). The crude association arises entirely from the fact that smokers drink more coffee.

Crude analysis: among the 500 coffee drinkers there are 80 cases in smokers (20% of 400) and 5 in non-smokers (5% of 100), so 85 cases, a risk of 17%. Among the 500 non-drinkers there are 20 cases in smokers (20% of 100) and 20 in non-smokers (5% of 400), so 40 cases, a risk of 8%. Crude RR = 17% / 8% = 2.125, and crude OR = $\frac{85\times460}{415\times40} = 2.36$. A hasty researcher would conclude that coffee doubles the risk. The analysis stratified by smoking, however, shows RR = 1 in each stratum: among smokers, 20% versus 20%; among non-smokers, 5% versus 5%. Coffee has no effect at all; smoking, which is associated with coffee, creates the illusion.

10.9.2 Stratification and the Mantel-Haenszel Estimator

Stratification means dividing the data into subgroups (strata) according to the values of the confounder and calculating the measure of association in each stratum. If the measures in the strata are similar to one another but differ from the crude measure, there is confounding. To obtain a single adjusted estimate, the weighted average proposed by Mantel and Haenszel is used:

$$\mathrm{OR(MH)} = \frac{\sum\left(\frac{a_i \times d_i}{n_i}\right)}{\sum\left(\frac{b_i \times c_i}{n_i}\right)}$$

where the index i refers to stratum i, and ni is the total of the stratum. For the coffee example, the smokers' stratum has a = 80, b = 320, c = 20, d = 80, n = 500, and the non-smokers' stratum a = 5, b = 95, c = 20, d = 380, n = 500. The numerator is $\frac{80\times80}{500}+\frac{5\times380}{500}=12.8+3.8=16.6$; the denominator is $\frac{320\times20}{500}+\frac{95\times20}{500}=12.8+3.8=16.6$. OR(MH) = 1.00, versus a crude OR of 2.36.

In mg2026, we can check whether sex confounds the association between obesity and diabetes. In women (n = 86): a = 10, b = 19, c = 5, d = 52, OR = 5.47. In men (n = 74): a = 8, b = 17, c = 6, d = 43, OR = 3.37. $\mathrm{OR(MH)}=\frac{\frac{10\times52}{86}+\frac{8\times43}{74}}{\frac{19\times5}{86}+\frac{17\times6}{74}}=\frac{6.047+4.649}{1.105+1.378}=\frac{10.695}{2.483}=4.31$, practically identical to the crude OR of 4.32. Sex is not a confounder for this relationship, which is consistent with the OR for sex from the logistic regression (0.97, Chapter 9).

Stratification works well for one or two factors with few categories. When we want to adjust simultaneously for several factors, some of them numeric, we use multiple regression (linear, logistic, Cox), which essentially performs the same operation within a model. This is why articles frequently report a "crude OR" (or "unadjusted OR") and an "adjusted OR" (aOR); comparing them shows how much the adjustment mattered.

10.9.3 Effect Modification

Sometimes the measures of association in the strata genuinely differ from one another: the exposure has a larger effect in one subgroup than in another. The phenomenon is called effect modification (or interaction) and is not an error to be corrected, but information to be reported. If a drug reduces the risk of stroke by 50% in older people and not at all in young people, a common estimate would hide an important clinical result. In such situations, the measure is reported separately for each stratum.

The differences between strata must, however, be tested, because in small samples they can arise from sampling variation alone. In mg2026, if we stratify the obesity-diabetes association by age (under 56 years versus 56 years and over), we obtain OR = 11.67 in the younger patients (only 6 cases of diabetes) and OR = 3.99 in the older ones. The difference looks large, but the Breslow-Day test of homogeneity gives p = 0.37: we have no evidence of real effect modification, only imprecise estimates in a stratum with very few cases. The age-adjusted OR(MH) is 5.02 (95% CI 2.04-12.35), close to the adjusted OR from the logistic regression (6.01). A cross-sectional study of 693 students formally tested an interaction term in logistic regression and reported significant effect modification for HbA1c (p = 0.023) [6]; this is the correct way to claim that an effect differs between subgroups.

Case study: what happens to an association after adjustment

Four articles by our colleagues show, with figures, why the crude and the adjusted association must be read together. In a prospective cohort of 278 patients hospitalized with decompensated heart failure, followed for 12 months, the hemoglobin to red cell distribution width ratio (median 0.89) was associated with event-free survival in the Kaplan-Meier analysis (log-rank p = 0.027), but the association was no longer significant after adjustment for age, sex, NT-proBNP, ejection fraction and glomerular filtration rate in a Cox model (p = 0.240) [23]. The primary event occurred in 167 patients (60.1%). The most likely explanation is confounding: a hematological marker is associated with disease severity, and once severity is taken into account through NT-proBNP and ejection fraction, the additional information disappears. A result "lost after adjustment" is a valuable result, not a failure.

In a prospective cohort of 133 patients with high-grade serous ovarian carcinoma, BRCA mutations (present in 39.1%) were associated with a lower risk of progression in the univariate analysis: HR = 0.52, 95% CI 0.27-0.99, p = 0.048 [117]. The hazard ratio (HR) is read like an

RR: 0.52 means a 48% lower instantaneous risk of progression. After adjustment, HR = 0.57, p = 0.124. The point estimate changed little (from 0.52 to 0.57), but the interval widened and came to contain 1. Here the problem is not necessarily confounding, but precision: with 133 patients and several adjustment variables, uncertainty increases. The crude interval, which almost reaches 1 (0.99), already signaled the fragility of the result.

In a retrospective cohort of 213 patients treated in the emergency department for suspected bacterial infection, women were older than men (median 77.0 versus 67.0 years, $p < 0.001$), and a urinary source of infection was more frequent in women (51.0% versus 34.5%, $p = 0.015$) [9]. The abstract reports an OR = 2.14 (95% CI 1.20-3.81; $p = 0.010$), which remains significant even after the Benjamini-Hochberg correction for multiple testing (adjusted p 0.039). For in-hospital mortality, 7 women (7.0%) and 9 men (8.0%), OR = 0.87 (95% CI 0.31-2.43); the abstract also contains an OR = 0.80 (95% CI 0.27-2.35; $p = 0.690$), again with a wide interval that contains 1. The age difference between the sexes is exactly the situation in which age can confound a crude comparison between women and men. The very wide intervals for mortality (from 0.27 to 2.35) show that, with 16 deaths, the study can exclude neither an important protective effect nor a harmful one.

An online cross-sectional survey of 753 young adults (93.5% women) reported the prevalences of eating behaviors with Wilson intervals and ORs adjusted for age and sex: aOR 2.34, 2.49 and 2.73 for different BMI categories, all with $p < 0.001$ [20]. Adjustment for sex, however, has limited power when 93.5% of the participants are women, and the convenience sample (who chooses to answer an online questionnaire about eating) limits generalization, no matter how correctly the OR is calculated.

10.10 From Association to Causation

10.10.1 Systematic Errors in Observational Studies

An observed association can have four explanations: chance, systematic error (bias), confounding and causation. Chance is assessed through confidence intervals and statistical tests. Confounding was discussed above. Systematic errors fall into two broad categories. Selection bias arises when the way participants are recruited makes the compared groups differ systematically: controls chosen from a population different from that of the cases, volunteers healthier than the general population, losses to follow-up related both to exposure and to disease. Information (measurement) bias arises when exposure or disease is measured differently in the compared groups: recall bias in case-control studies, a physician who looks more carefully for the disease in patients known to be exposed, misclassification of exposure. Unlike random error, systematic error does not decrease as the sample grows: a large, biased study is precise around a wrong value.

10.10.2 The Bradford Hill Criteria

In 1965, Austin Bradford Hill, who together with Richard Doll had demonstrated the link between smoking and lung cancer through case-control and cohort studies, proposed nine aspects to be considered before drawing a causal conclusion from an observational association [108]. We apply them to the smoking-COPD example.

Strength of association: the larger the RR, the less likely it is that the association is produced entirely by an unknown confounder (in our example, RR = 2.14). Consistency: the association has been observed in several studies, in different populations, with different methods. Specificity: the exposure leads to a particular disease (the weakest criterion, because smoking causes many diseases). Temporality: the exposure precedes the disease; it is the only mandatory criterion and is

guaranteed by the cohort design. Biological gradient: the risk increases with the dose (more pack-years, higher risk). Biological plausibility: there is a known mechanism (chronic inflammation of the airways). Coherence: the association does not contradict what is known about the natural history of the disease. Experimental evidence: quitting smoking slows the decline in lung function. Analogy: other inhaled exposures (occupational dusts) cause similar diseases.

Bradford Hill did not present them as a checklist, but as viewpoints which, weighed together, make a causal explanation more or less credible. A study that reports an adjusted OR with a dose-response gradient, for example an OR of 1.43 for each point of a dietary risk score (95% CI 1.23-1.66), provides a stronger argument than one that reports only a comparison between exposed and unexposed people [4].

10.10.3 Reporting Observational Studies: STROBE

For the reader to be able to judge all these aspects, observational studies must be reported completely. The STROBE guideline (Strengthening the Reporting of Observational Studies in Epidemiology) comprises 22 items for cohort, case-control and cross-sectional studies [25]. Among the requirements: stating the design in the title or abstract, describing the sources of participants and the selection criteria (and, in case-control studies, the way the controls were chosen), clearly defining the exposure, the outcome and the confounders, explaining how missing data were handled, reporting the number of participants at each stage, reporting crude and adjusted estimates with their confidence intervals, and discussing the limitations, including the likely direction of the systematic errors. An abstract that gives an OR without a confidence interval does not respect the spirit of this guideline.

Case study: odds ratios in our colleagues' studies, read critically

A retrospective comparative study of 1418 pregnant women (279 with gestational diabetes, 74 with type 1 diabetes, 107 with type 2 diabetes and 958 controls) analyzed the risk factors for complications, with ORs and 95% CIs [118]. Age has OR = 1.03 (95% CI 1.01-1.05; p = 0.002), obesity OR = 2.37 (95% CI 1.42-3.94; p < 0.001), and chronic hypertension OR = 2.51 (95% CI 1.26-5.01; p = 0.009). An OR of 1.03 seems small next to 2.37, but the comparison is misleading: the OR for age refers to one year, while that for obesity refers to the presence versus absence of a condition. For 10 years, the OR for age becomes 1.03^10 ≈ 1.34. The unit of the predictor must always be read before comparing ORs.

A cross-sectional study of 921 students found 24.1% to be overweight or obese and, for one behavioral pattern, aOR = 7.20 (95% CI 4.26-12.17), plus aOR = 1.98 (95% CI 1.26-3.10) for another factor and 1.43 for each point of a dietary risk score (95% CI 1.23-1.66; p < 0.001) [4]. An aOR of 7.20 is a strong association, but, with an outcome prevalence of 24%, it should not be translated as a "7 times higher risk" (Table 10.4). The cross-sectional design leaves the direction open: did eating behavior lead to excess weight, or did overweight people change their behavior? In the same population, another cross-sectional study of 693 students reported a prevalence of prediabetes of 21.1% and an OR = 1.78 for each category of an ordered dietary pattern (95% CI 1.38-2.30; p < 0.001), with prevalences of 15.4%, 20.0% and 27.3% in the three categories (chi-square p = 0.003) [6]. The monotonic increase in prevalence across categories is a gradient, one of the Bradford Hill arguments.

A retrospective cohort study of 486 patients with COVID-19 (200 without diabetes, 62 with type 1 diabetes and 224 with type 2 diabetes) reported an OR of 2.24 for admission to intensive care and an OR of 2.46 for mechanical ventilation [8]. The abstract gives neither the confidence intervals nor the p values for these ORs. As an exercise in critical reading, the natural question is: how precise are these estimates? With 62 patients with type 1 diabetes, an

OR of 2.24 could have a wide interval, possibly including 1; without the interval, the reader cannot judge. This is exactly the situation that the STROBE guideline tries to prevent.

Finally, a retrospective analysis of hospital admissions for oral and oropharyngeal squamous cell carcinoma in 2018-2024 (248 admissions) used an interrupted time series analysis: the mean monthly number of admissions fell from 4.9 before the pandemic to 2.6 during the pandemic and 1.6 afterward, with an incidence rate ratio IRR = 0.973 (95% CI 0.950-0.996; p = 0.024) for the trend, and cervical lymph node dissection was associated with the period through an adjusted OR of 1.87 (95% CI 1.06-3.29; p = 0.030) [119]. The IRR is interpreted as an RR for rates: 0.973 per month means a decrease of about 2.7% per month. The study combines an ecological approach (rates by period) with an individual one (logistic regression on patients).

10.11 Worked Problems

All the numerical results below were calculated with software and checked by hand; the CIs for the RR and OR are calculated by the logarithmic method (Katz for the RR, Woolf for the OR), the same one used by default in Epi Info and MedCalc.

Worked Problem 10.1. Smoking and COPD: A Cohort Study with 1200 Participants

Statement. A total of 1200 initially healthy people (smokers and non-smokers) are enrolled and followed for 5 years. Of the 300 smokers, 50 develop COPD; of the non-smokers, 70. (a) State the research hypotheses. (b) State the statistical hypotheses. (c) Construct the contingency table. (d) Is smoking a risk factor or a protective factor? Calculate the RR and OR. (e) Is the risk statistically significant? What test do we use? (f) What is the clinical interpretation?

(a) Research hypothesis: smoking increases the risk of developing COPD.

(b) Statistical hypotheses: H0: RR = 1, the risk of COPD is the same in smokers and in non-smokers (smoking is not associated with COPD); H1: RR ≠ 1, the risk differs.

(c) Type of study: prospective cohort (it starts from the exposure, in healthy people, and follows the onset of disease). The table is Table 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 (95% CI 1.53-3.01, calculation in Section 10.7.1). $\text{OR} = \frac{50\times830}{250\times70} = 2.37(;\ \text{CI}\,95\%\ 1.61 - 3.50)$. RR > 1 and the interval does not contain 1: smoking is a risk factor. Since this is a cohort study, the main measure is the RR; the OR is reported only as additional information.

(e) Chi-square test (all expected frequencies are well above 5): $\chi^2 = 19.75$, df 1, $p < 0.001$ (with the Yates correction 18.78, $p < 0.001$). H0 is rejected: the association is statistically significant. The wording "we accept H1" is tolerated in the lectures, but the correct wording is "we reject H0".

(f) Clinical interpretation: over the 5 years, smokers had a risk of COPD about 2.1 times higher than non-smokers (plausibly between 1.5 and 3 times). The absolute difference is 8.9 percentage points (95% CI 4.3-13.5), NNH = 12: for every 12 smokers followed for 5 years, one additional case of COPD occurs. The attributable fraction in the exposed is 53%, and the population attributable fraction 22%. Beware of the wrong calculation with the denominator 1200 for non-smokers, which gives a false RR of 2.86.

Worked Problem 10.2. The Same Risk, a Sample Ten Times Smaller

Statement. Suppose that only 120 participants are enrolled, of whom 30 are smokers. Of these, 5 smokers and 7 non-smokers become ill. (a) Calculate the RR and OR. (b) Is the risk significant? (c) Justify the difference from Problem 10.1.

Solution. The table: a = 5, b = 25, c = 7, d = 83 (non-smokers 120 - 30 = 90; healthy 90 - 7 = 83).

(a) $R_1 = \frac{5}{30} = 16.7\%$; $R_0 = \frac{7}{90} = 7.8\%$; RR = 2.14, exactly as in Problem 10.1. OR $= \frac{5\times83}{25\times7} = \frac{415}{175} = 2.37$, also identical.

(b) $\text{SE}[\ln(\text{RR})] = \sqrt{\frac{1}{5} - \frac{1}{30} + \frac{1}{7} - \frac{1}{90}} = \sqrt{0.2 - 0.0333 + 0.1429 - 0.0111} = \sqrt{0.2984} = 0.546$. 95% CI: $\exp(0.762 \pm 1.96 \times 0.546) = \exp(0.762 \pm 1.071)$, that is, $e^{-0.309} = 0.73$ and $e^{1.833} = 6.25$. 95% CI for the RR: 0.73-6.25. $\text{SE}[\ln(\text{OR})] = \sqrt{\frac{1}{5} + \frac{1}{25} + \frac{1}{7} + \frac{1}{83}} = \sqrt{0.3950} = 0.628$; 95% CI for the OR: 0.69-8.13. Both intervals contain 1. The expected frequencies are 3, 27, 9 and 81; one is below 5, so the appropriate test is Fisher's: p = 0.17 (uncorrected chi-square 1.98, p = 0.16). H0 is not rejected.

(c) The measures of association are identical, but with 12 cases in total the information is insufficient to distinguish an RR of 2 from an RR of 1. The correct conclusion is not "smoking is not a risk factor", but "the study does not have enough statistical power to demonstrate the association". Absence of evidence is not evidence of absence. Figure 10.2 shows the two intervals one below the other: the same estimate, very different precision.

Worked Problem 10.3. Monitor Exposure and Visual Acuity (Cross-Sectional Study)

Statement. In a cross-sectional study of 1000 people, 200 work at least 8 hours a day in front of a monitor, of whom 90 have reduced visual acuity; of the 800 who work less, 150 have reduced acuity. (a) What is the risk factor and what might its influence be? (b) Calculate the appropriate measure of association. (c) What are the odds of reduced acuity in those who work less than 8 hours a day? (d) Is there a significant association? (e) Calculate the 95% CI for the OR.

Solution. (a) The risk factor investigated is prolonged exposure to the monitor; the possible influence is an increase in the frequency of reduced visual acuity. The table: a = 90, b = 110, c = 150, d = 650.

(b) The study is cross-sectional, so we measure prevalences: 90 / 200 = 45.0% in the exposed and 150 / 800 = 18.75% in the unexposed. The prevalence ratio PR $= \frac{45.0}{18.75} = 2.40$(; CI 95% 1.94 − 2.96). OR $= \frac{90\times650}{110\times150} = \frac{58\,500}{16\,500} = 3.55$.

(c) The odds in those exposed for less than 8 hours: 150 / 650 = 0.231 (about 1 person with reduced acuity to 4.3 with normal acuity). In the exposed, the odds are 90 / 110 = 0.818. If the problem asks for the OR of those with short exposure versus those with long exposure, it is the inverse: 1 / 3.55 = 0.28.

(d) $\chi^2 = 60.44$, df 1, $p < 0.001$; the association is significant.

(e) ln(3.545) = 1.266; $\text{SE} = \sqrt{\frac{1}{90} + \frac{1}{110} + \frac{1}{150} + \frac{1}{650}} = \sqrt{0.01111 + 0.00909 + 0.00667 + 0.00154} = \sqrt{0.02841} = 0.1685$; 1.266 ± 0.330 gives 0.935 and 1.596; 95% CI for the OR: 2.55-4.93.

Interpretation: the prevalence of reduced acuity is 2.4 times higher in those who work at least 8 hours a day at a monitor. The OR (3.55) greatly exceeds the PR (2.40) because the condition is common (24% overall), so the OR should not be read as a risk. The cross-sectional design does not allow conclusions about direction: it is possible, for example, that people in certain occupations also have other characteristics (age) associated with reduced acuity.

Worked Problem 10.4. Oral Contraceptives and Rheumatoid Arthritis (Case-Control Study)

Statement. A researcher selected 100 women with confirmed rheumatoid arthritis and 200 women treated at the same institution for other musculoskeletal diseases. The use of oral contraceptives before diagnosis was established from the medical records: 40 of the cases and 120 of the controls had used them. (a) What type of study is it? (b) Calculate the OR. (c) Calculate the 95% CI for the OR. Can the RR be calculated? (d) Is there a significant association?

Solution. (a) Case-control study: selection was made by disease (cases with arthritis, controls without), and the exposure was looked for in the past. The controls come from the same institution, which is a reasonable choice. The table: a = 40 (exposed cases), b = 120 (exposed controls), c = 60 (unexposed cases), d = 80 (unexposed controls).

(b) $\mathrm{OR} = \frac{40\times80}{120\times60} = \frac{3200}{7200} = 0.44$. In the cases, the odds of having used contraceptives are 0.44 times those of the controls (40% of the cases versus 60% of the controls were exposed).

(c) $\ln(0.444) = -0.811$; $\mathrm{SE} = \sqrt{\frac{1}{40}+\frac{1}{120}+\frac{1}{60}+\frac{1}{80}} = \sqrt{0.025+0.00833+0.01667+0.0125} = \sqrt{0.0625} = 0.250$; -0.811 ± 0.490 gives -1.301 and -0.321; 95% CI for the OR: 0.27-0.73. The RR cannot be calculated: the proportion of ill women in the sample (100 of 300) was fixed by the researcher and does not reflect the risk in the population. A mechanical calculation of the "RR" (40/160 versus 60/140, that is, 0.58) has no interpretation as a risk. If rheumatoid arthritis is rare in the population, the OR = 0.44 nevertheless approximates the RR.

(d) $\chi^2 = 10.71$, df 1, p = 0.001; the interval for the OR does not contain 1. The association is significant. Since OR < 1, oral contraceptive use appears as a protective factor in this study. A causal interpretation calls for caution: controls with other musculoskeletal diseases might, for reasons related to their diseases, have a different contraceptive use from the general population (selection bias).

Worked Problem 10.5. ECG Abnormalities and Coronary Death: A Borderline Result

Statement. A total of 47 men aged 40-64 years with ECG abnormalities at the initial examination and 144 men of the same age without abnormalities were followed for 20 years. Of these, 8 in the first group and 10 in the second died of coronary heart disease. (a) The type of study. (b) Test the hypothesis that ECG abnormalities are a risk factor. (c) Calculate and interpret the RR. (d) Calculate the OR and the 95% CI.

Solution. (a) Prospective cohort: the groups are defined by exposure (ECG abnormalities) and followed for death. The table: a = 8, b = 39, c = 10, d = 134, N = 191.

(b) H0: RR = 1; H1: RR ≠ 1. Expected frequencies: $\frac{a\,47\times18}{191} = 4.43$ (below 5), b 42.57, c 13.57, d 130.43. Uncorrected chi-square: 4.22, p = 0.040; with the Yates correction: 3.12, p = 0.077; Fisher's exact test: p = 0.0485. Because one expected frequency is below 5, the recommended test is Fisher's, and p = 0.0485 is just below the 0.05 threshold.

(c) $R_1 = \frac{8}{47} = 17.0\%$; $R_0 = \frac{10}{144} = 6.9\%$; RR = 2.45. SE $= \sqrt{\frac{1}{8} - \frac{1}{47} + \frac{1}{10} - \frac{1}{144}} = \sqrt{0.1968} = 0.444$; CI 95% $= \exp(0.897 \pm 0.870) = 1.03 - 5.85$.

(d) OR $= \frac{8\times134}{39\times10} = \frac{1072}{390} = 2.75$; SE $= \sqrt{\frac{1}{8} + \frac{1}{39} + \frac{1}{10} + \frac{1}{134}} = \sqrt{0.2581} = 0.508$; CI 95% $= \exp(1.011 \pm 0.996) = 1.02 - 7.44$.

Interpretation: ECG abnormalities are associated with a risk of coronary death about 2.5 times higher, but the evidence is borderline. The confidence intervals start just above 1 (1.03 and 1.02), the uncorrected test and Fisher's test are barely significant, and the test with the Yates correction is not. With 18 deaths in total, an honest conclusion is: "data compatible with an increased risk, but imprecise; the result must be confirmed in a larger sample". The student who writes only "p < 0.05, so ECG abnormalities are a risk factor" loses half the information.

Worked Problem 10.6. Dust Exposure and Age: Two Risk Factors

Statement. Part I: 320 retirees aged 70-75 years are enrolled, of whom 80 worked for at least 5 years in a dusty environment. On investigation, in both the exposed and the unexposed group, 30 people each have reduced lung capacity. Part II: to study age as an independent risk factor, 200 people aged 30-50 years, not exposed to dust, are also investigated, of whom 12 have reduced lung capacity. Analyze both parts.

Solution, Part I. Type of study: cross-sectional, with occupational exposure reconstructed from the past (it can also be viewed as a retrospective cohort). The table: a = 30, b = 50 (80 - 30), c = 30, d = 210 (240 - 30). Note the trap in the problem statement: "30 in both groups" does not mean equal risk, because the groups are of different sizes. $R_1 = \frac{30}{80} = 37.5\%$; $R_0 = \frac{30}{240} = 12.5\%$. PR = 3.00 (95% CI 1.94-4.65). OR $= \frac{30\times210}{50\times30} = \frac{6300}{1500} = 4.20$(; CI 95% $2.32 - 7.60$). χ^2 = 24.62, p < 0.001. Dust exposure is associated with a three times higher frequency of reduced lung capacity; the absolute difference is 25 percentage points, and the attributable fraction in the exposed $\frac{\text{RR}-1}{\text{RR}} = \frac{2}{3} = 67\%$.

Part II. To isolate the effect of age, we compare only unexposed people, so that dust exposure does not confound the comparison: unexposed older people 30 / 240 = 12.5% versus unexposed young adults 12 / 200 = 6.0%. The table: a = 30, b = 210, c = 12, d = 188. RR $= \frac{12.5}{6.0} = 2.08$(; CI 95% $1.10 - 3.96$); OR $= \frac{30\times188}{210\times12} = \frac{5640}{2520} = 2.24$(; CI 95% $1.11 - 4.50$); χ^2 = 5.34, p = 0.021. Advanced age is also a risk factor, roughly twofold. If we compared the exposed older people with the unexposed young people, we would obtain RR $= \frac{37.5}{6.0} = 6.25$, but this figure mixes the effect of age with that of dust. The lack of exposed young people prevents us from checking whether the effect of dust is the same at all ages (effect modification). The design lesson: the choice of comparison group decides which question gets answered.

Worked Problem 10.7. Obesity and Diabetes in mg2026: PR, OR and Attributable Fraction

Statement. At baseline in mg2026, of the 54 obese patients (BMI $\geq$ 30), 18 have diabetes confirmed by OGTT; of the 106 non-obese patients, 11. Calculate the PR, the OR, the 95% CIs, the significance test and the population attributable fraction. Comment on the difference between the OR and the PR.

Solution. The table: a = 18, b = 36, c = 11, d = 95, N = 160. The study is cross-sectional (baseline analysis), so the prevalences are 18 / 54 = 33.3% in the obese and 11 / 106 = 10.4% in the non-obese.

$\mathrm{PR} = \frac{33.3}{10.4} = 3.21.$ $\mathrm{SE}[\ln(\mathrm{PR})] = \sqrt{\frac{1}{18} - \frac{1}{54} + \frac{1}{11} - \frac{1}{106}} = \sqrt{0.0556 - 0.0185 + 0.0909 - 0.0094} = \sqrt{0.1185} = 0.344;$ $\mathrm{CI}\,95\% = \exp(1.167 \pm 0.675) = 1.64 - 6.31.$

$\mathrm{OR} = \frac{18\times95}{36\times11} = \frac{1710}{396} = 4.32.$ $\mathrm{SE}[\ln(\mathrm{OR})] = \sqrt{\frac{1}{18} + \frac{1}{36} + \frac{1}{11} + \frac{1}{95}} = \sqrt{0.1847} = 0.430;$ $\mathrm{CI}\,95\% = \exp(1.463 \pm 0.842) = 1.86 - 10.03.$

The expected frequencies are 9.8, 44.2, 19.2 and 86.8 (all above 5). $\chi^2 = 12.70$, $p < 0.001$ (Yates 11.20, p = 0.0008; Fisher p = 0.0008). The association is significant.

Population attributable fraction: pe = 54 / 160 = 0.3375; $\mathrm{PAF} = \frac{0.3375\times2.21}{1+0.3375\times2.21} = \frac{0.746}{1.746} = 0.43.$ Equivalently, (18.1% - 10.4%) / 18.1% = 0.43. If the association were causal and obesity could be eliminated, about 43% of the cases of diabetes in this sample would be avoided. The conditional wording is mandatory: the data are cross-sectional and do not establish the temporal sequence.

The OR (4.32) is much larger than the PR (3.21) because diabetes is common in obese people (one third). Reporting the OR as a "more than four times higher risk" would be an exaggeration. Adjustment for sex by Mantel-Haenszel does not change the OR (4.31), and the logistic regression adjusted for age and sex gives an OR of 6.01 (Chapter 9). In Figure 10.2, the OR interval is wider than that of the PR, as usual.

10.12 Proposed Exercises

1. In 1985, the link between the intrauterine device (IUD) and infertility was studied. Of 283 infertile women, 89 had used an IUD; of 3833 control women, 640. (a) What type of study is it? (b) Construct the 2 x 2 table. (c) Calculate the OR with its 95% CI and interpret it. (d) Apply the chi-square test at $\alpha = 0.05$.
2. In a study of 1024 mothers, 23 of the 107 mothers who smoked and 117 of the 917 non-smoking mothers had children with Down syndrome, the cases being identified from birth certificates. Calculate the OR and its 95% CI and discuss whether the OR can be approximated by the RR.
3. In the mg2026 study, at 6 months, 16 of 53 patients in arm B and 6 of 53 in arm A lost at least 5% of their weight. Calculate the RR, the absolute risk difference and the NNT. Compare with the result for C versus A (RR 2.62; NNT 6).
4. In mg2026, of the 53 smokers, 10 have diabetes confirmed by OGTT; of the 107 non-smokers and former smokers, 19. Calculate the PR and the OR with 95% CIs and state the statistical and the clinical conclusion.
5. In mg2026, of the 54 obese patients, 21 have SBP ≥ 140 mmHg; of the 106 non-obese patients, 20. Calculate the PR, the OR, the 95% CIs and χ^2. Why does the OR differ from the PR?
6. A cohort of workers was followed for contact dermatitis. Those exposed to a solvent accumulated 400 person-years and 10 cases; the unexposed, 800 person-years and 8 cases. Calculate the incidence rate in each group (per 100 person-years) and the incidence rate ratio.

7. A chronic disease has an incidence rate of 5 cases per 1000 person-years and a mean duration of 4 years. Estimate the prevalence. How would the prevalence change if a new treatment doubled the mean duration of the disease without curing it?
8. A drug produces an adverse effect in 6% of treated patients, versus 2% on placebo. Calculate the risk difference, the RR and the NNH. What does the NNH mean clinically?
9. An exposure has RR = 3 and is present in 40% of the population. Calculate the attributable fraction in the exposed and the population attributable fraction.
10. A study is stratified by a confounder. Stratum 1: a = 20, b = 80, c = 10, d = 60. Stratum 2: a = 30, b = 20, c = 60, d = 60. Calculate the OR in each stratum, the crude OR (from the combined table) and the Mantel-Haenszel OR. What conclusion do you draw?
11. In mg2026, of the 65 patients from rural areas, 10 have diabetes; of the 95 from urban areas, 19. Calculate the OR with its 95% CI. Is it correct to say that a rural setting "protects" against diabetes?

Common Pitfalls

- Constructing the 2 x 2 table incorrectly: rows and columns swapped or a total put in place of a cell; the risk in the unexposed is calculated only among the unexposed (smoking-COPD: correct RR 2.14, not 2.86).
- Calculating the RR in a case-control study; there the proportion of ill people is fixed by the researcher and only the OR is calculated.
- Reading an OR for a common disease as a relative risk: the OR exaggerates the effect when the outcome is common (OR 4.32 versus PR 3.21 in mg2026).
- Concluding that "the factor has no effect" from a nonsignificant result in a small sample; the correct statement is "we could not demonstrate an effect" (Problem 10.2).
- Ignoring expected frequencies below 5 and mechanically applying the chi-square test instead of Fisher's.
- Interpreting an OR per unit (per year, per point) as if it were an OR for a binary variable.
- Confusing association with causation in cross-sectional studies, where the temporal sequence cannot be established.
- Reporting an OR or RR without a confidence interval.
- Confusing the confounder with effect modification: the former is corrected by adjustment, the latter is reported by stratum.

Key Points

- Table notation: a = exposed ill, b = exposed healthy, c = unexposed ill, d = unexposed healthy.
- RR $= \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$; OR $= \frac{a \times d}{b \times c}$; RR > 1 risk, RR = 1 indifferent, RR < 1 protection.
- Cohort and experimental studies: RR (and OR); cross-sectional: PR (and OR); case-control: OR only.
- OR ≈ RR only for rare diseases; when the disease is common, the OR is farther from 1 than the RR.
- 95% CI: $\mathrm{SE}[\ln(\mathrm{RR})] = \sqrt{\left[\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}\right]}$; $\mathrm{SE}[\ln(\mathrm{OR})] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$; an interval that contains 1 means a nonsignificant result.
- The attributable risk (R1 - R0) measures the absolute impact; NNT or NNH $= \frac{1}{|R_1 - R_0|}$, rounded up.
- A confounder is associated with both the exposure and the disease; it is controlled by randomization, matching, stratification (Mantel-Haenszel) or multiple regression.
- The Bradford Hill criteria weigh the causal arguments; temporality is the only mandatory one. Observational studies are reported according to the STROBE guideline.

- In mg2026: obesity-diabetes PR 3.21 (1.64-6.31), OR 4.32 (1.86-10.03); arm C versus A: RR 2.62 (1.11-6.17), difference 18.3 percentage points, NNT 6.

Self-Assessment Questions

1. (SC) A study selects 150 patients with gastric cancer and 300 people without cancer and investigates their past consumption of smoked foods. The appropriate measure of association is:
    a) the relative risk
    b) the cumulative incidence
    c) the odds ratio
    d) the prevalence ratio
    e) the incidence rate
2. (SC) In a cohort, the risk of disease is 20% in the exposed and 10% in the unexposed. The number needed to produce one additional case (NNH) is:
    a) 2
    b) 5
    c) 10
    d) 20
    e) 0,1
3. (SC) An RR = 0.70 with a 95% CI of 0.55-0.89 indicates:
    a) a significant risk factor
    b) a significant protective factor, with a risk reduction of about 30%
    c) a nonsignificant result, because the RR is below 1
    d) a risk reduction of 70%
    e) an indifferent factor
4. (SC) In a cross-sectional study, the prevalence of the disease is 45% in the exposed and 18.75% in the unexposed. The correct statement is:
    a) the OR will be approximately equal to the prevalence ratio, because the disease is rare
    b) the prevalence ratio is 2.40, and the OR will be larger than 2.40
    c) the study allows the temporal sequence to be established
    d) the cumulative incidence can be calculated
    e) the OR will be smaller than the prevalence ratio
5. (SC) The 95% confidence interval for an OR = 2.37 is 0.69-8.13. The correct conclusion is:
    a) the exposure certainly doubles the odds of disease
    b) the association is significant at $\alpha = 0.05$
    c) the data do not allow H0 to be rejected; the study has not demonstrated an association
    d) the exposure is a protective factor
    e) the interval is wrong, because it is not symmetric around 2.37
6. (MC) The following statements about the cohort study are true:
    a) it starts from people without the disease, classified by exposure
    b) it allows direct calculation of the incidence and the relative risk
    c) it is the most efficient design for very rare diseases
    d) it can be prospective or retrospective
    e) it is exposed to losses to follow-up

7. (MC) The following statements about the odds ratio are true:
    a) it is calculated as (a x d) / (b x c)
    b) in a case-control study, the OR of exposure is equal to the OR of disease
    c) for rare diseases, the OR approximates the relative risk
    d) the OR is always equal to the RR
    e) an OR < 1 indicates a protective factor
8. (MC) A confounder:
    a) is associated with the exposure
    b) is a risk factor for the disease, independently of the exposure
    c) necessarily lies on the causal pathway between exposure and disease
    d) can be controlled by stratification and the Mantel-Haenszel estimator
    e) can be controlled at the design stage by randomization or matching
9. (MC) The following statements about the case-control study are true:
    a) it is suitable for rare diseases and for those with a long latency
    b) it is susceptible to recall bias
    c) the controls must come from the same population as the cases
    d) it allows the incidence of disease in the exposed to be calculated
    e) matching for age prevents age from being studied as a risk factor
10. (MC) The Bradford Hill criteria for assessing causation include:

a) temporality
b) biological gradient (dose-response relationship)
c) a p value below 0.05
d) consistency across studies
e) biological plausibility

Chapter 11. Statistical Methods in Medical Decision-Making: Bayesian Analysis and Evidence-Based Medicine

Objectives

By the end of this chapter, the student will be able to:

- explain the difference between probability understood as frequency and probability understood as degree of belief, and recognize both meanings in clinical reasoning;
- calculate unconditional and conditional probabilities from a disease-symptom table and apply Bayes' theorem, including through the law of total probability;
- turn a diagnostic problem into natural frequencies and construct the corresponding tree;
- convert from probability to odds and back and use likelihood ratios to obtain the post-test probability, both by hand calculation and with the Fagan nomogram;
- judge when two tests can be combined in series or in parallel and what the assumption of conditional independence implies;
- solve a differential diagnosis table with three diseases and several clinical signs;
- describe the testing threshold and the treatment threshold and calculate them in a simple example;
- interpret a credible interval and a Bayes factor (BF10) reported by JASP;
- apply the five steps of evidence-based medicine, read a forest plot and recognize the main cognitive errors in clinical reasoning.

11.1 Uncertainty in Clinical Decision-Making

A 58-year-old woman comes into the family practice for a prescription. She has a BMI of 31, her father had diabetes, and she does not exercise. She has no complaints. The physician wonders whether she has undiagnosed diabetes. There is no way to be sure without a test, but the physician does not start from zero either: in the population of adults at cardiometabolic risk, roughly one in five or six has undiagnosed diabetes, and her profile suggests that the risk is even higher. From this moment on, every new piece of information (the FINDRISC score, fasting plasma glucose, HbA1c) moves this estimate up or down.

This, in short, is the subject of the chapter. Diagnosis is not an act of instant recognition, but a succession of revisions of a probability. Experienced physicians make these revisions intuitively, sometimes very well, at other times with systematic errors that we will describe in Section 11.11. Statistics provides the tool that makes the revision explicit and verifiable: Bayes' theorem [19,120].

11.1.1 From Symptom to Diagnosis: A Problem of Probabilities

In the previous chapters we used probability mainly to describe sampling variability: how probable it is to obtain such a large difference between two means if H0 were true. In clinical decision-making, the question changes. The physician does not need the probability of the data under the hypothesis that the patient is healthy, but the probability that the patient is ill, given the data. The two probabilities are different, and confusing them is one of the oldest reasoning errors in medicine.

A classic example: 80% of patients with pneumonia have fever. This does not mean that 80% of patients with fever have pneumonia. The first statement describes p(fever / pneumonia), the second would describe p(pneumonia / fever). In a family practice, in winter, most febrile patients have trivial viral respiratory infections, so the second probability is much smaller than the first. Bayes' theorem is precisely the formula that links the two probabilities and tells us what additional information is needed to move from one to the other: how frequent the disease is in the population from which the patient comes.

11.1.2 Probability as Frequency and Probability as Degree of Belief

There are two ways of understanding the word "probability". In the frequentist sense, the probability of an event is the limit of the relative frequency of that event in a very large number of repetitions of the same experiment. When we say that the prevalence of undiagnosed diabetes in the mg2026 dataset is 18.1% (29 of 160), we are describing an observed frequency, which estimates a frequency in the population.

In the subjective or Bayesian sense, probability expresses a person's degree of belief in a statement, based on the information they have. When the physician says "I think there is a 30% probability that this patient has diabetes", the physician is not referring to a series of repetitions of the same patient (the patient either has or does not have diabetes), but to how convinced they are. This probability can change as new information appears, and Bayes' theorem says how it must change in order to remain coherent.

In practice, the two meanings meet. The physician's degree of belief before the test (the pre-test probability) should preferably be anchored in a frequency: the prevalence of the disease in similar patients. And the performance of tests (sensitivity, specificity) is estimated in a frequentist way, from diagnostic accuracy studies, as we will see in Chapter 12. Bayesian analysis in clinical decision-making therefore uses frequencies to update degrees of belief [63,64].

11.2 Conditional Probability, Revisited

11.2.1 The Disease-Symptom Table

The lectures of our discipline start from a contingency table with the presence (B+) or absence (B-) of the disease in the rows and the presence (S+) or absence (S-) of the symptom in the columns. The same table can be constructed for any disease-symptom pair, and in an expert diagnostic system dozens or hundreds of such tables are constructed.

Table 11.1. The disease-symptom table, with the notation used in the lectures.

	S+	**S-**	**Total**
B+	a	b	a + b
B-	c	d	c + d
Total	a + c	b + d	$N = a + b + c + d$

Two types of probability are calculated from this table. Unconditional (marginal) probabilities refer to the whole sample:

$$p(B^+) = \frac{a + b}{N}$$

$$p(S^+) = \frac{a + c}{N}$$

Conditional probabilities are calculated within a row or a column. For example, the probability of having the symptom among those without the disease is

$$p(S^+/B^-) = \frac{c}{c+d}$$

and the probability of having the disease among those with the symptom is

$$p(B^+/S^+) = \frac{a}{a+c}$$

The slash is read as "given" or "conditional on". Everything after the slash restricts the population we are referring to. Here the first practical rule appears: to know which denominator to use, look after the slash. If B- is written after the slash, the denominator is the total of row B-, that is, c + d.

Let us take an example with numbers. In a family practice, 200 adults with acute cough were evaluated during one winter. In 30 of them, the radiograph confirmed pneumonia. Fever above 38 °C was present in 24 of those with pneumonia and in 34 of those without pneumonia (fictitious data).

Table 11.2. Pneumonia (B) and fever (S) in 200 patients with acute cough (fictitious data).

	S+ (fever)	**S-**	**Total**
B+ (pneumonia)	24	6	30
B-	34	136	170
Total	58	142	200

We calculate:

$$p(B^+) = \frac{30}{200} = 0.15$$

$$p(S^+) = \frac{58}{200} = 0.29$$

$$p(S^+/B^+) = \frac{24}{30} = 0.80$$

$$p(S^+/B^-) = \frac{34}{170} = 0.20$$

$$p(B^+/S^+) = \frac{24}{58} = 0.414$$

Note the asymmetry: fever occurs in 80% of patients with pneumonia, but only 41.4% of febrile patients have pneumonia. The remaining 58.6% are febrile patients without pneumonia, who, being much more numerous in the population with cough, "dilute" the group of those with fever.

11.2.2 The Multiplication Rule and the Law of Total Probability

The probability that two events occur simultaneously is obtained with the multiplication rule:

$$p(B^+ \text{ and } S^+) = p(S^+/B^+) \times p(B^+) = p(B^+/S^+) \times p(S^+)$$

In our example, $p(B^+ \text{ and } S^+) = \frac{24}{200} = 0.12$. Let us check: 0.80 x 0.15 = 0.12 and 0.414 x 0.29 = 0.12. The two expressions are equal because they describe the same cell of the table, namely cell a.

The second tool is the law of total probability. The symptom can occur either in an ill person or in a person who is not ill, and the two situations are mutually exclusive. Therefore:

$$p(S^+) = p(S^+/B^+) \times p(B^+) + p(S^+/B^-) \times p(B^-)$$

For Table 11.2: $p(S^+) = 0.80 \times 0.15 + 0.20 \times 0.85 = 0.12 + 0.17 = 0.29$, exactly what we obtained from the margin of the table. The formula seems trivial when we have the complete

table, but it becomes indispensable when we do not: most of the time, the physician knows the sensitivity of a test (from the literature) and the prevalence of the disease (from experience or from epidemiological studies), but does not have a table of their own patients.

If there are several possible diseases, for example B1, B2 and B3, mutually exclusive and covering all possibilities, the formula extends naturally:

$$p(S) = p(S/B_1) \times p(B_1) + p(S/B_2) \times p(B_2) + p(S/B_3) \times p(B_3)$$

We will use this form in Section 11.6, for differential diagnosis.

11.3 Bayes' Theorem

11.3.1 The Derivation

We start from the two expressions of the multiplication rule for the same cell:

$$p(B/S) \times p(S) = p(S/B) \times p(B)$$

We divide both sides by p(S) and obtain Bayes' theorem, in the form used in the lectures of our discipline:

$$p(B/S) = \frac{p(S/B) \times p(B)}{p(S)}$$

Replacing the denominator with the law of total probability, we obtain the form we will use most often:

$$p(B^+/S^+) = \frac{p(S^+/B^+) \times p(B^+)}{p(S^+/B^+) \times p(B^+) + p(S^+/B^-) \times p(B^-)}$$

Each term has a name. p(B+) is the prior or pre-test probability: how probable the disease was before we knew anything about the symptom or the test. p(S+ / B+) is the likelihood of the symptom in ill people; when S is the positive result of a test, this is the sensitivity. p(S+ / B-) is the frequency of the symptom in people who are not ill; for a test, it is 1 - specificity, that is, the false positive rate. p(B+ / S+) is the posterior or post-test probability; for a positive test, it is the positive predictive value.

Applied to Table 11.2, without looking at cell a:

$$p(B^+/S^+) = \frac{0.80 \times 0.15}{0.80 \times 0.15 + 0.20 \times 0.85} = \frac{0.12}{0.29} = 0.414$$

We have recovered the value calculated directly. The power of the formula appears when we change only the prevalence. Suppose that the same fever, with the same frequency in ill and non-ill people, occurs in patients with cough admitted to a pulmonology ward, where pneumonia has a prevalence of 50%. Then $p(B^+/S^+) = \frac{0.80 \times 0.50}{0.80 \times 0.50 + 0.20 \times 0.50} = \frac{0.40}{0.50} = 0.80$. The same clinical sign has a predictive value of 41.4% in the family practice and of 80% in the hospital.

11.3.2 Natural Frequencies

Bayes' formula is correct, but not very intuitive. Studies of physicians' reasoning have repeatedly shown that many of them greatly overestimate the positive predictive value when they are given the problem in the form of percentages. The same problem becomes much easier if it is stated in natural frequencies: whole numbers of patients in an imaginary group [120].

Figure 11.1 shows the method for a test with a sensitivity of 80% and a specificity of 90%, applied in a population with a disease prevalence of 10%. We take 1000 patients. A prevalence of 10% means 100 ill and 900 non-ill people. Of the 100 ill people, the test is positive in 80 (sensitivity 80%) and negative in 20. Of the 900 non-ill people, a specificity of 90% means 810

negative tests and 90 false positive tests. In total we have 80 + 90 = 170 positive tests, of which only 80 belong to ill people.

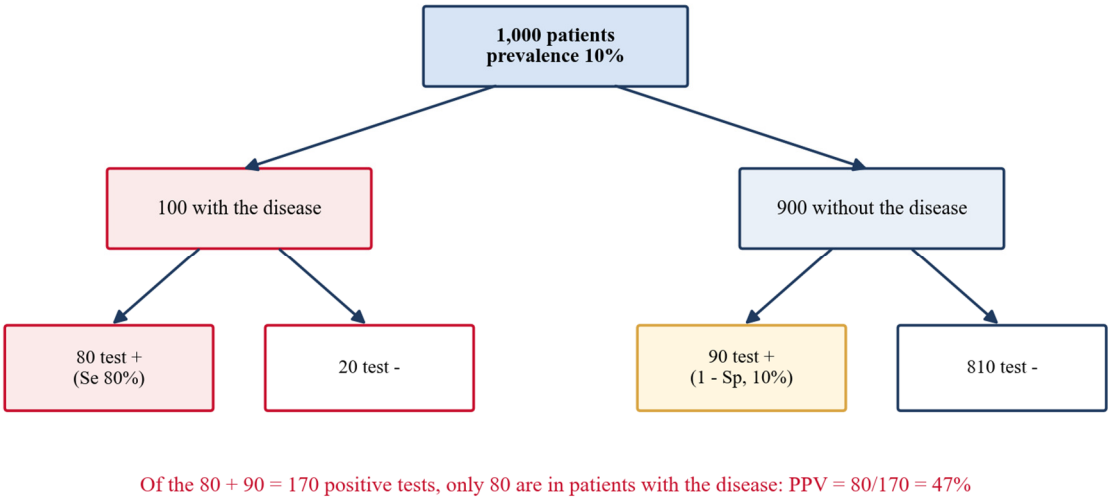


Figure 11.1. The natural frequency tree for a test with a sensitivity of 80% and a specificity of 90%, applied to 1000 patients with a disease prevalence of 10%. Of 170 positive tests, only 80 come from ill people, so the positive predictive value is 80/170 = 47%. Teaching example.

The positive predictive value is 80/170 = 47.1%. The negative predictive value is 810/(810 + 20) = 810/830 = 97.6%. A test that "errs" in only 10% of healthy people produces, in a population with a rare disease, more false positive results than true ones. The explanation is purely arithmetical: 10% of 900 (90) is more than 80% of 100 (80).

In the practical sessions I always recommend the same thing: when you are given a problem with percentages, rewrite it for 1000 or 10,000 patients and draw the tree. It takes two minutes and eliminates almost all reasoning errors.

11.3.3 The Pre-Test Probability: Prevalence and Clinical Context

The pre-test probability is not a property of the disease, but of the patient, in the context in which they are seen. The same diagnosis has very different pre-test probabilities in family medicine, in the specialist outpatient clinic and in the hospital. Family physicians see unfiltered patients, with vague symptoms and diseases in early stages; the prevalence of serious diseases is low. The hospital receives patients already selected by other physicians, so the prevalence is much higher. This is why a test with excellent performance in a university clinic can generate numerous false positive results when it is moved to primary care.

The pre-test probability is estimated from several sources: the prevalence of the disease in the population from which the patient comes (epidemiological data, registries, prevalence studies), the clinical data already known (age, sex, risk factors, symptoms, physical examination), validated clinical scores and, inevitably, the physician's experience. In the mg2026 dataset, the prevalence of diabetes confirmed by OGTT is 18.1% (95% Wilson CI 12.9-24.8), and this is the reasonable pre-test probability for an adult at cardiometabolic risk in a similar practice, before any other information. For an average adult without risk factors, the probability would be much lower; for an obese patient over 60 with a family history, it would be higher. The cross-sectional analysis of the mg2026 dataset shows, for example, a prevalence of 33.3% in obese people versus 10.4% in non-obese people.

Case study: colorectal cancer screening and Bayesian reasoning [101]

Iovanescu and colleagues (with M. Frandeș from our discipline) evaluated the reliability of combining the fecal immunochemical test (FIT) with flexible sigmoidoscopy in screening for colorectal neoplasia. A total of 2201 consecutive symptomatic patients referred for FIT were included; a sample of 252 people subsequently underwent colonoscopy (the reference standard), and real sigmoidoscopy was performed in a subsample of 57 patients. Colonoscopy found advanced neoplasia in 75 people (29.76%), and FIT was positive in 30.3% of the cases of advanced neoplasia. The compliance rate with referral for colonoscopy was 52% among those with a positive FIT, and the paired comparison of the two methods was made with the McNemar test ($P < 0.001$). The authors calculated the prior and posterior probabilities and showed that, at low prevalence levels, the negative predictive values lay between 81.5% and 90.12%. What does this study teach us? First, that the prevalence of almost 30% in the sample that underwent colonoscopy is not the prevalence in the screening population: the patients were symptomatic and selected. Second, that predictive values must be recalculated for the prevalence of the population in which the test will be used, precisely with Bayes' theorem. Third, that testing in series (FIT, then sigmoidoscopy) and in parallel changes the trade-off between sensitivity and specificity, a subject we return to in Section 11.5. Finally, the fact that only half of the FIT-positive patients accepted colonoscopy raises the problem of verification bias, discussed in Chapter 12.

11.4 Odds, Likelihood Ratios and the Odds Form of Bayes' Theorem

11.4.1 Converting between Probability and Odds

The odds of an event are the ratio between the probability that the event occurs and the probability that it does not occur:

$$\text{odds} = \frac{p}{1-p}$$

Conversely, the probability is obtained from the odds:

$$p = \frac{\text{odds}}{1 + \text{odds}}$$

A probability of 0.20 corresponds to odds of 0.20/0.80 = 0.25, that is, "1 to 4". A probability of 0.50 corresponds to odds of 1 ("one to one"), and a probability of 0.90 corresponds to odds of 9. For small probabilities, odds and probability are almost equal ($p = 0.02$ gives odds of 0.0204), but as the probability increases, the odds increase much faster and have no upper limit. You met odds in the chapter on risk analysis, with the odds ratio (OR); here we use them for a single patient.

Table 11.3. Correspondence between probability and odds.

Probability	Odds	Reading
0,05	0,053	about 1 to 19
0,10	0,111	1 to 9
0,181	0,221	about 1 to 4.5
0,25	0,333	1 to 3
0,50	1	1 to 1
0,75	3	3 to 1
0,90	9	9 to 1

11.4.2 Likelihood Ratios

The likelihood ratio (LR) of a test result tells us how many times more probable that result is in an ill person than in a person who is not ill [121]. For a dichotomous test we have two ratios:

$$\mathrm{LR}^{+} = \frac{p(T^{+}/B^{+})}{p(T^{+}/B^{-})} = \frac{\mathrm{Se}}{1-\mathrm{Sp}}$$

$$\mathrm{LR}^{-} = \frac{p(T^{-}/B^{+})}{p(T^{-}/B^{-})} = \frac{1-\mathrm{Se}}{\mathrm{Sp}}$$

An LR equal to 1 means that the result is equally frequent in ill and non-ill people, so it brings no information. A large LR+ (above 10) makes the disease much more probable; a small LR- (below 0.1) makes it much less probable. Between 2 and 5 for LR+, and between 0.2 and 0.5 for LR-, the test moves the probability moderately, sometimes sufficiently, sometimes not. These benchmarks are indicative and are used by many authors [120,121].

The major advantage of likelihood ratios over predictive values is that, like the sensitivity and specificity from which they derive, they do not (in principle) depend on prevalence. We can therefore carry them over from a published study into our own practice and apply them to the pre-test probability of our patient. The "in principle" caveat matters: if the spectrum of patients differs greatly (milder stages of the disease, comorbidities), both sensitivity and specificity change, and so do the LRs. In the case study in Section 11.8 we will see an example from our colleagues' research in which the specificity of a biomarker changes greatly depending on the time of sampling.

For the FINDRISC score in the mg2026 dataset, at the cut-off ≥ 12, the sensitivity is 75.9% (22 of 29), and the specificity 74.8% (98 of 131). Then:

$$\mathrm{LR}^{+} = \frac{0.759}{1-0.748} = \frac{0.759}{0.252} = 3.01$$

$$\mathrm{LR}^{-} = \frac{1-0.759}{0.748} = \frac{0.241}{0.748} = 0.32$$

A score of at least 12 is 3 times more frequent in people with diabetes than in those without diabetes; a score below 12 is about 3 times rarer in people with diabetes.

11.4.3 Bayes' Theorem in Odds Form

If we write Bayes' theorem once for B+ and once for B-, with the same test result T, and divide the two relations, the denominator p(T) cancels out and a remarkably simple formula remains:

post-test odds = pre-test odds x LR

The calculation algorithm has three steps: (1) convert the pre-test probability into odds; (2) multiply by the LR of the result obtained; (3) convert the post-test odds back into a probability.

For a person in the mg2026 dataset with FINDRISC ≥ 12, the pre-test probability is 0.181. The pre-test odds are 0.181/0.819 = 0.221. The post-test odds are 0.221 x 3.01 = 0.667. The post-test probability is 0.667/1.667 = 0.40, that is, 40.0%. This is exactly the positive predictive value calculated directly from the table, 22/55 = 40.0%, as expected. For a person with FINDRISC below 12: 0.221 x 0.32 = 0.071, so the post-test probability is 0.071/1.071 = 6.7%.

Figure 11.2 shows the relationship between the pre-test and the post-test probability for a few values of the LR. The diagonal line (LR = 1) is the useless test. The curves with a large LR rise quickly above the diagonal, especially in the region of low and medium pre-test probabilities; the curves with a small LR fall well below the diagonal, especially at high pre-test probabilities.

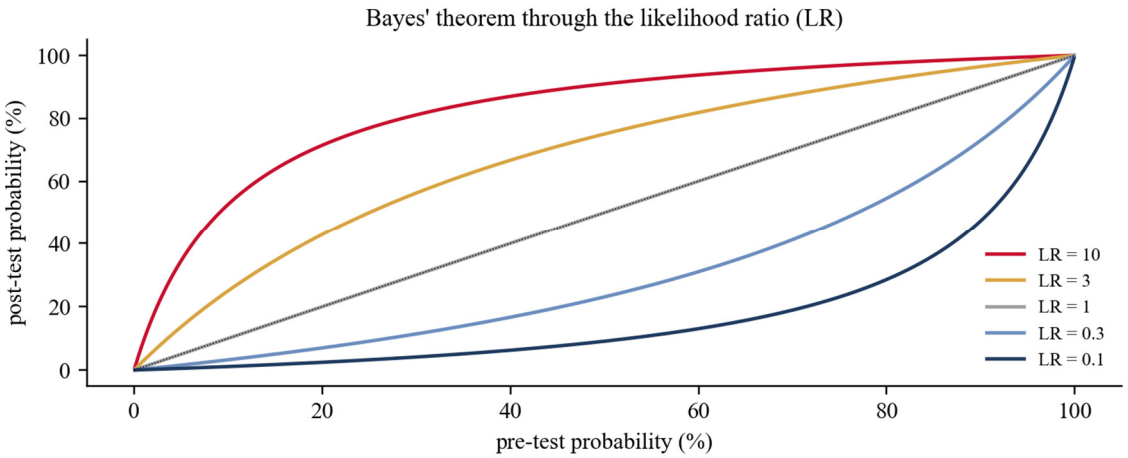


Figure 11.2. Post-test probability as a function of pre-test probability, for likelihood ratios between 0.1 and 10. The same test moves the probability much more in the middle range than at the extremes. Teaching example.

The figure also shows something important for everyday practice: a test contributes most when the pre-test probability is intermediate. If the patient has a pre-test probability of 2%, not even an LR+ of 10 takes the probability much above 17%; if it is 95%, a negative test with an LR- of 0.1 still leaves a probability of about 66%. This is why testing without a clear clinical question and without an estimate of the pre-test probability often produces more confusion than information.

11.4.4 The Fagan Nomogram

Before every physician had a calculator in their pocket, Fagan proposed a nomogram that performs the calculation graphically [122]. It has three vertical axes: on the left the pre-test probability, in the middle the likelihood ratio (on a logarithmic scale), on the right the post-test probability. A straight line is drawn from the pre-test point through the LR value, and the post-test probability is read where the line meets the right-hand axis. The construction works because the logarithm of the post-test odds is the sum of the logarithm of the pre-test odds and the logarithm of the LR, and a sum of logarithms can be represented with a straight ruler.

Figure 11.3 shows the nomogram with the FINDRISC example: from 18.1%, the line passing through LR+ = 3.01 reaches about 40%, and the one passing through LR- = 0.32 reaches about 6.7%.

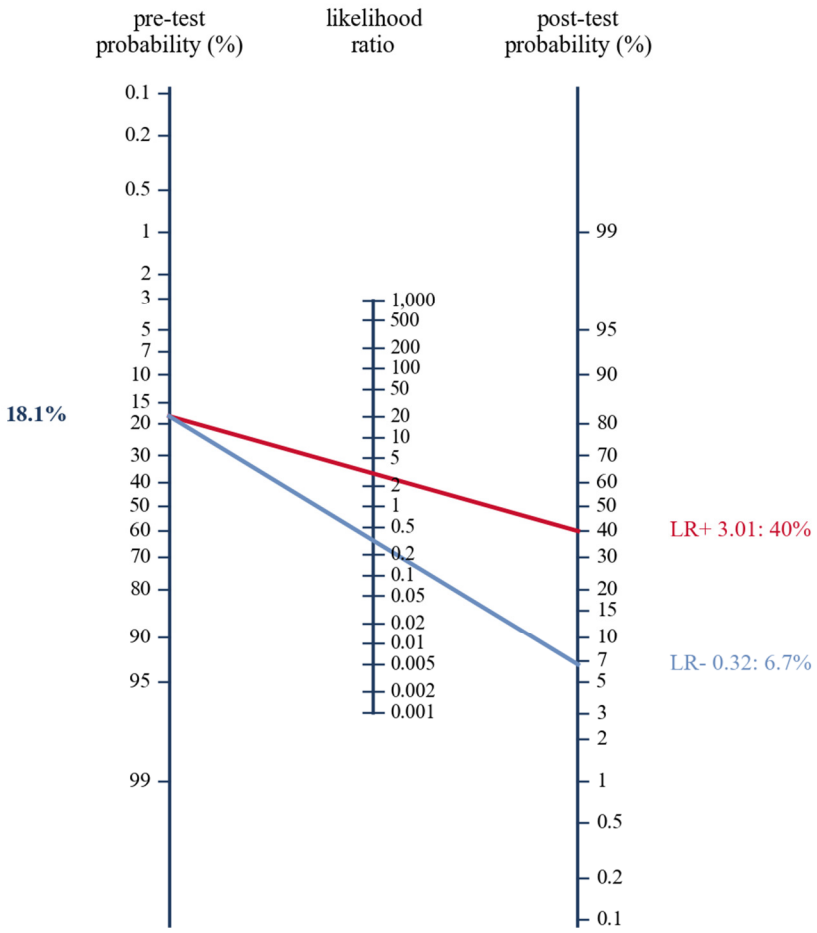


Figure 11.3. The Fagan nomogram applied to the FINDRISC score at the cut-off ≥ 12 in the mg2026 dataset: starting from a pre-test probability of 18.1%, a positive result (LR+ 3.01) leads to 40%, and a negative result (LR- 0.32) to 6.7%. Synthetic data (mg2026 dataset).

Today the nomogram has mainly teaching value, but it is also useful at the bedside for a quick estimate. The reading is accurate to within a few percent, which is sufficient for decision-making. For exact calculations, in Practical Session 11 we will build a Bayesian calculator in Google Sheets.

Case study: two cut-offs for the same elastography, one for ruling out and one for ruling in [3]

Moga and colleagues (with R. Lupușoru from our discipline) evaluated time-harmonic elastography for liver fibrosis, with vibration-controlled transient elastography (VCTE) as the reference method, in 165 patients. In the sample, 28.1% had cirrhosis (F4) according to VCTE. The authors reported not one cut-off but two. The rule-out cut-off for cirrhosis, below 1.61 m/s, had Se 90.7%, Sp 66.6%, PPV 55.7% and NPV 93.9%. The rule-in cut-off, above 1.83 m/s, had Se 65.1%, Sp 96.7%, PPV 90.3% and NPV 85.7%. The AUROC was 0.90 (95% CI 0.82-0.93). Let us translate this into likelihood ratios. For the rule-out cut-off, $LR^- = \frac{1-0.907}{0.666} = 0.14$, that is, a result below the cut-off greatly reduces the odds of cirrhosis. For the rule-in cut-off, $LR^+ = \frac{0.651}{1-0.967} = 19.7$, that is, a result above the cut-off increases the odds almost 20-fold. If we apply these LRs to the prevalence in the sample (28.1%), the odds form gives a PPV of about 88.5% at the rule-in cut-off and an NPV of about 94.8% at the rule-out cut-off, values close to those reported (the small differences come from rounding and from the exact number of patients analyzed at each cut-off). At a prevalence of 10%, as might be the case in a

> screening population, the same cut-offs would give a PPV of about 68.7% for the rule-in cut-off. Between the two cut-offs there remains a gray zone in which the test does not decide, and the patient needs another investigation. The study also reported reproducibility (ICC between 0.92 and 0.97), a piece of information that is often forgotten but decisive for a test used by different examiners.

11.5 Sequential Testing and Combining Tests

11.5.1 Tests in Series and in Parallel

A single test rarely decides a diagnosis. We usually combine tests, and the way they are combined changes the overall performance. There are two basic strategies.

In parallel testing, both tests are performed, and the patient is considered positive if at least one of the tests is positive (the "OR" rule). The strategy increases sensitivity (we miss fewer ill people), but decreases specificity (the number of false positives increases). It is used when missing an ill person is serious, for example in the emergency department, when the patient will not come back.

In serial testing, the second test is performed only in those positive on the first, and the patient is considered positive only if both tests are positive (the "AND" rule). The strategy increases specificity and PPV, but decreases sensitivity. It is economical, because the second test (usually more expensive or more invasive) is applied only to some of the patients. Colorectal cancer screening (FIT, then colonoscopy) and the diagnosis of HIV infection (a screening test, then a confirmatory test) are classic examples.

If the two tests are conditionally independent given disease status (we will explain shortly what this means), the combined performance is calculated simply:

in series: Se = Se1 x Se2; Sp = 1 - (1 - Sp1) x (1 - Sp2)

in parallel: Se = 1 - (1 - Se1) x (1 - Se2); Sp = Sp1 x Sp2

And in odds form, sequential testing means successive multiplication by the likelihood ratios:

final post-test odds = pre-test odds x LR1 x LR2

The post-test probability after the first test becomes the pre-test probability for the second. This is the most elegant property of Bayesian reasoning: information accumulates.

11.5.2 The Assumption of Conditional Independence

Two tests are conditionally independent if, among ill people, the result of one test tells us nothing about the result of the other, and likewise among people who are not ill. The assumption is often violated, especially when the tests measure closely related aspects of the same pathophysiology. The FINDRISC score and fasting plasma glucose, for example, both depend on obesity, age and sedentary lifestyle. A person without diabetes but with abdominal obesity is more likely to be "positive" on both.

The mg2026 dataset allows us to check this. We consider FINDRISC ≥ 12 (Se 75.9%, Sp 74.8%) and fasting plasma glucose ≥ 110 mg/dL (Se 82.8%, Sp 70.2%; 24 of 29 ill people positive, 92 of 131 non-ill people negative). If the tests were conditionally independent, the serial combination would have:

Se = 0.759 x 0.828 = 0.628; Sp = 1 - 0.252 x 0.298 = 0.925

The real data (patients positive on both tests), however, show Se 69.0% (20 of 29) and Sp 85.5% (112 of 131). The real specificity is lower than predicted, because among the 131 non-ill people, 19 are positive on both tests, many more than would be expected if the tests were

independent $\left(\frac{33\times39}{131} = 9.8\right)$. The practical consequence can be seen in the post-test probability. By successively multiplying the likelihood ratios (LR+ 3.01 for FINDRISC, LR+ 2.78 for blood glucose), we would estimate a post-test probability of 64.9% for a patient positive on both tests. The value observed in the data is only 20/39 = 51.3%. The independence assumption overestimated the gain brought by the second test.

Table 11.4. Combining FINDRISC ≥ 12 with fasting plasma glucose ≥ 110 mg/dL for diabetes confirmed by OGTT (synthetic data, mg2026 dataset, n = 160, 29 with diabetes).

Strategy	TP	FP	FN	TN	Se (%)	Sp (%)	PPV (%)	NPV (%)
FINDRISC ≥ 12	22	33	7	98	75,9	74,8	40,0	93,3
Blood glucose ≥ 110	24	39	5	92	82,8	70,2	38,1	94,8
In series (both +)	20	19	9	112	69,0	85,5	51,3	92,6
In parallel (at least one +)	26	53	3	78	89,7	59,5	32,9	96,3

The table confirms the general rule: the serial strategy gains specificity and PPV, the parallel strategy gains sensitivity and NPV. The methodological lesson is that multiplying likelihood ratios is correct only for conditionally independent tests. For correlated tests, the performance of the combination must be measured directly, in patients who have had both tests, or estimated with a multivariable model, for example the logistic regression from Chapter 9.

Case study: clinical rules that combine several measurements [110,123]

Two studies by the hepatology group in Timișoara, with R. Lupușoru from our discipline, show how several measurements are combined into a decision. Barbulescu and colleagues extended and validated the Baveno VI criteria for ruling out high-risk esophageal varices in 839 patients (derivation cohort 728, validation cohort 111). The Baveno VI criteria had an accuracy of 86.2% in the first cohort. The authors derived a rule-in rule (platelets below 150,000/mm³ and liver stiffness above 35.3 kPa) and a rule-out rule (platelets above 150,000/mm³ and liver stiffness below 19.6 kPa), and for the gray zone they proposed an additional albumin cut-off, below 3.4 g/dl. In the validation cohort, the rule-out cut-offs had an accuracy of 100%. The "AND" rule in each criterion is a serial combination; leaving a gray zone acknowledges that, between the two cut-offs, the post-test probability remains too close to the middle to decide without endoscopy. Fofiu and colleagues approached the same problem differently: in 124 analyzed patients with compensated cirrhosis, they combined spleen stiffness (cut-off 13.2 kPa, AUROC 0.84), liver stiffness (12.1 kPa, AUROC 0.86) and spleen size (12.9 cm, AUROC 0.71) into a score obtained by regression: 0.053 x SS + 0.054 x LS + 0.059 x SSZ - 1.84, with a cut-off > 0.34 and an AUROC of 0.93, superior to each factor taken separately. The regression model estimates the weights while taking into account the correlation between the measurements, which is exactly the problem of conditional independence. Both studies must be read with a caveat: an accuracy of 100% in a validation cohort of 111 patients has a wide confidence interval, and a score derived on 124 patients needs external validation before routine use.

11.6 Differential Diagnosis with Several Diseases

11.6.1 The Bayes Table with Three Diagnoses

In practice, the physician does not ask only "does the patient have disease X or not?", but "which of diseases X, Y, Z best explains what I see?". Bayes' theorem extends directly to several hypotheses, provided they are mutually exclusive (the patient has only one) and exhaustive (one of them is true). For each disease Di:

$$p(D_i/S) = \frac{p(S/D_i) \times p(D_i)}{\sum[p(S/D_j) \times p(D_j)]}$$

The numerator is calculated for each disease, and the denominator is the sum of the numerators. In practice, the posterior probability of each disease is the share of its numerator in the total.

Let us take an example from family medicine, with fictitious but plausible figures. A 68-year-old man, a former smoker, comes in for progressive exertional dyspnea over the last few months. For simplicity, we assume that three diagnoses cover almost all cases of this type in that practice: heart failure (HF), chronic obstructive pulmonary disease (COPD) and anemia. From the practice's experience, in patients with this profile, the prior probabilities are 0.35 for HF, 0.45 for COPD and 0.20 for anemia. The physician notes three signs: bilateral leg edema (E), wheezing on auscultation (W) and skin pallor (P). The frequency of each sign in each disease is given in Table 11.5.

Table 11.5. Prior probabilities and frequency of the clinical signs in each diagnosis (fictitious data).

Diagnosis	**p(D)**	$p(E/D)$	$p(W/D)$	$p(P/D)$
Heart failure	0,35	0,60	0,20	0,15
COPD	0,45	0,15	0,60	0,10
Anemia	0,20	0,10	0,05	0,70

The patient has edema (E+), no wheezing (W-) and is not pale (P-). We assume that, within each disease, the signs occur independently of one another (the assumption of conditional independence, called the "naive Bayes" model in computer science). The likelihood of the observed combination in each disease is the product of the probabilities: for a sign that is present we use p(S / D), for an absent sign we use 1 - p(S / D).

For HF: 0.60 x (1 - 0.20) x (1 - 0.15) = 0.60 x 0.80 x 0.85 = 0.408

For COPD: 0.15 x (1 - 0.60) x (1 - 0.10) = 0.15 x 0.40 x 0.90 = 0.054

For anemia: 0.10 x (1 - 0.05) x (1 - 0.70) = 0.10 x 0.95 x 0.30 = 0.0285

We multiply each likelihood by the prior probability:

HF: 0.35 x 0.408 = 0.1428

COPD: 0.45 x 0.054 = 0.0243

Anemia: 0.20 x 0.0285 = 0.0057

The sum is 0.1728. The posterior probabilities are:

$$p(\text{HF/data}) = \frac{0.1428}{0.1728} = 0.826$$

$$p(\text{COPD/data}) = \frac{0.0243}{0.1728} = 0.141$$

$$p(\text{ anemia }/\text{data}) = \frac{0.0057}{0.1728} = 0.033$$

Table 11.6. Calculation of the posterior probabilities for the patient with dyspnea (E+, W-, P-).

Diagnosis	Prior	Likelihood	Product	Posterior
Heart failure	0,35	0,408	0,1428	82,6%
COPD	0,45	0,054	0,0243	14,1%
Anemia	0,20	0,0285	0,0057	3,3%
Total	1,00		0,1728	100%

The ranking of the diagnoses has been reversed. Before the examination, COPD was the most probable diagnosis (45%); after the examination, heart failure reaches 82.6%. The result does not mean that the diagnosis is established: a probability of 82.6% justifies a targeted test (for example NT-proBNP and echocardiography), not definitive treatment without confirmation. The calculation can also be done step by step, sign by sign, using the posterior from the previous step as the prior for the next step. After edema alone, the probabilities would be 70.6% / 22.7% / 6.7%; after adding the absence of wheezing, 78.5% / 12.6% / 8.9%; after the absence of pallor, 82.6% / 14.1% / 3.3%. The final result is the same regardless of the order in which we add the signs, provided that the independence assumption holds.

11.6.2 Limitations of the "Naive" Model

The model in the example above has three weaknesses that you should keep in mind. The first is exhaustiveness: if the patient in fact has a fourth disease (for example pulmonary fibrosis or valvular heart disease), no figure in the table will signal it; the probabilities will be redistributed among the three diagnoses listed. The second is the independence of the signs: edema and orthopnea occur together in heart failure much more often than the product of their frequencies would suggest, and multiplying them as if they were independent exaggerates the evidence. The third is the quality of the input numbers: the prior probabilities and the frequencies of the signs are estimates, often from studies in different populations.

Nevertheless, naive Bayes models were among the first clinical decision support systems and remain surprisingly robust for ranking diagnoses, even if the absolute probabilities are not exact [124]. The lectures of our discipline present them, alongside pattern recognition, as statistical classification methods. The basic idea, updating probabilities by multiplying the prior by the likelihood, also underlies many modern classifiers, including some used in medical artificial intelligence.

11.7 Decision Thresholds: To Test, to Treat

The post-test probability is not an end in itself. It serves a decision: do we treat, investigate further or do nothing? To link probability to decision we need the treatment threshold, that is, the probability above which the expected benefit of treatment exceeds its risk.

Let B denote the net benefit of treating an ill person (for example the complications avoided) and R the net risk of treating a person who is not ill (adverse effects, costs, anxiety). The treatment threshold is

$$p_{\text{treatment}} = \frac{R}{R + B}$$

If needlessly treating a healthy patient "costs" one unit, and failing to treat an ill person "costs" four units (B = 4), the threshold is 1 / (1 + 4) = 0.20. Above 20%, treatment is worthwhile; below 20%, it is not. For a very safe and very effective treatment (for example a well-tolerated

oral antibiotic for an infection with serious complications), the threshold is low. For a risky treatment (chemotherapy, anticoagulation in an older patient at risk of bleeding), the threshold is high.

Once we have a test, two additional thresholds appear. The (lower) testing threshold is the pre-test probability below which not even a positive result would bring the patient above the treatment threshold; below it, the test is useless and the patient is not treated. The (upper) test-treatment threshold is the pre-test probability above which not even a negative result would bring the patient below the treatment threshold; above it, the patient is treated without a test. Between the two thresholds, the test can change the decision and is worth doing [120].

With the odds form, the calculation is direct. Let the treatment threshold be 20% (odds 0.25) and take a test with LR+ = 6 and LR- = 0.2.

Testing threshold: pre-test odds = 0.25 / 6 = 0.0417, so the probability is 0.0417 / 1.0417 = 4.0%.

Upper threshold: pre-test odds = 0.25 / 0.2 = 1.25, so the probability is 1.25 / 2.25 = 55.6%.

Patients with a pre-test probability between 4.0% and 55.6% benefit from the test. Below 4.0%, they are neither tested nor treated; above 55.6%, they are treated directly. The model is simplified (it ignores the risk and cost of the test itself), but it illustrates the main idea: the decision to order a test depends on the pre-test probability, on the performance of the test and on the consequences of errors. A better test (higher LR+, lower LR-) widens the interval in which testing is useful.

11.8 Screening in Family Medicine: The FINDRISC Example

The FINDRISC score (Finnish Diabetes Risk Score) was constructed to identify, by means of a simple questionnaire, people at high risk of type 2 diabetes [125]. It takes values between 0 and 26 and includes age, BMI, waist circumference, physical activity, daily consumption of vegetables and fruit, antihypertensive treatment, previously elevated blood glucose and diabetes in first- or second-degree relatives. It is exactly the kind of tool that a family physician can use in a patient who comes in for something else.

In the mg2026 dataset, diabetes was confirmed by OGTT (the reference standard, according to the ADA criteria [80]), and at the FINDRISC cut-off ≥ 12 we have LR+ 3.01 and LR- 0.32. To see what these figures mean in different patients, we apply the odds form to several pre-test probabilities.

Table 11.7. Post-test probability of diabetes after FINDRISC at the cut-off ≥ 12 (LR+ 3.01; LR- 0.32), for different pre-test probabilities. The likelihood ratios come from the mg2026 dataset (synthetic data); the pre-test probabilities are scenarios.

Pre-test probability	Possible context	Pre-test odds	Post-test after FINDRISC ≥ 12	Post-test after FINDRISC < 12
2%	young adult, no risk factors	0,020	5,8%	0,7%
5%	general adult population	0,053	13,7%	1,7%
10%	middle-aged adult, overweight	0,111	25,1%	3,5%
18,1%	the mg2026 sample (cardiometabolic risk)	0,221	40,0%	6,7%
30%	obese, known prediabetes	0,429	56,3%	12,1%

Pre-test probability	Possible context	Pre-test odds	Post-test after FINDRISC ≥ 12	Post-test after FINDRISC < 12
50%	patient with suggestive symptoms	1,000	75,1%	24,4%

The table tells us several important things. A positive FINDRISC does not "make the diagnosis": even in the high-risk sample, only 4 out of 10 positive patients have diabetes. A negative FINDRISC is more useful for reassuring than a positive one is for raising the alarm, at least at low and moderate pre-test probabilities: below 10% pre-test, a score below 12 brings the probability below 3.5%. At high pre-test probabilities (50%), a negative score still leaves a probability of 24.4%, much too high to forgo the laboratory test. The role of FINDRISC in the family practice is therefore triage: it decides who deserves a fasting plasma glucose or HbA1c test, not who has diabetes. The post-test column for the positive score is, incidentally, identical to the table of PPV as a function of prevalence that you will meet again in Chapter 12: the PPV at prevalence p is exactly the post-test probability for a patient with pre-test probability p.

When the patient already has a FINDRISC result and blood glucose is also measured, the reasoning continues: the post-test probability after FINDRISC becomes the pre-test probability for the blood glucose test. We saw in Section 11.5 that, in this particular case, the independence assumption does not hold, so simple multiplication of the likelihood ratios overestimates the final probability. In practice, guidelines in any case require confirmation of the diagnosis of diabetes by laboratory tests, not by probabilistic calculation [80].

Case study: predictive values depend on prevalence and on the spectrum of patients [75,126,127]

Three studies by our colleagues show the same lesson from different angles. Albai and colleagues (with M. Frandeş) validated the ISARIC-4C score in 159 patients with type 2 diabetes hospitalized for SARS-CoV-2 infection, with a mortality of 25.15%. The area under the ROC curve was 0.875 (95% CI 0.820-0.930), and at the chosen cut-off the score correctly classified 77.36% of the cohort, with Se 87.80%, Sp 73.72%, PPV 53.73% and NPV 94.56%. If we use Bayes' theorem to recalculate the PPV at a prevalence of 25.15%, we obtain about 52.9%, close to the reported value; the same PPV would fall greatly in a population with a mortality of 10%. A note for critical reading: the confidence intervals reported for Se and Sp (for example 87.66-87.93) are unusually narrow for 159 patients, of whom only about 40 died; a Wilson interval for a proportion of 88% calculated on 40 cases would have a width of more than 20 percentage points. Sipos and colleagues (with R. Moleriu) reported for IL-6 in the diagnosis of neonatal sepsis an optimal cut-off of 181 pg/ml, with Se 80.1%, Sp 85.7%, PPV 84.6% and NPV 81.8%. The abstract does not give the number of newborns, but we can do an inverse Bayesian exercise: from Se and Sp, $LR^+ = \frac{0.801}{0.143} = 5.6$; the post-test odds corresponding to a PPV of 84.6% are 0.846 / 0.154 = 5.49; dividing, the pre-test odds are about 0.98, that is, a prevalence of sepsis of about 49.5% in the sample analyzed. Among all newborns tested in a maternity unit, the prevalence of sepsis is much lower, so in practice the PPV would be lower than 84.6%. Pavel and colleagues (with A. Tudor) evaluated H-FABP in 83 patients with suspected acute myocardial infarction, separately according to the time since the onset of pain. At less than 4 hours, sensitivity and specificity were 89.67 and 95.65, which corresponds to an LR+ of about 20.6; beyond 4 hours, sensitivity was 86.73 and specificity only 49.84, that is, an LR+ of about 1.73, with almost no value for confirmation. The same biomarker, in the same department, performs completely differently depending on when the patients are tested. Therefore, LRs are not universal constants: they apply to patients similar to those in the study that estimated them.

11.9 Bayesian Inference Versus Frequentist Inference

So far we have applied Bayes' theorem to the individual patient. The same logic can also be applied to the parameters of a population: a mean, a proportion, a difference between treatments. This is Bayesian inference, an alternative to the frequentist hypothesis tests and confidence intervals of Chapters 4-10.

11.9.1 Prior, Likelihood, Posterior

In frequentist inference, the parameter (for example the prevalence of diabetes in the population, π) is an unknown constant, and probability describes only the variability of samples. In Bayesian inference, uncertainty about the parameter is expressed directly through a probability distribution. Before the data we have the prior distribution, which summarizes what we know or believe about the parameter. The data contribute through the likelihood (how probable the observed data are for each possible value of the parameter). Their combination gives the posterior distribution:

posterior $\propto$ likelihood x prior

The symbol $\propto$ means "proportional to"; a normalizing constant makes the area under the posterior distribution equal to 1. The formula is Bayes' theorem applied to an entire distribution rather than to a single probability.

For a proportion, the calculation is especially simple if we use a Beta distribution as the prior. A Beta(α; β) distribution can be thought of as the equivalent of α - 1 "successes" and β - 1 "failures" observed beforehand. If we observe k cases out of n, the posterior distribution is Beta(α + k; β + n - k). For diabetes in mg2026, k = 29 and n = 160.

With a uniform prior, Beta(1; 1), which states that any prevalence between 0 and 100% is equally plausible, the posterior is Beta(30; 132). Its mean is 30/162 = 18.5%, and the interval that contains 95% of the posterior probability, called the credible interval, is 12.9-24.8%. With an informative prior, Beta(10; 40), which expresses the prior belief that the prevalence is around 20% (with 95% of the probability between 10.2% and 32.0%), the posterior is Beta(39; 171), with a mean of 18.6% and a credible interval of 13.6-24.1%. Figure 11.4 shows the two situations.

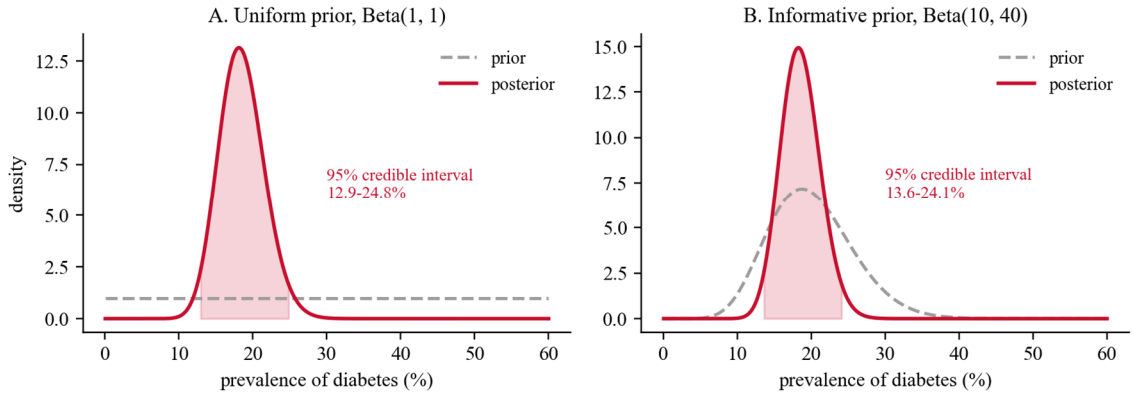


Figure 11.4. Prior distribution (dashed line) and posterior distribution (solid line) for the prevalence of OGTT-confirmed diabetes, with 29 cases out of 160. A: uniform prior, 95% credible interval 12.9-24.8%. B: informative prior centered on 20%, credible interval 13.6-24.1%. Synthetic data (mg2026 dataset).

Two remarks. With 160 patients, the data dominate: the two posterior distributions are very similar, even though the priors differ considerably. With 20 patients, the prior would have mattered much more. Second, the credible interval with a uniform prior (12.9-24.8%) virtually coincides with the Wilson confidence interval calculated in Chapter 4 (12.9-24.8%). For large samples and weakly informative priors, Bayesian and frequentist methods give very similar numerical results; what differs is the interpretation.

11.9.2 The Credible Interval versus the Confidence Interval

The frequentist 95% confidence interval has an awkward interpretation: if we repeated the study a very large number of times and calculated the interval each time, 95% of these intervals would contain the true value. We cannot say that "the true value lies in the interval 12.9-24.8% with probability 95%", because, in the frequentist sense, the parameter is not a random variable. This confusion is one of the most common misinterpretations in statistics [70].

The Bayesian credible interval has exactly the interpretation the clinician intuitively wants: given the prior and the data, the probability that the prevalence lies between 12.9% and 24.8% is 95%. The price paid is the choice of prior, which introduces an element of judgment that must be justified and reported transparently. A prior chosen to obtain the desired result is just as dishonest as a frequentist test repeated until it yields $p < 0.05$.

11.9.3 The Bayes Factor

The Bayesian counterpart of the hypothesis test is the Bayes factor. It compares how well two competing hypotheses predict the observed data, usually H1 (there is an effect) and H0 (there is no effect):

$$\mathrm{BF}_{10} = \frac{p(\mathrm{data}/H_1)}{p(\mathrm{data}/H_0)}$$

A BF10 of 10 means that the data are 10 times more probable under H1 than under H0. A BF10 of 0.1 means that the data are 10 times more probable under H0; equivalently, $\mathrm{BF}_{01} = \frac{1}{\mathrm{BF}_{10}} = 10$. Unlike the p value, the Bayes factor can provide evidence in favor of the null hypothesis, not only against it. It is, in fact, a likelihood ratio applied to hypotheses:

posterior odds of H1 = prior odds of H1 x BF10

You will recognize the formula from diagnosis: the hypotheses play the role of the diseases, and the data play the role of the test result.

JASP, the program used in the practical sessions, offers a Bayesian version of many frequentist tests (in each test's menu, under the Bayesian section) and uses a widely adopted interpretation scale [47,49].

Table 11.8. Commonly used interpretation scale for the Bayes factor BF10 (including in the JASP materials).

BF10	Evidence in favor of H1	BF10	Evidence in favor of H0
1-3	anecdotal (weak)	1/3-1	anecdotal (weak)
3-10	moderate	1/10-1/3	moderate
10-30	strong	1/30-1/10	strong
30-100	very strong	1/100-1/30	very strong
over 100	extreme	below 1/100	extreme

On the mg2026 dataset, several analyses from earlier chapters, redone in a Bayesian framework (with the JASP default priors, for example the Cauchy distribution with a width of 0.707 for the t-test), give the following results:

- SBP at enrollment, women versus men (independent-samples t-test): BF10 = 0.18, that is, BF01 = 5.7. The data are about 5.7 times more probable if there is no difference between the sexes: moderate evidence for H0. The frequentist test (p = 0.81) only told us that we cannot reject H0; the Bayes factor tells us, in addition, that the data actively support the absence of a difference.
- Blood glucose at 0 versus 6 months (paired t-test): BF10 on the order of 10^18, extreme evidence for a change.

- BMI-SBP correlation (r = 0.405, n = 160): BF10 of approximately 116,000, extreme evidence.
- HDL-triglyceride correlation (r = 0.076): BF10 = 0.16, that is, BF01 = 6.4, moderate evidence for the absence of a correlation.
- Obesity and diabetes (2 x 2 contingency table, joint multinomial sampling, default prior): BF10 of approximately 85, very strong evidence for an association.

Bayesian analysis does not solve design problems (a poorly designed study remains poorly designed), nor does it replace clinical judgment about effect size. It does, however, have two teaching advantages: it answers directly the question "how strongly do the data support one hypothesis over another?", and it forces us to make prior information explicit. In the medical literature, Bayesian methods appear more and more often in adaptive clinical trials and in meta-analyses, and today's student will read them as a physician [42,63].

11.10 Evidence-Based Medicine

Evidence-based medicine has been defined as the conscientious, explicit and judicious use of the best available evidence in making decisions about the care of individual patients, by integrating clinical expertise with evidence from systematic research [15]. The definition has three components, and all three matter: research evidence, the clinician's experience, and the patient's values and preferences. Evidence-based medicine does not mean mechanically applied guidelines, nor statistics in place of clinical practice.

11.10.1 The Five Steps

The practice of evidence-based medicine is usually described in five steps, which are easy to remember in English as the five "A"s.

Step 1, formulating the question (ask). A well-constructed clinical question has the PICO structure [24]: P (the patient or population), I (the intervention, exposure or diagnostic test), C (the comparator) and O (the outcome). For example: "In adults at cardiometabolic risk in family medicine (P), does the use of a self-monitoring mobile app added to counseling (I), compared with standard counseling (C), increase the proportion of those who lose at least 5% of their body weight at 6 months (O)?". For a diagnostic question: "In previously untested overweight adults (P), what are the sensitivity and specificity of a FINDRISC score ≥ 12 (I) for type 2 diabetes (O), compared with the OGTT as the reference standard (C)?".

Step 2, searching for evidence (acquire). The search is run in bibliographic databases such as PubMed, in collections of systematic reviews and in guidelines, using the PICO terms. Efficient searching was discussed in the medical informatics volume of the first semester [1].

Step 3, critical appraisal (appraise). Is the study valid (design, randomization, blinding, losses to follow-up, reference standard)? What are the results (effect size, confidence interval)? Do they apply to my patient? For diagnostic studies, critical appraisal follows the STARD checklist, which we discuss in Chapter 12.

Step 4, application (apply). The evidence is integrated with the patient's specific situation: comorbidities, preferences, access to treatment, costs. This is where the Bayesian reasoning of this chapter comes in, because my patient's pre-test probability may differ from the prevalence in the study.

Step 5, evaluating the outcome (assess). Physicians evaluate their own practice: did the decision lead to the desired outcome? What can be done better next time?

11.10.2 The Hierarchy of Evidence and the GRADE System

Not all studies provide equally solid evidence. For questions about the efficacy of treatments, the classic hierarchy places expert opinion and case series at the bottom, followed by case-control studies, cohort studies, randomized clinical trials and, at the top, systematic reviews with meta-analysis of randomized trials. For diagnostic questions, the top is occupied by cross-sectional studies in consecutive patients, all of whom receive both the test under evaluation and the reference standard; case-control designs (known patients versus healthy volunteers) systematically overestimate accuracy. For prognostic questions, the preferred design is the prospective cohort.

The pyramid is useful as a first orientation, but it oversimplifies. A small randomized trial with many losses to follow-up may provide weaker evidence than a large, well-conducted cohort. This is why the current approach, GRADE (Grading of Recommendations, Assessment, Development and Evaluation), rates the certainty of the evidence separately for each outcome rather than for each study type [128]. Evidence from randomized trials starts at high certainty, and evidence from observational studies at low certainty. Certainty can be rated down for risk of bias, inconsistency across studies, indirectness (the population or the outcome differs from those we are interested in), imprecision (wide confidence intervals) and publication bias. It can be rated up, especially for observational studies, when the effect is very large or there is a dose-response gradient. The result is one of four categories: high, moderate, low or very low certainty.

11.10.3 Systematic Reviews and Meta-Analyses

A systematic review answers a clear question (usually of the PICO type), with a protocol established in advance, an exhaustive search of several databases, explicit inclusion and exclusion criteria, an assessment of the quality of each study and a transparent synthesis. It differs from a narrative review, in which the author selects studies according to undeclared criteria. A meta-analysis is the statistical component of a systematic review: it combines the results of the studies into an overall estimate, weighting each study by its precision (as a rule, inversely proportional to its variance).

The result of a meta-analysis is presented in a forest plot, which you have already encountered in Chapter 10 for the measures of association in the mg2026 dataset. Each row is a study: a square for the point estimate (with its area proportional to the weight of the study) and a horizontal line for the confidence interval. The vertical line marks the absence of an effect (1 for RR and OR, 0 for differences). The diamond at the bottom is the pooled estimate, with its width equal to its confidence interval. If the diamond does not touch the vertical line, the pooled effect is statistically significant.

Heterogeneity describes how much the results of the studies differ from one another beyond what chance would explain. It is tested with Cochran's Q test and quantified by I^2, the proportion of the total variability due to real differences between studies. As a rough guide, an I^2 of around 25% indicates low heterogeneity, 50% moderate and 75% high. When heterogeneity is high, combining the studies into a single number must be done with caution, using a random-effects model (which allows each study to have its own true effect), and explanations must be sought: different populations, different doses, different designs.

For reporting systematic reviews, the PRISMA 2020 guideline is used [129], which includes a checklist and a flow diagram of study selection (how many records were found, how many remained after removing duplicates, how many were excluded on reading the title and abstract, how many on reading the full text, and for what reasons). In meta-analyses of diagnostic accuracy, pairs of sensitivity and specificity are combined, and the result is usually presented as a summary ROC curve, a topic beyond the level of this course.

Unlike a systematic review, a narrative review does not declare its selection criteria and does not produce a pooled estimate. An example from our colleagues' research is the review by Suciu and colleagues on heart failure with preserved ejection fraction, which provides a narrative synthesis of studies published between 2009 and 2025 [130]. Such works are valuable for understanding a field, but they rank lower in the hierarchy of evidence than a systematic review with meta-analysis.

11.11 Cognitive Errors in Clinical Reasoning

Clinical reasoning uses two systems. The first is fast, intuitive and based on pattern recognition ("I've seen this before"). The second is slow, analytical and explicit, like the Bayesian calculation in this chapter. Experienced physicians rely mainly on the first system and, most of the time, successfully. Intuition, however, has systematic errors, called cognitive biases, which can be recognized and corrected [120].

Base-rate neglect is the central error of this chapter: we judge the probability of a disease by how well the clinical picture fits, ignoring how rare the disease is. A young man with palpitations, sweating and headache "fits" pheochromocytoma, but the disease is so rare that, even with a suggestive picture, the post-test probability remains low. The antidote is always to start with the question "how common is this disease in patients like this one?".

Anchoring means fixating on the first impression and adjusting insufficiently when new information appears. A patient sent from triage with the label "panic attack" will be viewed through this lens, and a slightly elevated troponin may be downplayed. In Bayesian terms, anchoring means failing to update the probability sufficiently after a result with a high LR. Closely related is premature closure: stopping the search after the first plausible diagnosis.

Availability means overestimating the probability of diseases that come easily to mind: last week's dramatic case, the disease recently discussed at a conference, the missed diagnosis that led to a complaint against us. After a missed case of pulmonary embolism, many physicians order CT angiography in patients with a very low pre-test probability. The real probability has not changed; only the ease with which the disease comes to mind has.

Other common errors are confirmation bias (we seek only the information that supports our hypothesis), the framing effect (the same information, presented as "90% survival" or as "10% mortality", leads to different decisions) and overconfidence. No list eliminates these errors, but two habits reduce them considerably: explicitly estimating the pre-test probability before ordering a test, and asking "what else could it be?" before closing the diagnosis.

11.12 Communicating Risk to the Patient

Patients, like physicians, understand natural frequencies much better than conditional percentages. "Out of 100 people like you who have a FINDRISC score of at least 12, about 40 have diabetes and 60 do not" is a sentence anyone can understand. "The positive predictive value of the test is 40%" is not.

The second rule is to communicate absolute risk, not just relative risk. In the simulated mg2026 study, the proportion of patients who lost at least 5% of their weight was 29.6% in arm C and 11.3% in arm A. The relative risk is 2.62 ("2.6 times as many"), but the absolute difference is 18.3 percentage points, and the number needed to treat (NNT) is 5.5: about 6 patients must be enrolled in program C instead of A for one additional patient to reach the target. For a rare adverse effect, the difference between relative and absolute is even more important: a drug that "doubles the risk" of thrombosis may increase the absolute risk from 2 to 4 per 10,000 patients per year. Both statements are true, but only the second allows the patient to weigh the decision.

The third rule is to use the same denominator for all options ("out of 1000 people") and, whenever possible, a visual aid: a grid of 100 figures, some of them colored. Communicating uncertainty is also part of communicating risk. Saying "the test came back positive, which means that the probability that you have the disease has risen from about 1 in 5 to 2 in 5; we need a confirmatory test" is more honest and, in the long run, more reassuring than a premature verdict.

11.13 Worked Problems

Worked Problem 11.1. The Disease-Symptom Table and Bayes' Theorem

In an emergency department, 400 patients with chest pain were analyzed retrospectively. In 60 of them, an acute coronary syndrome was confirmed (B+). ST-segment depression on the initial ECG (S+) was present in 36 of the patients with coronary syndrome and in 34 of the others (fictitious data). Calculate p(B+), p(S+ / B+), p(S+ / B-), p(S+), p(B+ / S+) and p(B+ / S-). Check p(B+ / S+) with Bayes' theorem.

Solution. We build the table using the notation from the lecture. In row B+: a = 36, b = 60 - 36 = 24. In row B-: c = 34, d = 340 - 34 = 306.

	S+	S-	Total
B+	36	24	60
B-	34	306	340
Total	70	330	400

Unconditional probabilities: $p(B^+) = \frac{60}{400} = 0.15$; $p(S^+) = \frac{70}{400} = 0.175$.

Probabilities conditional on disease: $p(S^+/B^+) = \frac{36}{60} = 0.60$; $p(S^+/B^-) = \frac{34}{340} = 0.10$.

Probabilities conditional on the symptom: $p(B^+/S^+) = \frac{36}{70} = 0.514$; $p(B^+/S^-) = \frac{24}{330} = 0.073$.

Check with Bayes, without cell a: $p(B^+/S^+) = \frac{0.60 \times 0.15}{0.60 \times 0.15 + 0.10 \times 0.85} = \frac{0.090}{0.090 + 0.085} = \frac{0.090}{0.175} = 0.514$.

Clinical interpretation. ST depression raises the probability of coronary syndrome from 15% to 51.4%, but does not make it certain. Its absence lowers it to 7.3%, a value still too high to discharge the patient without further investigations (serial troponins). The likelihood ratio of the sign is $\text{LR}^+ = \frac{0.60}{0.10} = 6$, and $\text{LR}^- = \frac{0.40}{0.90} = 0.44$: the sign is useful for confirmation, much less so for exclusion.

Worked Problem 11.2. Natural Frequencies for a Rapid Test

A rapid test for anti-HCV antibodies has a sensitivity of 98% and a specificity of 97%. It is proposed for use (a) in the general population, where the prevalence of infection is 1%, and (b) in a center for people who inject drugs, where the prevalence is 20% (illustrative figures). Calculate the PPV and NPV in the two situations using natural frequencies.

Solution (a). We take 10,000 people. A prevalence of 1% gives 100 infected and 9900 uninfected people. Of the 100, the test is positive in 98 (Se 98%) and negative in 2. Of the 9900, a specificity of 97% gives 9603 negatives and 297 false positives. Total positives: 98 + 297 = 395. PPV $= \frac{98}{395} = 24.8\%$. Total negatives: 2 + 9603 = 9605. NPV $= \frac{9603}{9605} = 99.98\%$.

Solution (b). Out of 10,000 people, 2000 are infected and 8000 are not. True positives: 0.98 x 2000 = 1960; false negatives: 40. False positives: 0.03 x 8000 = 240; true negatives: 7760. PPV = $\frac{1960}{1960+240} = \frac{1960}{2200} = 89.1\%$. NPV = $\frac{7760}{7760+40} = 99.49\%$.

Interpretation. The same test, with an excellent specificity of 97%, produces three false positives for every true positive in the general population: about 3 out of 4 positive results are false. In the high-risk population, 9 out of 10 positive results are true. This is why a positive screening test in the general population must always be confirmed with a different test (for example, detection of viral RNA), and communication with the patient must explain this before the test is performed.

Worked Problem 11.3. Same Study, Different Prevalences: HPV Genotyping and p16/Ki-67 Dual Immunostaining

Moleriu and colleagues followed 184 patients with LSIL cytology, 64 of whom underwent biopsy [131]. Histology (the reference standard) showed high-grade lesions in 18 of 64 (28.13%). HPV genotyping had Se 94.44% (17/18) and Sp 58.7% (27/46); p16/Ki-67 dual immunostaining had Se 66.67% (12/18) and Sp 84.78% (39/46). Calculate LR+ and LR- for each test, then the PPV and NPV at the study prevalence and at a prevalence of 10%.

Solution. We reconstruct the 2 x 2 tables. For HPV: a = 17, b = 1, c = 46 - 27 = 19, d = 27. For p16/Ki-67: a = 12, b = 6, c = 46 - 39 = 7, d = 39. We note that 17 + 19 = 36 patients were HPV positive (56.25% of 64), and 12 + 7 = 19 were p16/Ki-67 positive (29.69%), which matches the proportions in the abstract.

The likelihood ratios:

HPV: $LR^+ = \frac{0.9444}{1-0.587} = \frac{0.9444}{0.413} = 2.29$; $LR^- = \frac{0.0556}{0.587} = 0.095$.

p16/Ki-67: $LR^+ = \frac{0.6667}{0.1522} = 4.38$; $LR^- = \frac{0.3333}{0.8478} = 0.393$.

At the study prevalence (28.13%, odds 0.391): for HPV, PPV = $\frac{17}{36} = 47.2\%$ and NPV = $\frac{27}{28} = 96.4\%$; for p16/Ki-67, PPV = $\frac{12}{19} = 63.2\%$ and NPV = $\frac{39}{45} = 86.7\%$. The same values are obtained with the odds form (for example, 0.391 x 2.29 = 0.896, and 0.896/1.896 = 47.2%).

At a prevalence of 10% (odds 0.111): a positive HPV result gives 0.111 x 2.29 = 0.254, that is, a PPV of 20.3%; a negative HPV result gives 0.111 x 0.095 = 0.0106, that is, a post-test probability of 1.0% (NPV 99.0%). For p16/Ki-67: 0.111 x 4.38 = 0.487, PPV 32.7%; 0.111 x 0.393 = 0.0437, post-test probability 4.2% (NPV 95.8%).

Interpretation. The two tests have opposite profiles. HPV genotyping, highly sensitive and not very specific, is good for exclusion: a negative result greatly lowers the probability of a high-grade lesion. p16/Ki-67 immunostaining, which is more specific, is better for confirmation. Combining them (HPV as the first test, p16/Ki-67 in those who test positive) would be a logical serial strategy, but the performance of the combination cannot be calculated from the abstract, because we do not know how many patients were positive on both tests.

Worked Problem 11.4. Sequential Testing with Likelihood Ratios

Golu and colleagues studied 2D-SWE elastography in parathyroid lesions, in 22 patients with hyperparathyroidism and 43 healthy controls [132]. At a cutoff of 12.5 kPa, sensitivity was 93% and specificity 86% (AUC 0.949). (a) Calculate LR+ and LR-. (b) What is the post-test probability after a positive result, at the study prevalence and at a pre-test probability of 5%? (c) In the patient with a pre-test probability of 5% and a positive elastography result, a second,

conditionally independent test is performed, with LR+ = 8.5 (hypothetical value). What is the final probability?

Solution. (a) $\mathrm{LR}^{+} = \frac{0.93}{1-0.86} = \frac{0.93}{0.14} = 6.64$. $\mathrm{LR}^{-} = \frac{0.07}{0.86} = 0.081$.

(b) The study prevalence is 22/65 = 33.8%, with $\frac{\text{odds } 0.338}{0.662} = 0.512$. The post-test odds are 0.512 x 6.64 = 3.40, and the probability is 3.40/4.40 = 77.3%. At a pre-test probability of 5%, the odds are 0.0526; the post-test odds are 0.0526 x 6.64 = 0.350; the probability is 0.350/1.350 = 25.9%.

(c) The probability of 25.9% becomes the new pre-test probability, with odds of 0.350. After the second positive test: 0.350 x 8.5 = 2.97, that is, a probability of 2.97/3.97 = 74.8%. Equivalently, in a single step: 0.0526 x 6.64 x 8.5 = 2.97.

Interpretation. The study prevalence of 33.8% is artificial, a result of the case-control design (22 patients and 43 selected controls). The predictive value of 77.3% does not carry over to practice. Moreover, the case-control design with healthy controls tends to overestimate sensitivity and specificity (spectrum bias, discussed in Chapter 12), so the real LRs may be more modest. The sequential calculation is valid only if the two tests are conditionally independent.

Worked Problem 11.5. Differential Diagnosis of Chest Pain in the Family Practice

A 55-year-old patient presents to the family physician with intermittent chest pain. The physician considers three diagnoses, assumed to be mutually exclusive and exhaustive, with the following prior probabilities (fictitious data, but plausible for primary care): musculoskeletal pain 0.50; gastroesophageal reflux disease (GERD) 0.30; stable angina pectoris 0.20. Three elements of the history and examination have the frequencies shown in the table.

Diagnosis	**p(D)**	$p(\text{ pain on exertion } / D)$	$p(\text{ pain reproduced on palpation } / D)$	$p(\text{ heartburn postprandial } / D)$
Musculoskeletal	0,50	0,20	0,70	0,05
GERD	0,30	0,10	0,10	0,70
Stable angina	0,20	0,80	0,15	0,15

The patient describes exertional pain, the pain is not reproduced by palpation, and he has no postprandial heartburn. Calculate the posterior probabilities, assuming conditional independence of the three elements.

Solution. The likelihood of the combination (exertion +, palpation -, heartburn -) in each diagnosis:

Musculoskeletal: 0.20 x (1 - 0.70) x (1 - 0.05) = 0.20 x 0.30 x 0.95 = 0.0570

GERD: 0.10 x (1 - 0.10) x (1 - 0.70) = 0.10 x 0.90 x 0.30 = 0.0270

Angina: 0.80 x (1 - 0.15) x (1 - 0.15) = 0.80 x 0.85 x 0.85 = 0.5780

The products with the prior probabilities: 0.50 x 0.0570 = 0.0285; 0.30 x 0.0270 = 0.0081; 0.20 x 0.5780 = 0.1156. The sum is 0.1522.

The posterior probabilities: musculoskeletal 0.0285/0.1522 = 18.7%; $\frac{\text{GERD } 0.0081}{0.1522} = 5.3\%$; angina 0.1156/0.1522 = 76.0%.

Interpretation. Angina, the least probable diagnosis before the history was taken (20%), becomes the most probable (76.0%). The result justifies referral for cardiological evaluation (exercise ECG or other functional testing) and the initiation of secondary prevention measures. The result does depend, however, on the assumption that the three elements are independent

within each disease and on the diagnostic list being complete; an acute coronary syndrome, for example, is not included and must be ruled out clinically before any calculation.

Worked Problem 11.6. The Testing Threshold and the Treatment Threshold

For acute pharyngitis in an adolescent, the physician considers that antibiotic treatment is worthwhile if the probability of streptococcal etiology exceeds 25%. In our scenario, a rapid antigen test has a sensitivity of 86% and a specificity of 95%. Between which pre-test probabilities is the test useful?

Solution. $\mathrm{LR}^{+} = \frac{0.86}{0.05} = 17.2$. $\mathrm{LR}^{-} = \frac{0.14}{0.95} = 0.147$. The treatment threshold of 25% corresponds to odds of 0.25/0.75 = 0.333.

The lower threshold (testing threshold): the pre-test odds at which a positive test reaches exactly the treatment threshold are 0.333 / 17.2 = 0.0194, that is, a probability of 0.0194/1.0194 = 1.9%.

The upper threshold (test-treatment threshold): the pre-test odds at which a negative test lowers the probability exactly to the treatment threshold are 0.333 / 0.147 = 2.27, that is, a probability of 2.27/3.27 = 69.4%.

Interpretation. For pre-test probabilities between 1.9% and 69.4%, the test result can change the decision, so the test is useful. Below 1.9%, even a positive test would not justify the antibiotic (and the probability that the positive result is false is high). Above 69.4%, even a negative test would leave the probability above 25%, so the test does not change the decision. The model ignores the cost and discomfort of the test and relies on a treatment threshold chosen by clinical judgment; real thresholds are set in guidelines.

Worked Problem 11.7. Bayesian Inference for a Proportion in a Small Sample

In a pilot study, 3 of 20 patients treated with a new drug developed an adverse skin reaction. (a) With a uniform prior Beta(1; 1), what is the posterior distribution, its mean and its 95% credible interval? (b) A previous study suggests a frequency of about 10%, summarized by a Beta(2; 18) prior (mean 10%, 95% prior interval from 1.3% to 26.0%). What changes?

Solution. (a) The posterior is Beta(1 + 3; 1 + 17) = Beta(4; 18). The mean is 4/22 = 18.2%. The 95% credible interval (from tables of the Beta distribution or from software) is 5.4-36.3%.

(b) The posterior is Beta(2 + 3; 18 + 17) = Beta(5; 35). The mean is 5/40 = 12.5%, and the 95% credible interval is 4.3-24.2%.

Interpretation. With only 20 patients, the prior matters: the estimate drops from 18.2% to 12.5%, and the interval narrows, because the informative prior effectively adds the information of another 18 patients observed beforehand. Compare this with the example in Section 11.9, where, with 160 patients, the two priors gave almost identical results. In a scientific report, both analyses should be presented, and the choice of prior must be justified, for example by reference to the previous study.

Worked Problem 11.8. From the Bayes Factor to the Probability of the Hypothesis

In the mg2026 dataset, the comparison of SBP between the sexes gives BF10 = 0.18, and the obesity-diabetes association gives a BF10 of approximately 85. Calculate the posterior probability of hypothesis H1 for each analysis, starting from (a) a prior probability of 50% and (b) one of 20%.

Solution. We use posterior odds = prior odds x BF10.

SBP and sex: (a) prior odds 1; posterior odds 0.18; $p(H_1/\text{data}) = \frac{0.18}{1.18} = 15.3\%$. (b) Prior odds 0.25; posterior odds 0.045; $p(H_1/\text{data}) = \frac{0.045}{1.045} = 4.3\%$.

Obesity and diabetes: (a) 1 x 85 = 85; $p(H_1/\text{data}) = \frac{85}{86} = 98.8\%$. (b) 0.25 x 85 = 21.25; $p(H_1/\text{data}) = \frac{21.25}{22.25} = 95.5\%$.

Interpretation. The Bayes factor describes the strength of the evidence in the data; the final probability of the hypothesis also depends on how plausible it was beforehand. The obesity-diabetes association is so well supported that it remains highly probable even with a skeptical prior. For SBP and sex, the data push the hypothesis of a difference toward improbability. Note the perfect analogy with diagnosis: BF10 is the likelihood ratio, and the prior probability is the pre-test probability.

11.14 Proposed Exercises

1. In a study of 500 patients, 60 have disease B. Symptom S is present in 45 of the diseased and in 88 of the non-diseased. Build the disease-symptom table and calculate p(B+), p(S+ / B+), p(S+ / B-) and p(B+ / S+). Check the last result with Bayes' theorem.
2. A screening mammogram has a sensitivity of 90% and a specificity of 93%. The prevalence of undiagnosed breast cancer in the screening population is 0.8%. Using 10,000 women and natural frequencies, calculate the PPV and NPV. How many women with a positive mammogram must undergo further investigation to find one cancer?
3. Convert the probabilities 0.40 and 0.05 into odds, and convert the odds 3 and 0.25 into probabilities.
4. Danila and colleagues (with F. Cioca) reported, for dNLR in patients with COVID-19 and diabetes, a cutoff of 2.685 with a sensitivity of 70.312%, a specificity of 65.978% and an AUC of 0.624, and for ALRI a sensitivity of 76.429% and a specificity of 69.541% (AUC 0.752) [105]. (a) Calculate LR+ and LR- for dNLR. (b) For a patient with a pre-test probability of 30%, what is the post-test probability after a dNLR above the cutoff and after one below the cutoff? (c) Comment on the number of decimal places reported for sensitivity and specificity.
5. Using the data from Worked Problem 11.4 (2D-SWE, Se 93%, Sp 86%), calculate the post-test probability after a negative result, for a pre-test probability of 33.8% and of 5%.
6. Test A has Se 0.90 and Sp 0.80; test B has Se 0.85 and Sp 0.95. Assuming conditional independence, calculate the sensitivity and specificity of the combination in series and in parallel. Which strategy would you choose for triage in the emergency room, and which for confirmation before a toxic treatment?
7. In an emergency department, a patient with acute headache has three possible diagnoses, with the following prior probabilities: migraine 0.70; tension-type headache 0.25; subarachnoid hemorrhage (SAH) 0.05. The frequencies of the clinical findings are: sudden "thunderclap" onset (0.05 / 0.02 / 0.80); neck stiffness (0.05 / 0.10 / 0.60); nausea (0.70 / 0.10 / 0.70), in the order migraine / tension-type / SAH. The patient has all three findings. Calculate the posterior probabilities, assuming conditional independence, and comment on the decision.
8. The treatment threshold for a disease is 10%. The available test has LR+ = 4 and LR- = 0.25. Calculate the testing threshold and the test-treatment threshold.

9. Ion and colleagues (with F. Cioca) compared clinical scores for predicting severe COVID-19 in patients with diabetes [106]. SOFA had an AUC of 0.771 on admission and 0.873 on day five; NEWS2 0.892 on admission and 0.729 on day five; APACHE II 0.698 on admission and 0.806 on day five. (a) Classify each value on the AUC scale from the lecture (Chapter 12). (b) Which score would you use on admission, and which after five days? (c) The study describes a logistic regression but reports hazard ratios (HR, for example SOFA on day five HR 3.07, 95% CI 2.29-4.12). What effect measure would have been expected from a logistic regression?
10. A Bayesian analysis gives BF10 = 0.16 for a correlation. (a) Express the result as BF01 and classify it on the scale in Table 11.8. (b) If the prior probability that the correlation exists was 50%, what is the posterior probability?
11. Formulate, in the PICO structure, a clinical question about the usefulness of HbA1c compared with the OGTT for diagnosing diabetes in the family practice, and identify the most appropriate type of study to answer it.

Common Pitfalls

- Confusing p(S / B) with p(B / S). The sensitivity of a test is not the probability that a patient with a positive test is diseased.
- Ignoring prevalence (base-rate neglect): a test with Se and Sp of 95% can have a PPV below 50% if the disease is rare.
- Using the PPV or NPV reported in a case-control study as if they applied in the family practice. The prevalence in such a study is fixed by the researcher.
- Multiplying likelihood ratios for correlated tests (for example, FINDRISC and blood glucose), which overestimates the final probability.
- Confusing odds with probability: odds of 0.25 correspond to a probability of 0.20, not 0.25.
- Interpreting the frequentist confidence interval as a credible interval ("95% probability that the parameter is here").
- Reading a small BF10 as an "inconclusive result". A BF10 of 0.1 is evidence for H0, not an absence of evidence.

Key Points

- Bayes' theorem: $p(B/S) = \frac{p(S/B) \times p(B)}{p(S)}$, with p(S) from the law of total probability.
- The odds form: post-test odds = pre-test odds x LR; $LR^{+} = \frac{Se}{1-Sp}$; $LR^{-} = \frac{1-Se}{Sp}$.
- The pre-test probability depends on the patient and the setting (family practice, outpatient clinic, hospital); the predictive values depend on it.
- Natural frequencies (1000 patients, a tree) are the safest way to solve a diagnostic problem and to explain a result to a patient.
- In series: Sp and PPV increase; in parallel: Se and NPV increase. The simple formulas assume conditional independence.
- Bayesian differential diagnosis: the product prior x likelihood for each disease, divided by the sum of the products.
- Bayesian inference: posterior $\propto$ likelihood x prior; the credible interval and the Bayes factor have direct interpretations, but they require a justified prior.
- Evidence-based medicine: five steps (PICO, search, critical appraisal, application, evaluation), certainty of evidence rated with GRADE, systematic reviews reported according to PRISMA 2020.

Self-Assessment Questions

1. (SC) In the disease-symptom table with the notation from the lecture (a, b in row B+; c, d in row B-), the conditional probability p(S+ / B-) is:
   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. (SC) A test with a sensitivity of 80% and a specificity of 90% is applied in a population with a disease prevalence of 10%. The positive predictive value is approximately:
   a) 80%
   b) 47%
   c) 90%
   d) 8%
   e) 72%
3. (SC) A pre-test probability of 0.20 corresponds to odds of:
   a) 0,20
   b) 0,25
   c) 0,80
   d) 4
   e) 0,16
4. (SC) A test result with a likelihood ratio equal to 1:
   a) confirms the disease
   b) rules out the disease
   c) doubles the pre-test probability
   d) does not change the probability of the disease
   e) has a sensitivity of 100%
5. (SC) A Bayesian analysis reports BF10 = 0.05. The correct interpretation is:
   a) strong evidence in favor of hypothesis H1
   b) p = 0.05, a borderline significant result
   c) strong evidence in favor of hypothesis H0 (BF01 = 20)
   d) an inconclusive result; nothing can be said
   e) the probability that H0 is true is 5%
6. (MC) Regarding serial testing (the second test is performed only in those who are positive on the first, and the patient is considered positive if both tests are positive):
   a) it increases the specificity of the combination compared with each test taken separately
   b) it increases the sensitivity of the combination compared with each test taken separately
   c) it increases the positive predictive value
   d) it is more economical, because the second test is applied to only some of the patients
   e) the formula Se = Se1 x Se2 is valid only if the tests are conditionally independent
7. (MC) Regarding likelihood ratios:

a) $\mathrm{LR}^{+} = \frac{\mathrm{Se}}{1-\mathrm{Sp}}$

b) $\mathrm{LR}^{-} = \frac{1-\mathrm{Se}}{\mathrm{Sp}}$

c) they depend strongly on the prevalence of the disease, just like the predictive values

d) post-test odds = pre-test odds x LR

e) an LR smaller than 1 increases the probability of the disease

8. (MC) Regarding evidence-based medicine:

a) it integrates research evidence with clinical expertise and the patient's values

b) in the PICO structure, C stands for the comparator

c) in the GRADE system, evidence from observational studies starts at high certainty

d) a narrative review ranks above a systematic review in the hierarchy of evidence

e) heterogeneity in a meta-analysis can be quantified by I^2

9. (MC) Regarding cognitive errors in clinical reasoning:

a) base-rate neglect means ignoring the prevalence of the disease

b) anchoring means insufficient adjustment of the first impression when new data appear

c) availability leads to overestimating the probability of diseases that are easy to recall

d) verification bias is a cognitive error made by the physician at the bedside

e) premature closure means continuing investigations after the diagnosis has been established

10. (MC) Regarding Bayesian inference:

a) the posterior distribution is proportional to the likelihood multiplied by the prior distribution

b) as the sample grows, the influence of the prior on the result decreases

c) a 95% credible interval can be interpreted as a 95% probability that the parameter lies in the interval, given the prior and the data

d) the Bayes factor can provide evidence in favor of the null hypothesis

e) a Beta(1; 1) prior favors small values of a proportion

Chapter 12. Evaluating the Performance of Diagnostic Tests: Sensitivity, Specificity, Predictive Values and the ROC Curve

Objectives

By the end of this chapter, the student will be able to:

- define the reference standard and the index test, and describe the correct design of a diagnostic accuracy study;
- build the 2 x 2 table (T+/T- versus B+/B-) with the notation a, b, c, d, and calculate prevalence, sensitivity, specificity, predictive values, accuracy, misclassification rate, the Youden index, likelihood ratios and the diagnostic odds ratio;
- reconstruct a complete 2 x 2 table starting from a few indicators and a few known cells;
- explain why predictive values depend on prevalence and recalculate them for a different population;
- recognize spectrum bias, verification bias and incorporation bias in a published article;
- construct an ROC curve by hand from a table of cutoffs, interpret the area under the curve and its confidence interval, and compare two ROC curves;
- choose a decision cutoff (Youden index, rule-out cutoff, rule-in cutoff) according to the clinical purpose;
- calculate and interpret Cohen's kappa coefficient and recognize the situations in which the intraclass correlation coefficient is used;
- apply the same indicators to the evaluation of an artificial intelligence classifier, and use the STARD 2015 and STARD-AI checklists in the critical reading of a study.

12.1 The Index Test and the Reference Standard

Any diagnostic test, from a question in the medical history to a biomarker or an artificial intelligence algorithm, must answer a simple question: how well does it separate patients who have the disease from those who do not? To answer it, we need two things. The first is the index test, that is, the new, simpler, cheaper or less invasive test whose performance we want to know. The second is the reference standard (traditionally called the "gold standard"), that is, the method considered the best available for establishing the presence or absence of the disease.

In the mg2026 dataset, the index test is the FINDRISC score, and the reference standard is the oral glucose tolerance test (OGTT). In the research of our colleagues in Timişoara, the reference standards are highly varied: liver biopsy with Metavir staging for fibrosis [107], histology for cervical lesions [131], upper gastrointestinal endoscopy for esophageal varices [110], and the final diagnosis established on all available data for malignant biliary obstruction [57]. Sometimes the reference standard is itself a noninvasive method, chosen because it is better validated: vibration-controlled transient elastography (VCTE) for time-harmonic elastography [3] or the controlled attenuation parameter (CAP) for the new methods of quantifying steatosis [58,65].

The reference standard is almost never perfect. Liver biopsy analyzes a fragment a few millimeters long from an entire organ; histology depends on the pathologist; the OGTT has modest reproducibility. When the reference standard is wrong, the index test will appear weaker

than it is, because disagreements will automatically be attributed to the new test. This should be kept in mind when a new test "loses" against an imperfect reference.

12.1.1 The Design of a Diagnostic Accuracy Study

The ideal design is cross-sectional: a series of consecutive patients, representative of those in whom the test will be used in practice (for example, adults with suspected disease in the family practice), who undergo both the index test and the reference standard, independently and without the interpreter of either knowing the result of the other. The prevalence of the disease in the sample is then an estimate of the true prevalence, and all the indicators, including the predictive values, are meaningful for that population.

The diagnostic case-control design (known patients compared with healthy volunteers) is easier to carry out and is often used in the early phases of evaluating a test. It has, however, two major problems. First, the prevalence is fixed by the researcher (for example, cases and controls in a 1:2 ratio), so the calculated predictive values have no clinical meaning. Second, comparing typical patients, often in advanced stages, with perfectly healthy people exaggerates accuracy, because precisely the difficult cases are missing (early disease, patients with other diseases that mimic the disease under study). We will return to this spectrum bias in Section 12.5.

12.2 The 2 x 2 Table and the Performance Indicators

12.2.1 Notation

The course of our discipline uses a table with the disease in the rows and the test result in the columns. We keep this notation throughout the chapter.

Table 12.1. The 2 x 2 table for evaluating a diagnostic test, with the notation used in the lectures.

	T+	**T-**	**Total**
B+	a (true positives, TP)	b (false negatives, FN)	a + b
B-	c (false positives, FP)	d (true negatives, TN)	c + d
Total	a + c	b + d	$N = a + b + c + d$

Cell a contains the diseased with a positive test (true positives), b the diseased with a negative test (false negatives), c the people without the disease under study but with a positive test (false positives), and d the people without the disease and with a negative test (true negatives). In some textbooks and programs (for example MedCalc), the table is transposed, with the disease in the columns and the test in the rows; letters b and c then swap places. Therefore, before any calculation, always check what each cell represents, not which letter it has.

12.2.2 Prevalence, Sensitivity and Specificity

The prevalence (in the sample) is the proportion of diseased individuals:

prevalence = (a + b) / N

Sensitivity (Se) is the probability that a diseased person has a positive test, that is, the proportion of diseased people correctly identified:

$$\text{Se} = p(T^+/B^+) = \frac{a}{a+b}$$

Specificity (Sp) is the probability that a person without the disease under study has a negative test:

$$\text{Sp} = p(T^-/B^-) = \frac{d}{c+d}$$

Sensitivity is calculated only on the row of the diseased, specificity only on the row of the non-diseased. Consequently, in principle, they do not depend on prevalence: if we double the number of healthy people in the study, specificity remains the same, and sensitivity does not even notice. We say "in principle" because sensitivity and specificity depend on the spectrum of patients (disease stage, comorbidities), and the spectrum often changes along with prevalence.

Two English mnemonics are useful at the bedside. SnNout (high Sensitivity, Negative result, rules out): a highly sensitive test, if negative, rules out the disease, because a sensitive test misses few diseased patients. SpPin (high Specificity, Positive result, rules in): a highly specific test, if positive, confirms the disease, because a specific test gives few false positives. The mnemonics are approximations: a test with Se 95% and Sp 40% "rules out" the disease well only if the pre-test probability is not very high, as we saw in Chapter 11.

12.2.3 Predictive Values

The positive predictive value (PPV) is the probability that a person with a positive test has the disease, that is, the proportion of diseased people among those with a positive test:

$$\text{PPV} = p(B^+/T^+) = \frac{a}{a+c}$$

The negative predictive value (NPV) is the proportion of people without the disease among those with a negative test:

$$\text{NPV} = p(B^-/T^-) = \frac{d}{b+d}$$

Predictive values are calculated on the columns. They answer the question the patient asks ("my test came back positive; how likely is it that I have the disease?"), but they depend strongly on prevalence, as we will see in Section 12.4. Predictive values are, in fact, the post-test probabilities from Chapter 11, calculated for a pre-test probability equal to the prevalence in the sample [133].

12.2.4 Accuracy and the Misclassification Rate

Accuracy (Acc) is the proportion of all people correctly classified by the test:

$$\text{Acc} = \frac{a+d}{N}$$

The misclassification rate, or error rate (ER), is the proportion of people classified incorrectly, that is, of false negatives and false positives:

$$\text{ER} = \frac{b+c}{N} = 1 - \text{Acc}$$

Accuracy seems the most intuitive indicator, but it can be misleading. It is a weighted average of sensitivity and specificity, with weights given by prevalence: Acc = Se x prevalence + Sp x (1 - prevalence). When the disease is rare, accuracy is dominated by specificity. A "test" that declares everyone healthy has, at a prevalence of 2%, an accuracy of 98%, even though it detects no diseased person. For this reason, accuracy is never reported on its own.

> **The rule that prevents the error**
>
> The misclassification rate is calculated from the cells that contain errors: the false negatives (b) and the false positives (c). The correct formula is $\text{ER} = \frac{b+c}{N} = \frac{\text{FN}+\text{FP}}{N}$. The formula (b + d) / N, which sometimes appears in lecture notes, is wrong: d contains the true negatives, that is, people classified correctly. A quick check: Acc + ER must be exactly 1. For the table

> 150 / 30 / 4 / 256 in Worked Problem 12.1, $\frac{a+d}{N} = \frac{406}{440} = 0.923$, and $\frac{b+c}{N} = \frac{34}{440} = 0.077$; the sum is 1.

12.2.5 The Youden Index, Likelihood Ratios and the Diagnostic Odds Ratio

The Youden index combines sensitivity and specificity into a single number [134]:

$$J = \text{Se} + \text{Sp} - 1$$

J ranges from 0 (the test discriminates no better than tossing a coin) to 1 (a perfect test). Its advantage is that it does not depend on prevalence; its disadvantage is that it gives equal importance to false negatives and false positives, which is rarely true clinically. We will use it to choose the cutoff on the ROC curve.

The likelihood ratios, introduced in Chapter 11, are also calculated from Se and Sp [121]:

$$\text{LR}^+ = \frac{\text{Se}}{1 - \text{Sp}}$$

$$\text{LR}^- = \frac{1 - \text{Se}}{\text{Sp}}$$

The diagnostic odds ratio (DOR) is the ratio of the odds of a positive test in the diseased to the odds of a positive test in the non-diseased:

$$\text{DOR} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \times d}{b \times c} = \frac{\text{LR}^+}{\text{LR}^-}$$

It is the same calculation as the odds ratio in risk analysis, applied to the diagnostic table. A DOR of 1 means a useless test; values in the tens or hundreds indicate a powerful test. The DOR is useful for comparing tests in meta-analyses, but it is hard to interpret clinically and does not tell us whether the test is better at confirmation or at exclusion: two tests with Se 90% / Sp 50% and Se 50% / Sp 90% have exactly the same DOR (9).

Table 12.2. Summary of diagnostic performance indicators (notation used in the lectures).

Indicator	Formula	Depends on prevalence?	Question it answers
Prevalence	$\frac{a+b}{N}$	(it is the prevalence)	How common is the disease in the sample?
Sensitivity	$\frac{a}{a+b}$	no (in principle)	What proportion of the diseased does the test detect?
Specificity	$\frac{d}{c+d}$	no (in principle)	What proportion of the healthy are negative?
PPV	$\frac{a}{a+c}$	yes	Given a positive result, how likely is disease?
NPV	$\frac{d}{b+d}$	yes	Given a negative result, how likely is health?
Accuracy	$\frac{a+d}{N}$	yes	What proportion is classified correctly?
Error rate	$\frac{b+c}{N}$	yes	What proportion is misclassified?
Youden index	Se + Sp - 1	no	How well does it discriminate overall?
LR+	$\frac{\text{Se}}{1-\text{Sp}}$	no	By what factor do the odds increase with a positive result?
LR-	$\frac{1-\text{Se}}{\text{Sp}}$	no	By what factor do the odds decrease with a negative

Indicator	Formula	Depends on prevalence?	Question it answers
			result?
DOR	$\frac{a \times d}{b \times c}$	no	Odds ratio of a positive test

12.2.6 The Indicators in the mg2026 Dataset and Their Confidence Intervals

At the cutoff FINDRISC ≥ 12, the table for OGTT-confirmed diabetes is: a = 22 (TP), b = 7 (FN), c = 33 (FP), d = 98 (TN), with N = 160 and 29 diseased patients (prevalence 18.1%).

Table 12.3. FINDRISC ≥ 12 versus OGTT, mg2026 dataset (synthetic data).

	FINDRISC ≥ 12	FINDRISC < 12	Total
Diabetes (OGTT)	22	7	29
No diabetes	33	98	131
Total	55	105	160

Se $= \frac{22}{29} = 75.9\%$; Sp $= \frac{98}{131} = 74.8\%$; PPV $= \frac{22}{55} = 40.0\%$; NPV $= \frac{98}{105} = 93.3\%$; Acc $= \frac{120}{160} = 75.0\%$; ER $= \frac{40}{160} = 25.0\%$; J = 0.759 + 0.748 - 1 = 0.507; LR+ = 3.01; LR- = 0.32; DOR $= \frac{22 \times 98}{7 \times 33} = \frac{2156}{231} = 9.33$.

Each of these indicators is a proportion estimated from a sample and must be accompanied by a confidence interval. For proportions, the Wilson method, discussed in Chapter 4, is preferred. Sensitivity is based on only 29 diseased patients, so its interval is wide: Se 75.9% (95% CI 57.9-87.8). Specificity, calculated on 131 non-diseased individuals, is more precise: 74.8% (95% CI 66.7-81.5). PPV 40.0% (95% CI 28.1-53.2), NPV 93.3% (95% CI 86.9-96.7). For the likelihood ratios, a method on the logarithmic scale is used: LR+ 3.01 (95% CI 2.10-4.31), LR- 0.32 (95% CI 0.17-0.62), and DOR 9.33 (95% CI 3.65-23.84).

This is a lesson you will see again in every diagnostic accuracy article: sensitivity is almost always estimated less precisely than specificity, because there are fewer diseased patients. A study with 30 diseased patients cannot estimate sensitivity more precisely than about ±15 percentage points, however many decimal places it reports [135].

12.3 Reconstructing a 2 x 2 Table

A type of problem specific to our course, and frequent in the exam, asks you to reconstruct the entire table starting from a few indicators and a few cells. These problems look complicated, but they are always solved with the same four-step strategy.

Step 1. Write out the empty table, with the letters a, b, c, d and the marginal totals, and write down the formula for each given indicator: Se $= \frac{a}{a+b}$, Sp $= \frac{d}{c+d}$, PPV $= \frac{a}{a+c}$, NPV $= \frac{d}{b+d}$, prevalence = (a + b)/N.

Step 2. Fill in the cells that are given directly (for example, "TP = 90" means a = 90).

Step 3. Choose an indicator that contains a single unknown and solve the equation. For example, if PPV $= \frac{9}{12}$ and a = 90, then $\frac{90}{90+c} = \frac{9}{12}$, whence 9(90 + c) = 12 x 90, so c = 30. Repeat until all cells are known.

Step 4. Calculate the totals and check the given indicators. Then calculate the requested indicators.

A common mistake is to use an indicator with two unknowns before finding one of them. Another kind of mistake occurs when ratios are given as fractions (2/3, 6/8): do not convert them into rounded decimals (0.66) before solving; work with the fraction, otherwise the cells will no longer come out as whole numbers. Worked Problems 12.2-12.5 apply the strategy step by step.

12.4 The Dependence of Predictive Values on Prevalence

Sensitivity and specificity are properties of the test (for a given spectrum of patients). Predictive values are properties of the test applied to a particular population. Bayes' theorem, written with the indicators from this chapter, shows the exact relationship:

$$\mathrm{PPV} = \frac{\mathrm{Se} \times P}{\mathrm{Se} \times P + (1 - \mathrm{Sp}) \times (1 - P)}$$

$$\mathrm{NPV} = \frac{\mathrm{Sp} \times (1 - P)}{\mathrm{Sp} \times (1 - P) + (1 - \mathrm{Se}) \times P}$$

where P is the prevalence. As prevalence increases, the PPV increases and the NPV decreases.

Figure 12.1 shows this dependence for FINDRISC ≥ 12 (Se 76%, Sp 75%), and Table 12.4 gives the exact values calculated on the mg2026 dataset for several prevalences.

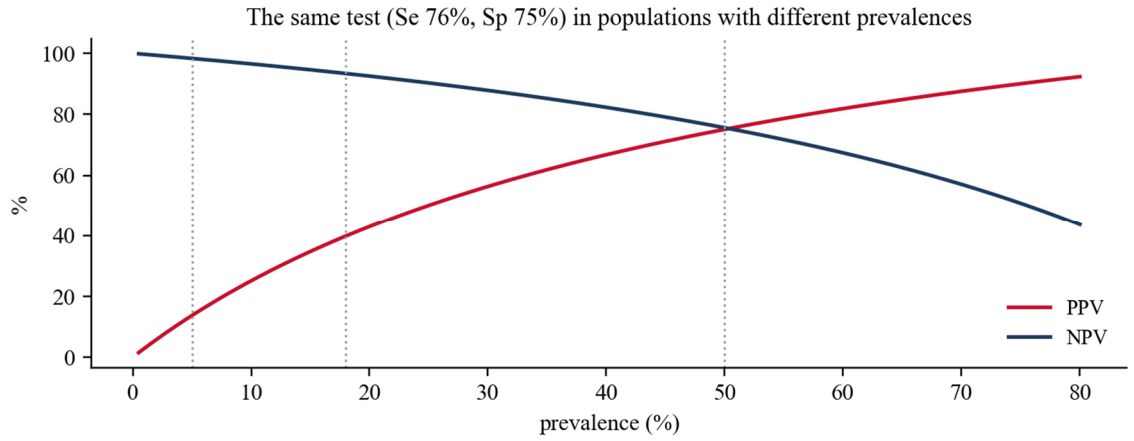


Figure 12.1. Positive and negative predictive value of the same test (Se 76%, Sp 75%, FINDRISC ≥ 12) in populations with different prevalences. The dotted lines mark prevalences of 5%, 18% and 50%. Synthetic data (mg2026 dataset).

Table 12.4. PPV and NPV of FINDRISC ≥ 12 (Se 75.9%, Sp 74.8%) for different prevalences of undiagnosed diabetes. Synthetic data (mg2026 dataset).

Prevalence	**2%**	**5%**	**10%**	**18%**	**30%**	**50%**
PPV (%)	5,8	13,7	25,1	39,8	56,3	75,1
NPV (%)	99,3	98,3	96,5	93,4	87,9	75,6

At a prevalence of 2% (for example, young adults without risk factors), out of 100 people with FINDRISC ≥ 12, only about 6 have diabetes. At a prevalence of 50% (for example, patients with polyuria, polydipsia and weight loss), 75 out of 100 have diabetes. The test is the same; the population is different. The practical conclusion has two sides. Predictive values reported in an article apply only to populations with a similar prevalence. To transfer them to another population, they are recalculated from Se, Sp and the new prevalence (or, equivalently, with the likelihood ratios) [133,136].

This dependence explains why screening in the general population produces so many false-positive results and why confirmatory tests are indispensable. It also explains why the same biomarker looks excellent in a hospital study and disappointing in the family practice.

12.5 Bias in Diagnostic Accuracy Studies

Indicators calculated correctly from a correct table can still be misleading if the table comes from a study with a flawed design. Three biases occur so often that every physician must be able to recognize them.

Spectrum bias occurs when the patients in the study do not resemble those in whom the test will be used. The most common form is the case-control design with patients in advanced stages and perfectly healthy controls. Patients with advanced disease have markedly abnormal test values (apparently high sensitivity), and healthy people have no disease that could produce false-positive results (apparently high specificity). In practice, the test will encounter early cases and patients with other diseases, and performance will drop. Spectrum bias is not just a design problem: it reflects the fact that sensitivity and specificity really do vary with the type of patient. The study by Pavel and colleagues on H-FABP, discussed in Chapter 11, showed a sensitivity and specificity of 89.67 and 95.65 within 4 hours of pain onset, but 86.73 and 49.84 after 4 hours [127]. Sipos and colleagues observed, for IL-6 in neonatal sepsis, a sensitivity of 90% in one subgroup and 71.4% in another, depending on culture status [126].

Verification bias (also called work-up bias) occurs when the reference standard is performed in only some of the patients, and the selection depends on the result of the index test. Typically, patients with a positive test are referred for biopsy or colonoscopy, and those with a negative test are not. The false negatives then remain undetected (cell b is underestimated), and sensitivity appears higher than it is. In the study by Moleriu and colleagues, of 184 patients with LSIL cytology, only 64 underwent biopsy [131]; in the study by Iovanescu and colleagues, 252 of 2201 patients underwent colonoscopy, and adherence to colonoscopy among FIT-positive patients was 52% [101]. In both situations, the criteria used to decide on verification influence the estimates, and the authors must describe them.

Incorporation bias occurs when the result of the index test is part of the definition of the reference standard. If the final diagnosis of pancreatitis is established, among other things, on the basis of lipase, we cannot then correctly evaluate the accuracy of lipase: the test "confirms itself". The problem is common when the reference standard is a final clinical diagnosis built from all available data, including the test under evaluation. The study by Rațiu and colleagues compared ERCP, MRCP and histology with the final diagnosis of malignant biliary obstruction [57]; with such a design, the reader must check whether the final diagnosis used the results of the tests being compared.

Other sources of bias are interpreting the test with knowledge of the reference standard result (interpretation bias, prevented by blinding), excluding uninterpretable results (for example, invalid elastography measurements), and using the same sample both to choose the cutoff and to estimate performance at that cutoff (optimistic overestimation, corrected by validation in a separate sample).

> **Case study: the case-control design and the apparent performance of tests [2,116,132]**
>
> Three studies by our colleagues use a case-control design and provide a good exercise in critical reading. Lupușoru and colleagues evaluated contrast-enhanced ultrasound with parametric imaging of arrival time (CEUS-PAT) for liver cirrhosis in 64 people: 37 with cirrhosis and 27 healthy volunteers. CEUS-PAT values were 0.83 ± 0.09 in cirrhotic patients

> and 0.49 ± 0.11 in healthy individuals ($p < 0.0001$), the correlation with VCTE was $r = 0.81$, and at the cutoff > 0.7 the area under the ROC curve was 0.98, with Se 89.19%, Sp 100%, PPV 100% and NPV 87.1%. The figures can be reconstructed exactly: 33 of the 37 cirrhotic patients were positive (89.19%), all 27 volunteers were negative (100%), hence 31 negatives, of whom 27 were healthy $\left(\frac{\text{NPV } 27}{31} = 87.1\%\right)$. Golu and colleagues (with L. Moleriu and A. Tudor) studied 2D-SWE elastography in 22 patients with hyperparathyroidism and 43 healthy controls; the mean stiffness values reported in the abstract were 10.2 ± 4.9 kPa and 19.5 ± 7.6 kPa ($p = 0.007$), and at the 12.5 kPa cutoff, Se 93%, Sp 86% and AUC 0.949. Feier and colleagues (with C. Muntean) compared 145 cases of gangrenous cholecystitis with 290 cases of chronic cholecystitis, in a fixed 1:2 ratio, matched for age and sex; the best inflammatory index had an AUC of 0.889, Se 79.5% and Sp 86.3%. What do the three studies have in common? The prevalence is fixed by the design (57.8%, 33.8% and 33.3%), so the predictive values do not apply directly in practice. When the controls are healthy volunteers (the first two studies), the very high AUCs (0.98 and 0.949) probably represent the upper limit of performance; patients with non-cirrhotic chronic liver disease or with thyroid nodules would have been the clinically relevant comparators. Feier's study is closer to the clinical situation, because the controls also have cholecystitis, but the 1:2 ratio remains artificial. All three are necessary steps in the evaluation of a test; the next step would be a cohort of consecutive patients with clinical suspicion.

12.6 Tests with a Continuous Result and the Choice of Cutoff

12.6.1 The Trade-off between Sensitivity and Specificity

Many tests do not naturally produce a "positive" or "negative" result, but a numerical value: blood glucose, troponin, liver stiffness, a clinical score. To use them for decision-making, we choose a cutoff: values above the cutoff are considered positive. Choosing the cutoff changes the 2 x 2 table and, with it, all the indicators.

The mechanism can be seen in panel B of Figure 12.3, below: the distribution of the FINDRISC score in people with diabetes (mean 14.76, SD 3.96) is shifted to the right relative to the distribution in people without diabetes (mean 9.40, SD 3.70), but the two distributions overlap. Wherever we place the cutoff, some diseased people will remain below it (false negatives) and some healthy people will be above it (false positives). If we move the cutoff to the left (lower values), we catch more diseased people (sensitivity increases), but we also declare more healthy people positive (specificity decreases). If we move it to the right, the opposite happens. There is no cutoff that increases both; there are only trade-offs. The course of our discipline describes this as a relationship of "inverse proportionality" between sensitivity and specificity; more precisely, it is an inverse monotonic relationship: when one increases, the other can only decrease or stay the same.

12.6.2 The Table of Cutoffs for FINDRISC

Table 12.5 shows, for several cutoffs of the FINDRISC score, the four cells and the main indicators, calculated on the mg2026 dataset.

Table 12.5. Performance of FINDRISC for OGTT-confirmed diabetes at different cutoffs (n = 160; 29 with diabetes, 131 without). Synthetic data (mg2026 dataset).

Cutoff	TP	FP	FN	TN	Se (%)	Sp (%)	1 - Sp	PPV (%)	NPV (%)	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

Cutoff	TP	FP	FN	TN	Se (%)	Sp (%)	1 - Sp	PPV (%)	NPV (%)	J
≥ 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

Read the table from top to bottom. As the cutoff increases, TP decrease and FN increase (sensitivity drops from 96.6% to 41.4%), while FP decrease and TN increase (specificity rises from 28.2% to 94.7%). The PPV increases, the NPV decreases. The Youden index reaches its maximum, 0.595, at the cutoff ≥ 11 (Se 93.1%, Sp 66.4%). Note that at all cutoffs the prevalence remains 18.1%; only the way the test divides the patients changes.

12.7 The ROC Curve

12.7.1 Step-by-Step Construction

The ROC curve (receiver operating characteristic, a name inherited from radar signal theory) graphically represents all possible trade-offs between sensitivity and specificity. The horizontal axis shows 1 - specificity (the proportion of false positives among the healthy), and the vertical axis shows sensitivity. Each cutoff becomes a point. The lecture writes the relationship as Se = f(1 - Sp).

Construction by hand has five steps.

Step 1. Order the cutoffs from the strictest (the highest, which gives the fewest positive results) to the most permissive.

Step 2. For each cutoff, calculate Se and 1 - Sp from the corresponding 2 x 2 table (Table 12.5).

Step 3. Add the two extreme points: (0; 0), which corresponds to a cutoff so high that no one is positive, and (1; 1), which corresponds to a cutoff so low that everyone is positive.

Step 4. Plot the points on the graph and join them with straight-line segments, in the order of the cutoffs.

Step 5. Draw the diagonal from (0; 0) to (1; 1): it represents a worthless test, for which the proportion of positives is the same in the diseased and the healthy.

For FINDRISC, with the six cutoffs ≥ 16, ≥ 15, ≥ 14, ≥ 12, ≥ 10 and ≥ 8, the points (1 - Sp; Se) are: (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). Figure 12.2 shows the resulting curve, superimposed on the complete curve constructed with all possible cutoffs of the score.

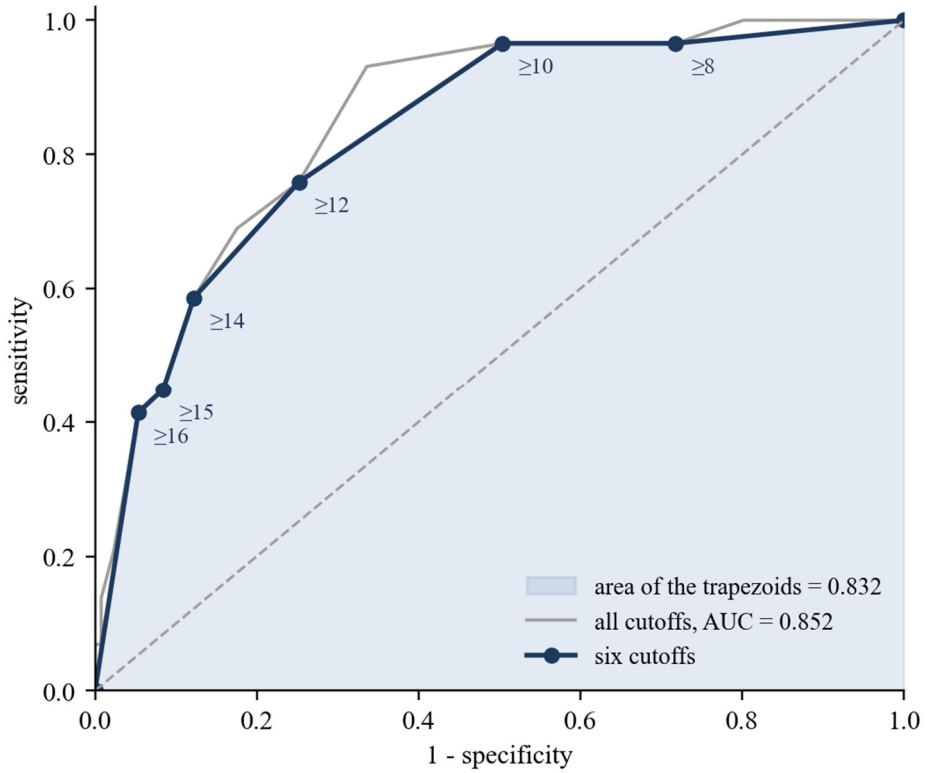


Figure 12.2. ROC curve of the FINDRISC score for OGTT-confirmed diabetes, constructed by hand from six cutoffs (line with markers), superimposed on the complete curve with all cutoffs (gray line). The area of the trapezoids under the six-cutoff curve is 0.832; the area under the complete curve is 0.852. Synthetic data (mg2026 dataset).

The area under the curve can be calculated by hand using the trapezoidal method. Between two consecutive points (x1; y1) and (x2; y2), the area of the trapezoid is (x2 - x1) x (y1 + y2) / 2. For the seven segments we obtain 0.0111 + 0.0132 + 0.0197 + 0.0873 + 0.2172 + 0.2064 + 0.2776 = 0.832. With all possible cutoffs, the area is 0.852, the value reported by software. The difference arises because, by joining sparse points with straight lines, we "cut the corners" of the curve; the more cutoffs we use, the closer the area of the trapezoids gets to the true area.

12.7.2 What the Area under the ROC Curve Means

The area under the ROC curve (AUC), also called the c statistic in the lectures, is an overall indicator of the test's ability to separate the diseased from the healthy, independently of the cutoff [137,138]. It has a very concrete probabilistic interpretation: the AUC is the probability that a randomly chosen diseased person has a higher test value than a randomly chosen healthy person (pairs with equal values count as one half).

For FINDRISC in the mg2026 dataset, we can check this directly. There are 29 x 131 = 3799 diseased-healthy pairs. In 3154 pairs, the diseased person has the higher score; in 169 pairs, the scores are equal. $\text{AUC} = \frac{3154+\frac{169}{2}}{3799} = \frac{3238.5}{3799} = 0.852$. Interpretation: if we randomly pick one patient with diabetes and one without, in 85.2% of cases the one with diabetes has the higher FINDRISC score. The calculation is identical to the U statistic of the Mann-Whitney test from Chapter 7, divided by the product of the sizes of the two groups; the AUC is therefore an effect size measure for the difference between the distributions of a marker in the diseased and the healthy.

A perfect test has AUC = 1 (the curve rises vertically to Se = 1 at 1 - Sp = 0, then runs horizontally). A useless test has AUC = 0.5 (the curve coincides with the diagonal). An AUC

below 0.5 usually indicates a coding error: the test works "the wrong way round", and reversing the direction (low values are positive) gives AUC = 1 - AUC.

12.7.3 Interpretation Scales for the AUC

The course of our discipline uses the following scale for the c statistic:

Table 12.6. Interpretation scale for the area under the ROC curve used in the lectures of our discipline.

AUC	Interpretation
0,91-1	excellent
0,81-0,90	very good
0,71-0,80	good
0,61-0,70	satisfactory
below 0.6	poor

Other textbooks use slightly different scales. A variant common in the clinical literature considers 0.9-1 excellent, 0.8-0.9 good, 0.7-0.8 acceptable (or moderate), 0.6-0.7 poor and 0.5-0.6 failed. The differences between scales show that the thresholds are conventions, not laws. Interpretation must also take the clinical context into account: an AUC of 0.75 may be very useful for a cheap triage test and insufficient for a test that decides on surgery. On the scale used in the lectures, FINDRISC (0.852) and fasting plasma glucose (0.844) are "very good" tests for OGTT-confirmed diabetes in the mg2026 dataset, HbA1c (0.764) is "good", and BMI (0.731) and waist circumference (0.737) are "good", toward the lower limit.

Examples from our colleagues' research cover the whole scale: CEUS-PAT for cirrhosis, AUC 0.98 (excellent) [2]; the combined score for high-risk varices, AUROC 0.93 [110]; ISARIC-4C for mortality, 0.875 [75]; global longitudinal strain for heart failure with preserved ejection fraction, 0.833 [90]; ALRI in patients with COVID-19 and diabetes, 0.752, compared with only 0.624 for dNLR (satisfactory) [105].

12.7.4 The Confidence Interval for the AUC

The AUC is a sample estimate and has a standard error. Hanley and McNeil proposed a formula that depends only on the AUC and on the numbers of diseased and healthy individuals [137]; the DeLong method, used by default in many programs, estimates the standard error nonparametrically, from the actual data. For FINDRISC in mg2026, the 95% CI is 0.762-0.943 with the Hanley-McNeil method and approximately 0.782-0.923 with the DeLong method. For fasting plasma glucose, the AUC is 0.844, with a 95% CI of 0.752-0.937 (Hanley-McNeil). Both intervals are wide, because they are based on only 29 diseased patients. It is the lower limit of the interval that matters for a cautious statement: we can say that FINDRISC has at least "good" performance (0.76), not necessarily "very good".

A confidence interval that includes 0.5 means that the test does not discriminate significantly better than chance. The associated statistical test (H0: AUC = 0.5) appears in all programs; for FINDRISC, $p < 0.001$.

12.7.5 Comparing Two ROC Curves

When two tests are performed in the same patients, their AUCs are correlated, and the comparison must take this correlation into account, just as the paired t-test takes into account the correlation between repeated measurements. The DeLong method does exactly this: it estimates the variance

of the difference between two AUCs on the same data and calculates a z statistic. When the tests are evaluated in different groups of patients, a comparison for independent samples is used.

Figure 12.3 compares FINDRISC with fasting plasma glucose in the mg2026 dataset.

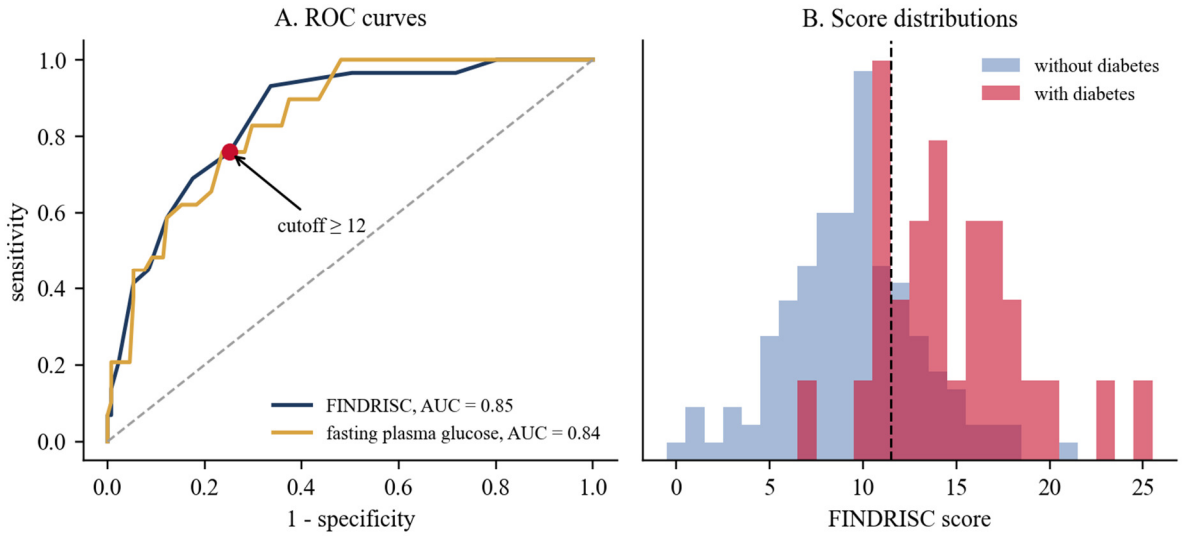


Figure 12.3. A: ROC curves of the FINDRISC score (AUC 0.85) and fasting plasma glucose (AUC 0.84) for OGTT-confirmed diabetes; the red dot marks the cutoff FINDRISC ≥ 12. B: distributions of the FINDRISC score in people without and with diabetes; the dotted line marks the cutoff between 11 and 12. Synthetic data (mg2026 dataset).

The difference between the AUCs is 0.852 - 0.844 = 0.008 (95% CI -0.060 to 0.077), z = 0.24, p = 0.81 (DeLong method, with a correlation of approximately 0.5 between the two AUCs). The two tests have virtually identical discriminative ability for OGTT-confirmed diabetes. The choice between them is made on other criteria: FINDRISC requires no blood sampling and can be completed by the patient in the waiting room, whereas blood glucose is an objective measurement but requires fasting and a laboratory. Note also the shape of the curves: they cross, so at some cutoffs one test is better and at others the other. The AUC summarizes the entire curve, including cutoff regions we would never use in practice.

Case study: comparing several tests by AUROC [57,107]

Rațiu and colleagues (with R. Lupușoru) compared three methods against the final diagnosis of malignancy in 160 patients with biliary obstruction (mean age 69.31 ± 10.96 years; 53.1% men): the AUROC was 0.88 for MR cholangiopancreatography (95% CI 0.75-0.90), 0.94 for ERCP (95% CI 0.85-0.99) and 0.80 for histology (95% CI 0.70-0.82). Two observations for critical reading. First, the confidence intervals are highly asymmetric around the estimates (0.88 with an upper limit of 0.90, 0.80 with an upper limit of 0.82), which is unusual for the standard calculation methods and is worth checking in the full text. Second, the intervals overlap, but overlapping intervals are not a test of the difference; for tests performed in the same patients, the correct comparison is the DeLong method on paired data. Bâldea and colleagues (with A. Tudor) compared transient elastography (TE) with ARFI-based VTQ quantification in 157 patients with chronic hepatitis B or C, with liver biopsy as the reference standard. The distribution of fibrosis stages was F0 19.7%, F1 22.3%, F2 27.4%, F3 17.8%, F4 12.7%, and the correlation between the two methods was strong (Spearman ρ = 0.826; 95% CI 0.769-0.870). For the successive fibrosis thresholds, the AUROCs were 0.876 versus 0.832 (p = 0.358), 0.826 versus 0.862 (p = 0.313), 0.907 versus 0.880 (p = 0.434) and 0.981 versus 0.974 (p = 0.423), TE versus VTQ. None of the differences was significant. The correct interpretation is not "the two methods are equivalent" but "the study found no differences; the observed differences, of the order of 0.03-0.04, are compatible with chance". Demonstrating

equivalence would have required an equivalence design, with a margin set in advance, or the confidence intervals of the differences.

12.7.6 Choosing the Cutoff: The Youden Index and Clinical Cutoffs

The cutoff with the maximum Youden index is the point on the ROC curve that lies farthest vertically from the diagonal. For FINDRISC in mg2026, the Youden cutoff is ≥ 11, with Se 93.1%, Sp 66.4% and J = 0.595. For fasting plasma glucose, the Youden cutoff is ≥ 110 mg/dL, with Se 82.8%, Sp 70.2% and J = 0.530. The Youden index, however, assumes that a false negative and a false positive are equally important and that prevalence does not matter, which rarely matches clinical reality.

In practice, the cutoff is chosen according to the purpose of the test. For a triage test, where a negative result ends the investigations, what matters is not missing diseased patients: a rule-out cutoff with high sensitivity and high NPV is chosen. In mg2026, FINDRISC ≥ 10 has Se 96.6% and NPV 98.5%: a patient with a score below 10 has a probability of diabetes of about 1.5% and can be reassured. For a test that triggers an invasive or costly treatment, what matters is not treating healthy people: a rule-in cutoff with high specificity and high PPV is chosen. FINDRISC ≥ 16 has Sp 94.7% and PPV 63.2%. Between the two cutoffs lies a gray zone, in which the decision requires additional information.

The glucose example shows one more important point. The clinical diagnostic cutoff for diabetes for fasting plasma glucose, 126 mg/dL [80], has in the mg2026 dataset a Se of only 37.9% and a Sp of 94.7% (PPV 61.1%, NPV 87.3%) against the OGTT. The cutoff is chosen to confirm the diagnosis (high specificity), not for screening; used alone, it would miss almost two thirds of the diabetic patients identified by the OGTT. The "statistically optimal" cutoff (110 mg/dL) and the "clinical" cutoff (126 mg/dL) answer different questions.

The same reasoning appears in our colleagues' studies. Moga and colleagues reported for time-harmonic elastography a rule-out cutoff for cirrhosis (< 1.61 m/s: Se 90.7%, NPV 93.9%) and a rule-in cutoff (> 1.83 m/s: Sp 96.7%, PPV 90.3%) [3]. Barbulescu and colleagues constructed rule-out and rule-in criteria for high-risk varices, with a gray zone resolved by albumin [123]. Popa and colleagues (with A. Tudor) reported for GFAP, in 85 patients with a stroke code alert in the emergency department, an accuracy of 94.2% in differentiating hemorrhagic from ischemic stroke, at a cutoff of 77.15 pg/mL; GFAP was 142.91 ± 102.19 pg/mL in patients with acute stroke versus 37.76 ± 19.92 pg/mL in those without stroke ($p < 0.001$) [61]. A standard deviation close to the mean suggests a skewed distribution, for which the median and the Mann-Whitney test would have been more appropriate, and the cutoff with four significant figures, derived from 69 stroke patients, needs validation.

Case study: ROC cutoffs for diseases with several grades [58,65,90]

Hepatic steatosis has four grades (S0-S3), and fibrosis five stages (F0-F4). For such ordinal diseases, ROC analysis is performed separately for each severity threshold: S0 versus S1-S3, S0-S1 versus S2-S3, and so on. Bende and colleagues (with R. Lupușoru) evaluated the UGAP parameter against CAP in 176 subjects classified as follows: S0 27.2%, S1 31.6%, S2 7.3%, S3 33.9%. The mean UGAP values increased monotonically with grade ($198.3 \pm \frac{25.7 d_B}{m}$ in S0 and $270.8 \pm \frac{31.62 d_B}{m}$ in S3; $p < 0.001$), the correlation with CAP was r = 0.73, and the cutoffs were 192.5 dB/m for S1 (AUC 0.83), 231 dB/m for S2 (AUC 0.90) and 248 dB/m for S3 (AUC 0.91). Nica and colleagues (with R. Lupușoru) evaluated the ultrasound fat fraction (USFF) in 169 patients, with CAP as the reference (CAP cutoffs 248, 268 and 280 dB/m; 52.6% with at least moderate steatosis). USFF correlated with CAP (r = 0.81), and the cutoffs > 5%, > 7.5% and > 10% had AUROCs of 0.90, 0.92 and 0.91 for at least mild, moderate and

> severe steatosis, respectively. An interesting detail for critical reading: the abstract describes as "moderate" the correlation r = 0.83 between the TAI parameter and CAP, although on the Colton scale used in the lectures it would be "very good to strong". Stoichescu-Hogea and colleagues (with A. Tudor) compared global longitudinal strain (GLS) in three groups (62 patients with heart failure with preserved ejection fraction, 46 hypertensive patients, 40 controls): -17.6 ± 2.3% in the first group, -18.4 ± 3.0% in the hypertensive patients and -20.2 ± 1.4% in the controls, and the optimal cutoff of -19.35% had an AUC of 0.833. Here "positive" means a less negative value, and the software must be set correctly for the direction of the test (otherwise the AUC would come out as 0.167). In all three studies, the reference standard (CAP or the clinical diagnosis) is imperfect, and the cutoffs are derived from the same sample on which performance is estimated, so they need external validation.

12.8 Interrater Agreement

A useful diagnostic test must be not only valid (measure what it is supposed to measure) but also reliable: it must give the same result when repeated or interpreted by another examiner. For categorical results (for example, "left ventricular hypertrophy present or absent" on the ECG), agreement is measured with Cohen's kappa coefficient. For numerical results (liver stiffness in kPa, blood pressure), the intraclass correlation coefficient and the Bland-Altman method are used [102].

12.8.1 Cohen's Kappa Coefficient

The proportion of observed agreement (Po) is not enough, because two raters would often agree by chance as well, especially when one of the categories is very common. Kappa corrects the observed agreement for the agreement expected by chance (Pe):

$$\kappa = \frac{\mathrm{Po} - \mathrm{Pe}}{1 - \mathrm{Pe}}$$

The expected agreement is calculated from the marginal totals, just like the expected frequencies in the chi-square test: for each category, the product of the proportions with which the two raters used it, then the sum. Kappa is 1 for perfect agreement, 0 when agreement does not exceed chance, and negative when agreement is lower than expected by chance. The Landis and Koch scale is commonly used for interpretation [103].

Table 12.7. The Landis and Koch scale for interpreting the kappa coefficient.

κ	Agreement
below 0	poor (below chance level)
0-0,20	slight
0,21-0,40	fair (modest)
0,41-0,60	moderate
0,61-0,80	substantial
0,81-1,00	almost perfect

The complete example is given in Worked Problem 12.8. Kappa has some limitations: it depends on the prevalence of the categories (at extreme prevalences, kappa can be low even with high observed agreement), and it does not distinguish between small and large disagreements in the case of ordered categories. For ordinal scales (for example, steatosis grades), weighted kappa is used.

12.8.2 The Intraclass Correlation Coefficient

For numerical measurements, the Pearson correlation coefficient is not suitable for agreement: two devices that differ systematically by 5 kPa can have r = 1 but are not in agreement. The intraclass correlation coefficient (ICC) measures the proportion of total variability due to real differences between patients, relative to that due to differences between measurements or raters. Values above 0.90 indicate excellent reliability, between 0.75 and 0.90 good, and between 0.50 and 0.75 moderate.

Moga and colleagues included in their study of time-harmonic elastography a reproducibility substudy in 27 patients: the overall agreement between examiners was 0.94 (95% CI 0.89-0.97), and the ICCs for different comparisons were 0.92 (95% CI 0.82-0.96), 0.94 (95% CI 0.87-0.97) and 0.97 (95% CI 0.95-0.99) [3]. A test with such reproducibility can be used by different examiners without the differences between them changing the classification of patients. The ICC is also used for the test-retest reliability of questionnaires, for example ICC values between 0.82 and 0.91 over a 14-day interval in a validation study by our colleagues [73].

12.9 Evaluating Artificial Intelligence Classifiers

Artificial intelligence algorithms that classify medical images (radiographs, retinal photographs, dermoscopic images) or signals (ECG) are, from a statistical point of view, diagnostic tests. They are evaluated with the same tools, even though the computer science literature uses different terms.

Here the 2 x 2 table is called a confusion matrix. Sensitivity is called recall (or the true positive rate), the PPV is called precision, and specificity remains specificity (rarely, the true negative rate). An indicator specific to the field is the F1 score, the harmonic mean of precision and sensitivity:

F1 = 2 x precision x sensitivity / (precision + sensitivity)

F1 is high only when both components are high, and it is used mainly when the positive class is rare. Unlike the AUC, F1 depends on prevalence (through precision) and on the chosen cutoff.

The classic problem in evaluating classifiers is the accuracy paradox with imbalanced classes. If only 8% of the retinal photographs in a screening program show diabetic retinopathy, an algorithm that declares all images normal has an accuracy of 92% but a sensitivity of 0. This is why an article that reports only "model accuracy" cannot be evaluated. Other common pitfalls are evaluation on the same dataset on which the model was trained (overfitting), the lack of external validation on data from another center, and choosing the cutoff after seeing the results. Worked Problem 12.10 calculates all these indicators for a concrete example. For prediction models based on regression or machine learning, there is a dedicated reporting guideline, TRIPOD+AI [139].

12.10 Reporting Diagnostic Accuracy Studies: STARD 2015 and STARD-AI

The STARD 2015 guideline (Standards for Reporting of Diagnostic Accuracy Studies) comprises 30 items that a diagnostic accuracy article should report [140]. Some matter more for critical appraisal: identification of the study as a diagnostic accuracy study in the title or abstract; description of the population, the eligibility criteria and the recruitment method (consecutive or not); description of the index test and the reference standard in sufficient detail to allow replication; a statement of whether the interpreters were blinded to the result of the other test; a statement of whether the cutoffs were set in advance or derived from the data; the participant flow diagram (how many were eligible, how many received each test, how many were excluded and why); the complete 2 x 2 table; accuracy estimates with confidence intervals; uninterpretable results and how they were handled.

STARD-AI extends STARD to studies in which the index test is an artificial intelligence algorithm [141]. It adds requirements concerning the source and characteristics of the training, validation and test data, the strict separation of these sets, the algorithm version, data preprocessing, how the cutoff was chosen, performance in subgroups (sex, age, center, device type) and error analysis.

When reading an article, use the checklist as a filter: if the 2 x 2 table cannot be reconstructed, if you do not know how the patients were recruited, or if the cutoff was chosen on the same data on which performance is reported, the results should be read with caution, however high the AUC and sensitivity may be.

12.11 Worked Problems

Worked Problem 12.1. All the Indicators from a Complete Table

A test was evaluated in 440 patients against the reference standard. Of the 180 diseased patients, 150 had a positive test; of the 260 non-diseased individuals, 4 had a positive test. Calculate all the performance indicators.

Solution. The table, with the notation from the lecture: 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

Prevalence = 180/440 = 0.409 (40.9%).

$\text{Se} = \frac{150}{180} = 0.833(83.3\%)$. $\text{Sp} = \frac{256}{260} = 0.985(98.5\%)$.

$\text{PPV} = \frac{150}{154} = 0.974(97.4\%)$. $\text{NPV} = \frac{256}{286} = 0.895(89.5\%)$.

$\text{Acc} = \frac{150+256}{440} = \frac{406}{440} = 0.923(92.3\%)$. $\text{ER} = \frac{30+4}{440} = \frac{34}{440} = 0.077(7.7\%)$; check: 0.923 + 0.077 = 1.

$J = 0.833 + 0.985 - 1 = 0.818$.

$\text{LR}^+ = \frac{0.833}{1-0.985} = \frac{0.833}{0.0154} = 54.2$. $\text{LR}^- = \frac{0.167}{0.985} = 0.17$.

$\text{DOR} = \frac{150\times256}{30\times4} = 38.400/120 = 320$.

Interpretation. The test has a very high specificity and an LR+ above 50: a positive result virtually confirms the disease (SpPin). A sensitivity of 83.3%, however, means that 1 in 6 diseased patients has a negative test; a negative result does not rule out the disease, and an LR- of 0.17 makes it only moderately less likely. The very good predictive values (97.4% and 89.5%) apply to the prevalence of 40.9%; in a screening population, with a much lower prevalence, the PPV would decrease and the NPV would increase.

Worked Problem 12.2. Reconstructing the Table from the PPV, the NPV and Two Cells

For a test, the following are known: $\text{PPV} = \frac{9}{12}$, $\text{NPV} = \frac{7}{8}$, the number of true positives TP = 90 and the number of false negatives FN = 10. Reconstruct the table and calculate Se, Sp and the prevalence.

Solution. Step 1: a = TP = 90, b = FN = 10.

Step 2: from PPV $= \frac{a}{a+c} = \frac{9}{12}$ it follows that $\frac{90}{90+c} = \frac{9}{12}$, so 9 x (90 + c) = 12 x 90, that is, 810 + 9c = 1080, whence c = 30.

Step 3: from NPV $= \frac{d}{b+d} = \frac{7}{8}$ it follows that $\frac{d}{10+d} = \frac{7}{8}$, so 8d = 70 + 7d, whence d = 70.

	T+	T-	Total
B+	90	10	100
B-	30	70	100
Total	120	80	200

Step 4 (check): PPV $= \frac{90}{120} = \frac{9}{12}$ and NPV $= \frac{70}{80} = \frac{7}{8}$.

Indicators: Se $= \frac{90}{100} = 0.90$; Sp $= \frac{70}{100} = 0.70$; prevalence = 100/200 = 0.50; Acc $= \frac{160}{200} = 0.80$; ER $= \frac{10+30}{200} = 0.20$; $LR^+ = \frac{0.90}{0.30} = 3.0$; $LR^- = \frac{0.10}{0.70} = 0.14$; J = 0.60.

Interpretation. The test is better at exclusion than at confirmation. A prevalence of 50% is typical of a case-control design with equal groups, which means that the PPV of 75% does not carry over to a real population with a low prevalence.

Worked Problem 12.3. Reconstructing the Table from Se, the PPV and Two Cells

The following are known: Se $= \frac{2}{3}$, PPV $= \frac{6}{8}$, FN = 30 and TN = 80. Reconstruct the table and calculate Sp and the NPV.

Solution. b = 30, d = 80.

From Se $= \frac{a}{a+b} = \frac{2}{3}$: $\frac{a}{a+30} = \frac{2}{3}$, so 3a = 2a + 60, whence a = 60.

From PPV $= \frac{a}{a+c} = \frac{6}{8}$: $\frac{60}{60+c} = \frac{6}{8}$, so 6 x (60 + c) = 480, that is, 360 + 6c = 480, whence c = 20.

	T+	T-	Total
B+	60	30	90
B-	20	80	100
Total	80	110	190

Sp $= \frac{80}{100} = 0.80$. NPV $= \frac{80}{110} = 0.727$. Prevalence = 90/190 = 0.474; Acc $= \frac{140}{190} = 0.737$; ER $= \frac{50}{190} = 0.263$; $LR^+ = \frac{0.667}{0.20} = 3.33$; $LR^- = \frac{0.333}{0.80} = 0.42$; DOR $= \frac{60\times80}{30\times20} = 8$.

Remark. A sensitivity of 2/3 is 0.667, not 0.66. Rounding the fraction prematurely (to 0.66) leads to a = 58.2, that is, to a cell with a non-integer value, a sure sign of error. Work with the fractions until the end.

Worked Problem 12.4. Building the Table from Prevalence, Se and Sp

A rapid test has Se = 85% and Sp = 90%. Build the table for 500 patients with a disease prevalence of 20% and calculate the PPV and NPV. Repeat the calculation for 2000 people with a prevalence of 5%.

Solution. A prevalence of 20% out of 500: a + b = 100 diseased, c + d = 400 non-diseased. 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

PPV $= \frac{85}{125} = 68.0\%$; NPV $= \frac{360}{375} = 96.0\%$; Acc $= \frac{445}{500} = 89.0\%$; ER $= \frac{55}{500} = 11.0\%$.

At a prevalence of 5%, out of 2000: 100 diseased (a = 85, b = 15) and 1900 non-diseased (d = 0.90 x 1900 = 1710, c = 190). PPV $= \frac{85}{85+190} = \frac{85}{275} = 30.9\%$; NPV $= \frac{1710}{1725} = 99.1\%$.

Interpretation. Sensitivity and specificity have not changed. When the prevalence fell from 20% to 5%, the PPV dropped from 68.0% to 30.9%, and the NPV rose from 96.0% to 99.1%. Accuracy increased slightly (from 89.0% to (85 + 1710)/2000 = 89.8%), because specificity, which is higher than sensitivity, carries more weight when healthy people predominate; one more reason not to rely on accuracy.

Worked Problem 12.5. Reconstruction from Se, the NPV, FN and FP

The following are known: Se = 0.80, NPV = 0.95, FN = 20 and FP = 60. Reconstruct the table.

Solution. b = 20, c = 60. From Se $= \frac{a}{a+20} = 0.80$: a = 0.80a + 16, so 0.20a = 16, a = 80. From NPV $= \frac{d}{20+d} = 0.95$: d = 19 + 0.95d, so 0.05d = 19, d = 380.

	T+	T-	Total
B+	80	20	100
B-	60	380	440
Total	140	400	540

Sp $= \frac{380}{440} = 86.4\%$; PPV $= \frac{80}{140} = 57.1\%$; prevalence = 100/540 = 18.5%; Acc $= \frac{460}{540} = 85.2\%$; ER $= \frac{80}{540} = 14.8\%$; $LR^+ = \frac{0.80}{0.136} = 5.87$; $LR^- = \frac{0.20}{0.864} = 0.23$.

Worked Problem 12.6. Complete Evaluation of FINDRISC ≥ 12 on the mg2026 Dataset

Using Table 12.3, formulate the question, calculate the indicators with confidence intervals and interpret the result for a family physician working with a population in which the prevalence of undiagnosed diabetes is 5%.

Solution. (a) The research question: how well does the FINDRISC score, at the cutoff ≥ 12, identify adults with undiagnosed diabetes, compared with the OGTT? (b) Design: a cross-sectional diagnostic accuracy study in 160 adults at cardiometabolic risk, who underwent both tests. (c) Indicators: Se 75.9% (95% CI 57.9-87.8); Sp 74.8% (95% CI 66.7-81.5); PPV 40.0% (95% CI 28.1-53.2); NPV 93.3% (95% CI 86.9-96.7); accuracy 75.0%; LR+ 3.01 (95% CI 2.10-4.31); LR- 0.32 (95% CI 0.17-0.62); DOR 9.33. (d) At a prevalence of 5%: pre-test odds 0.0526; after a positive score, 0.0526 x 3.01 = 0.158, so the PPV is 13.7%; after a negative score, 0.0526 x 0.32 = 0.017, so the post-test probability is 1.7% (NPV 98.3%), exactly the values in Table 12.4.

Clinical interpretation. In a population with a prevalence of 5%, out of 100 patients with FINDRISC ≥ 12, about 14 have diabetes; the remaining 86 will undergo a laboratory test and receive a reassuring result. Out of 100 patients with a score below 12, about 98 do not have diabetes. The score is useful for selecting the patients who are worth testing, not for making the diagnosis. The wide confidence interval for sensitivity (57.9-87.8) shows that, in a sample with 29 diseased patients, we do not know precisely how many diabetic patients the score misses.

Worked Problem 12.7. The ROC Curve of Fasting Plasma Glucose, Constructed by Hand

On the mg2026 dataset, for fasting plasma glucose against OGTT-confirmed diabetes, the following were obtained at the cutoffs ≥ 126, ≥ 120, ≥ 115, ≥ 110, ≥ 105 and ≥ 100 mg/dL: TP 11, 14, 19, 24, 26, 29 (out of 29 diseased) and FP 7, 15, 28, 39, 54, 68 (out of 131 non-diseased). Construct the points of the ROC curve, calculate the area using the trapezoidal method and compare it with the AUC reported by the software (0.844).

Solution. We calculate $\text{Se} = \frac{\text{TP}}{29}$ and $1 - \text{Sp} = \frac{\text{FP}}{131}$.

Cutoff (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

We add (0; 0) and (1; 1) and calculate the trapezoids (width x mean height): 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. Sum: 0.836.

Interpretation. The approximate area, 0.836, is slightly below the exact value of 0.844, for the same reason as with FINDRISC: sparse points "cut the corners" of the curve. The maximum Youden index among the cutoffs in the table is at ≥ 110 mg/dL (0.828 + 0.702 - 1 = 0.530). At the diagnostic cutoff ≥ 126 mg/dL, sensitivity is only 37.9%: the clinical cutoff is a rule-in cutoff, not a screening cutoff.

Worked Problem 12.8. The Kappa Coefficient for Two Family Physicians

Two family physicians independently interpreted 100 electrocardiograms for the criteria of left ventricular hypertrophy. Both considered the criteria present in 20 ECGs and absent in 65. Physician A said "present" and physician B "absent" in 10 ECGs; the reverse occurred in 5. Calculate kappa.

Solution. The agreement table:

	B: present	B: absent	Total
A: present	20	10	30
A: absent	5	65	70
Total	25	75	100

Observed agreement: $\text{Po} = \frac{20+65}{100} = 0.85$.

Agreement expected by chance: A says "present" in 30% of cases, B in 25%; the probability that both say "present" by chance is 0.30 x 0.25 = 0.075. The probability that both say "absent" by chance is 0.70 x 0.75 = 0.525. Pe = 0.075 + 0.525 = 0.60.

$$\kappa = \frac{0.85-0.60}{1-0.60} = \frac{0.25}{0.40} = 0.625.$$

Interpretation. On the Landis and Koch scale, agreement is substantial (0.61-0.80). Note that the raw agreement of 85% looks excellent, but 60 percentage points of it would also have arisen by chance, given the high frequency of normal ECGs. Of the 15 disagreements, 10 go in the same

direction (A sees hypertrophy more often), which suggests a systematic difference in criteria between the two physicians that is worth discussing.

Worked Problem 12.9. Reconstructing the Table from an Article: CEUS-PAT for Cirrhosis

Lupuşoru and colleagues reported, in 37 patients with cirrhosis and 27 healthy volunteers, the following for CEUS-PAT at the cutoff > 0.7: Se 89.19%, Sp 100%, PPV 100%, NPV 87.1% [2]. Reconstruct the table and calculate LR+, LR-, accuracy and the PPV at a cirrhosis prevalence of 10% among patients with chronic liver disease.

Solution. a + b = 37, c + d = 27. From $\text{Se} = \frac{a}{37} = 0.8919$, a = 33, so b = 4. From $\text{Sp} = \frac{d}{27} = 1$, d = 27, c = 0. Check: $\text{PPV} = \frac{33}{33} = 100\%$; $\text{NPV} = \frac{27}{4+27} = \frac{27}{31} = 87.1\%$.

	CEUS-PAT > 0.7	CEUS-PAT ≤ 0.7	Total
Cirrhosis	33	4	37
Healthy	0	27	27
Total	33	31	64

$\text{Acc} = \frac{60}{64} = 93.8\%$. $\text{LR}^- = \frac{0.108}{1} = 0.11$. LR+ cannot be calculated (division by 0, because 1 - Sp = 0); formally, it is infinite. Also formally, the PPV would be 100% at any prevalence.

Interpretation. A specificity of 100% observed in 27 healthy volunteers does not mean that the test never gives false positives. The Wilson confidence interval for 27/27 has a lower limit of around 87.5%, so the true specificity could be considerably lower, especially in patients with non-cirrhotic liver disease, who are absent from the sample. Here, a PPV of "100% at any prevalence" is an illusion produced by a small sample and a favorable spectrum. The correct recommendation for such a result is validation in consecutive patients with chronic liver disease.

Worked Problem 12.10. Evaluating an Artificial Intelligence Classifier

An algorithm for detecting diabetic retinopathy was tested on 1000 retinal photographs, 80 of which showed retinopathy (confirmed by two ophthalmologists). The algorithm correctly classified 72 of the images with retinopathy and 874 of those without. Calculate the confusion matrix, sensitivity (recall), specificity, precision (PPV), the F1 score and accuracy. Compare the accuracy with that of an "algorithm" that declares all images normal.

Solution. TP = 72, FN = 8, TN = 874, FP = 920 - 874 = 46.

	Algorithm: retinopathy	Algorithm: normal	Total
Retinopathy	72	8	80
No retinopathy	46	874	920
Total	118	882	1000

Sensitivity (recall) = 72/80 = 90.0%. Specificity = 874/920 = 95.0%. Precision = 72/118 = 61.0%. $F_1 = \frac{2\times0.610\times0.900}{0.610+0.900} = \frac{1.098}{1.510} = 0.727$. Accuracy = (72 + 874)/1000 = 94.6%.

An "algorithm" that declares everything normal would have TP = 0, TN = 920, an accuracy of 920/1000 = 92.0%, a sensitivity of 0 and an F1 of 0.

Interpretation. The accuracy of 94.6% is only 2.6 percentage points above that of a useless classifier; it is not a sufficient indicator with imbalanced classes. The sensitivity of 90% and the specificity of 95% describe the algorithm better. The precision of 61.0% tells us that almost 4 in

10 referrals to the ophthalmologist will be false positives, a cost that must be weighed when planning the screening program.

12.12 Proposed Exercises

1. A test was applied to 200 patients: 45 diseased patients with a positive test, 5 diseased patients with a negative test, 30 non-diseased individuals with a positive test and 120 non-diseased individuals with a negative test. Build the table and calculate the prevalence, Se, Sp, PPV, NPV, Acc, ER, J, LR+, LR- and DOR.
2. For a test, the following are known: PPV $= \frac{4}{5}$, NPV $= \frac{9}{10}$, TP = 80 and FN = 20. Reconstruct the table and calculate Se, Sp and N.
3. For a test, the following are known: Sp $= \frac{3}{4}$, NPV $= \frac{15}{16}$, TN = 150 and TP = 40. Reconstruct the table and calculate Se and the PPV.
4. A test has Se = 75% and Sp = 95%. Build the table for 1000 patients with a prevalence of 10% and calculate the PPV and NPV.
5. Using the data from the study by Moleriu and colleagues (Worked Problem 11.3): 18 high-grade lesions and 46 without, HPV with 17 positives among the diseased and 27 negatives among the others, p16/Ki-67 with 12 and 39, build the two tables and calculate the accuracy and the Youden index for each test [131]. Which test is better suited for exclusion and which for confirmation?
6. At five cutoffs, a marker has the points (1 - Sp; Se): (0.05; 0.40), (0.10; 0.60), (0.20; 0.80), (0.40; 0.90), (0.70; 0.98). Calculate the area under the ROC curve using the trapezoidal method (add the points (0; 0) and (1; 1)) and classify it on the scale used in the lectures.
7. Two radiologists read 200 chest radiographs for pneumonia. Both said "pneumonia" for 30 and both said "no pneumonia" for 145; radiologist A said "pneumonia" and B "no pneumonia" for 15, and the reverse for 10. Calculate kappa and interpret it on the Landis and Koch scale.
8. Albai and colleagues reported for ISARIC-4C, in 159 patients with a mortality of 25.15%, Se 87.80% and Sp 73.72% [75]. (a) Approximately reconstruct the 2 x 2 table (number of deaths, TP, FN, FP, TN, rounded to whole numbers). (b) Calculate the PPV from the reconstructed table. (c) Calculate the approximate Wilson confidence interval for sensitivity and compare it with the reported interval (87.66-87.93).
9. An algorithm for classifying skin lesions was tested on 500 images, 50 of which were melanomas. The algorithm correctly identified 40 melanomas and 360 benign lesions. Calculate sensitivity, specificity, precision, F1 and accuracy, and comment.
10. Danila and colleagues reported an AUC of 0.624 for dNLR and 0.752 for ALRI in patients with COVID-19 and diabetes [105]. Classify the two values on the scale used in the lectures and explain why an AUC significantly different from 0.5 ($p < 0.001$) does not guarantee clinical usefulness.
11. Sipos and colleagues reported for IL-6 in neonatal sepsis Se 80.1% and Sp 85.7% [126]. Calculate the PPV and NPV for a sepsis prevalence of 10% among the tested newborns.
12. In the mg2026 dataset, fasting plasma glucose ≥ 126 mg/dL has Se 37.9% and Sp 94.7% against the OGTT. Calculate LR+ and LR- and explain why this cutoff is suitable for confirmation but not for screening.

Common Pitfalls

- Calculating the error rate with the formula (b + d)/N. Correct: ER $= \frac{b+c}{N} = \frac{\text{FN+FP}}{N} = 1 - \text{Acc}$.
- Swapping rows and columns: in some programs the disease is in the columns, and b and c swap places. Check what each cell contains.
- Reporting the PPV and NPV from a case-control study as clinical values; the prevalence in the sample is fixed by the design.
- Rounding fractions (2/3 = 0.66) before reconstructing the table.
- Considering accuracy a sufficient indicator, especially when the disease is rare.
- Interpreting overlapping AUC confidence intervals as proof of equivalence, or non-overlapping ones as proof of a difference; the comparison is made with the DeLong method.
- Choosing the cutoff on the same sample on which performance is reported, without validation.
- Reporting a specificity of 100% without a confidence interval, in a small sample of healthy controls.

Key Points

- Notation used in the lectures: a = TP, b = FN, c = FP, d = TN; Se $= \frac{a}{a+b}$, Sp $= \frac{d}{c+d}$, PPV $= \frac{a}{a+c}$, NPV $= \frac{d}{b+d}$, Acc $= \frac{a+d}{N}$, ER $= \frac{b+c}{N}$.
- Se and Sp do not depend on prevalence (in principle), but they do depend on the spectrum of patients; the PPV and NPV depend on prevalence and are recalculated with Bayes' theorem.
- SnNout: a highly sensitive test, if negative, rules out the disease; SpPin: a highly specific test, if positive, rules it in.
- J = Se + Sp - 1; $\text{LR}^+ = \frac{\text{Se}}{1-\text{Sp}}$; $\text{LR}^- = \frac{1-\text{Se}}{\text{Sp}}$; DOR $= \frac{a \times d}{b \times c}$.
- ROC curve: Se as a function of 1 - Sp for all cutoffs; AUC = the probability that a randomly chosen diseased person has a higher value than a randomly chosen healthy person.
- The scale used in the lectures: 0.91-1 excellent; 0.81-0.90 very good; 0.71-0.80 good; 0.61-0.70 satisfactory; below 0.6 poor. The AUC is reported with its 95% CI.
- In mg2026: FINDRISC AUC 0.852 (95% CI 0.762-0.943), fasting plasma glucose 0.844, nonsignificant difference (p = 0.81); Youden cutoff ≥ 11 (Se 93.1%, Sp 66.4%).
- The cutoff is chosen according to the purpose: exclusion (high Se and NPV), confirmation (high Sp and PPV), with a gray zone in between.
- Kappa measures agreement beyond chance for categories (Landis and Koch scale); the ICC is used for numerical measurements.
- AI classifiers are evaluated with the same indicators (recall = Se, precision = PPV, F1); reporting follows STARD 2015 and STARD-AI.

Self-Assessment Questions

1. (SC) The misclassification rate of a test, with the notation used in the lectures (a = TP, b = FN, c = FP, d = TN), is:
   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. (SC) A test with a sensitivity of 99% comes back negative in a patient with a moderate pre-test probability. The most appropriate conclusion is:

   a) the disease is confirmed
   b) the disease becomes very unlikely (SnNout)
   c) the result provides no information
   d) the specificity must be calculated to interpret a negative result, the sensitivity being irrelevant
   e) the positive predictive value is 99%

3. (SC) Which of the following indicators depends most directly on the prevalence of the disease in the tested population?
   a) sensitivity
   b) specificity
   c) the Youden index
   d) the positive predictive value
   e) the positive likelihood ratio

4. (SC) The area under the ROC curve of the FINDRISC score for diabetes is 0.852. The correct interpretation is:
   a) 85.2% of patients are classified correctly
   b) the sensitivity of the score is 85.2%
   c) a randomly chosen patient with diabetes has a higher score than a randomly chosen patient without diabetes in about 85.2% of pairs
   d) the positive predictive value is 85.2%
   e) 85.2% of patients with a positive score have diabetes

5. (SC) Two raters obtain a kappa coefficient of 0.55. On the Landis and Koch scale, agreement is:
   a) slight
   b) fair
   c) moderate
   d) substantial
   e) almost perfect

6. (MC) Regarding spectrum bias:
   a) it occurs frequently in case-control studies that compare typical patients with healthy volunteers
   b) it occurs when the patients in the study differ from those in whom the test will be used in practice
   c) it occurs when the reference standard is applied only to patients with a positive test
   d) it can change both sensitivity and specificity
   e) it is eliminated by increasing the number of patients in the study

7. (MC) Regarding the ROC curve:
   a) it has sensitivity on the vertical axis and 1 - specificity on the horizontal axis
   b) the diagonal corresponds to a test with no discriminative value (AUC = 0.5)
   c) each point on the curve corresponds to a decision cutoff
   d) the area under the curve depends directly on the prevalence of the disease in the sample
   e) the point with the maximum Youden index is the one farthest vertically from the diagonal

8. (MC) For the table TP = 22, FN = 7, FP = 33, TN = 98:
   a) the sensitivity is 75.9%

b) the positive predictive value is 40.0%

c) the misclassification rate is (7 + 98)/160

d) the diagnostic odds ratio is approximately 9.33

e) the specificity is 98/105

9. (MC) Regarding the evaluation of artificial intelligence classifiers:

a) precision is equivalent to the positive predictive value

b) recall is equivalent to sensitivity

c) accuracy is a sufficient indicator when the positive class is rare

d) the F1 score is the harmonic mean of precision and sensitivity

e) the STARD-AI guideline requires a description of how the training, validation and test data were separated

10. (MC) Regarding the comparison of two ROC curves:

a) for two tests performed in the same patients, the DeLong method takes into account the correlation between the AUCs

b) overlapping AUC confidence intervals prove that the tests are equivalent

c) the FINDRISC - glucose difference (0.008; p = 0.81) in the mg2026 dataset shows that no significant difference in discrimination was found

d) a nonsignificant difference between AUCs demonstrates that the tests are equivalent

e) two ROC curves can cross, so that each test is better at different cutoffs

Chapter 13. Clinical Studies: Classification, Design, Protocol, Data Collection Instruments and Integrated Analysis of Biomedical Data

Objectives

By the end of this chapter, the student will be able to:

- distinguish an interventional clinical study from an observational one and assign a study to a development phase (0-IV);
- recognize the design of a study (parallel-group, crossover, factorial, cluster, pragmatic or explanatory, single-arm pilot) and its consequences for the statistical analysis;
- explain simple, block and stratified randomization, allocation concealment and blinding, with the statistical argument behind each;
- define the primary and secondary endpoints, distinguish a surrogate endpoint from a "hard" clinical endpoint, and calculate the sample size for a mean or a proportion;
- compare intention-to-treat analysis with per-protocol analysis and explain why missing data are not simply "deleted";
- describe the structure of a protocol, the registration obligations and the ethical and legal framework (Declaration of Helsinki, GCP, EU Regulation 536/2014, GDPR);
- design a data collection form and a questionnaire, choose the appropriate question type for each variable, and interpret indicators of reliability and validity (Cronbach's alpha, ICC, KMO, factor analysis);
- build an electronic form in Google Forms, with validation rules, sections and branching, and clean a raw dataset;
- write the analysis plan of a randomized trial and apply it to the mg2026 dataset, from "Table 1" to the conclusion and limitations, reporting according to CONSORT 2025.

13.1 What a Clinical Study Is and What It Requires of Statistics

So far we have worked with data that had already been collected. We have described variables, estimated confidence intervals, tested hypotheses, measured associations and judged the performance of a diagnostic test. In this chapter we take a step back, to where data are born: in the design of a study. An impeccable statistical result calculated on poorly collected data remains a poor result. Conversely, a well-thought-out study makes the analysis almost mechanical.

A clinical study is research in humans in which the investigator assigns, according to a plan established in advance, an intervention (a drug, a procedure, an educational program, a mobile app) and follows its effect on certain health outcomes. The word that matters is "assigns". In an observational study, the researcher only watches: patients smoke or do not, are obese or are not, receive a treatment because their attending physician decided so. In an interventional study, exposure is decided by the protocol, ideally by drawing lots. This difference changes everything in the interpretation: from observational studies we draw conclusions about association, while from well-conducted randomized trials we allow ourselves causal conclusions [63,64].

Why does a clinical study need statistics from the very beginning? Because, before enrolling the first patient, we must know how many patients we need (Chapter 4), which hypothesis we are testing and at what significance level (Chapter 5), which test we will use for the primary endpoint (Chapters 6-8), how we will express the effect (difference in means, relative risk, NNT, Chapter

10) and how we will handle patients who drop out. All these decisions are written into the protocol and the statistical analysis plan. If we make them after seeing the data, we risk choosing, consciously or not, the option that "comes out" significant. The methodological guidelines of our discipline insist on exactly this point: the plan precedes the data [18,111].

In the hierarchy of evidence of evidence-based medicine, the randomized controlled trial (RCT) stands above cohort, case-control and cross-sectional studies, and below the systematic synthesis of several RCTs. The hierarchy is not, however, a rigid ladder. A small RCT with many dropouts and no blinding may provide weaker evidence than a large, carefully adjusted cohort. This is why a physician reading an article must judge the actual quality of the study, not just the label on the cover. Modern systems for grading the certainty of evidence start from the design, but then lower or raise the level according to the risk of systematic error, imprecision and inconsistency [128].

Throughout the chapter we will rely on the mg2026 teaching dataset. It simulates a pragmatic randomized trial in 12 family practices: 160 adults at cardiometabolic risk, allocated to three arms (A, standard counseling; B, counseling plus a self-monitoring mobile app; C, counseling, the app and a monthly teleconsultation), followed up at 0, 3 and 6 months. We will look at it in turn through the eyes of the person who designs the study, the person who collects the data and the person who analyzes them. At the end, we will write the integrated analysis plan together.

13.2 Classification of Clinical Studies

13.2.1 Interventional or Observational

The first question to ask when you open an article is: who decided the exposure? If the investigator decided, the study is interventional (experimental). If the exposure occurred independently of the study, it is observational: cohort, case-control, cross-sectional (Chapter 10, Figure 10.1). Observational studies can be prospective or retrospective, analytical or descriptive, but none of them has the power of randomization to balance unknown confounders.

There are also intermediate forms. A nonrandomized study with two groups, in which patients chose the type of monitoring themselves or were allocated according to equipment availability, is interventional in the sense that the intervention is applied according to a protocol, but it resembles a cohort as far as the comparability of the groups is concerned. A before-after study without a control group is also interventional, but it cannot separate the effect of treatment from the natural course of the disease, regression to the mean or the placebo effect. We will encounter all of them in the case studies in this chapter.

13.2.2 The Phases of Drug Development

For drugs, clinical studies follow a standardized path, shown in Figure 13.1. Before human studies there is the preclinical stage (laboratory, animal models). Sometimes a phase 0 is added, with microdoses given to a very small number of volunteers, to check whether the substance behaves in the human body as the preclinical studies suggested; it has no therapeutic purpose and does not answer questions about efficacy.

Phase I usually includes 20-80 volunteers (healthy or, in oncology, patients) and examines safety, tolerability, pharmacokinetics and the maximum tolerated dose. The statistics are predominantly descriptive: frequencies of adverse effects, means and ranges for concentrations. Phase II, with 100-300 patients, looks for signals of efficacy and the optimal dose; this is where the first comparisons with a control group appear, along with the first estimates of effect size, which are then used to calculate the sample size for phase III. Phase III is the large randomized trial, with hundreds or thousands of patients, which compares the new treatment with the standard

treatment or with placebo on a relevant clinical endpoint; marketing authorization is granted on the basis of its results. Phase IV begins after authorization and monitors rare adverse effects, use in populations not included in earlier studies, and effectiveness in real-world practice (pharmacovigilance).

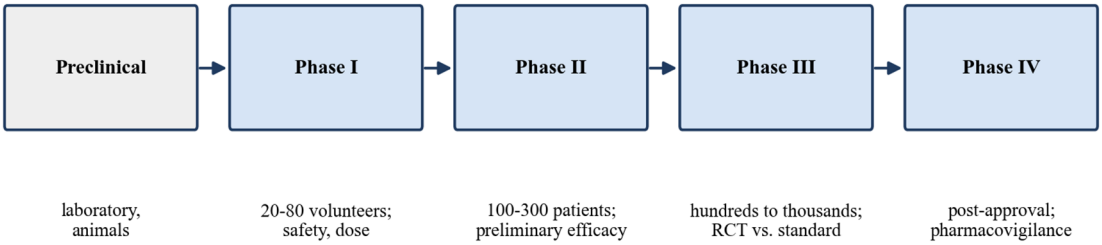


Figure 13.1. The stages of drug development: preclinical (laboratory, animals), phase I (20-80 volunteers, safety and dose), phase II (100-300 patients, initial efficacy), phase III (hundreds to thousands of patients, RCT versus standard treatment), phase IV (after authorization, pharmacovigilance).

The statistical logic of the phases is that of a funnel. An adverse effect with a frequency of 1 in 1000 is quite likely not to appear at all in a phase III study with 600 patients. The approximate "rule of 3" states that if we have observed no events in n patients, the upper limit of the 95% confidence interval for their frequency is approximately 3/n; for n = 600, this means 0.5%. The absence of an adverse effect in the authorization studies does not prove that it does not exist, only that it is probably rarer than that. Hence the need for phase IV.

Studies that do not test drugs (educational programs, mobile apps, surgical interventions, devices) do not always follow this numbering, but they follow the same progression: a feasibility or pilot study, then a large comparative study. The mg2026 scenario would correspond to a phase III-type study of a nonpharmacological intervention: it compares strategies already considered safe, on an efficacy endpoint, under conditions close to real-world practice.

13.2.3 By Objective: Superiority, Non-Inferiority, Equivalence

Most of the studies you will read are superiority studies: they aim to show that the new treatment is better than the comparator. The hypotheses are those of Chapter 5: H0, there is no difference; H1, there is a difference. Rejecting H0 supports superiority (if the difference is in the expected direction).

Sometimes we do not want a more effective treatment; we want one that is just as effective but cheaper, more convenient or with fewer adverse effects. We then design a non-inferiority study. We set in advance a non-inferiority margin Δ, the largest loss of efficacy we consider clinically acceptable. The new treatment is declared non-inferior if the limit of the confidence interval of the difference (new minus standard) does not go beyond $-\Delta$. An example: if teleconsultation were to replace in-person visits, we might accept a reduction in blood glucose that is at most 3 mg/dL smaller than with in-person visits; if the 95% CI of the difference is, say, -2.1 to 4.5 mg/dL, the lower limit of -2.1 is above -3 and we conclude non-inferiority. In an equivalence study, we require the entire interval to lie between $-\Delta$ and $+\Delta$.

Note that here it is the confidence interval, not the p value, that decides. A superiority study with a nonsignificant result does not demonstrate equivalence: "absence of evidence is not

evidence of absence" [79]. If the study was small, the interval is wide and includes both important effects in favor of the treatment and important effects against it. Only a study designed from the outset for non-inferiority, with a clinically justified margin and a sufficient sample, can support such a conclusion.

13.2.4 By Design

Figure 13.2 summarizes four designs you will encounter frequently. Each has a direct consequence for the choice of statistical test.

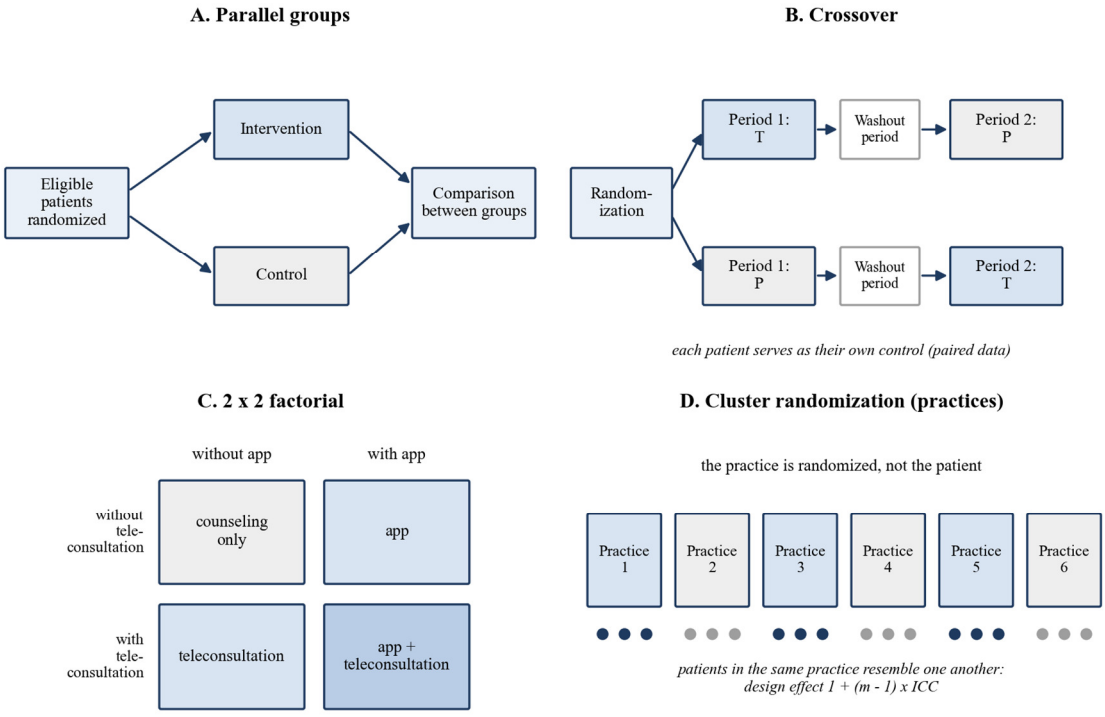


Figure 13.2. Four clinical study designs. A: parallel groups; each patient receives a single intervention and the groups are compared with each other. B: crossover design; each patient receives both treatments in random order, separated by a washout period, so the data are paired. C: 2 x 2 factorial design; two interventions tested simultaneously. D: cluster randomization; the practice is randomized, and patients in the same practice resemble one another.

Parallel groups. Each patient is allocated to a single arm and stays there until the end. This is the design of the mg2026 study and of most RCTs. The groups are independent, so we use the independent-samples t-test or the Mann-Whitney test for two arms, ANOVA or Kruskal-Wallis for three or more arms, and the chi-square test for proportions.

Crossover design. Each patient receives both treatments in turn, with the order decided by lot; a washout period is left between the periods so that the effect of the first treatment can wear off. Each patient serves as his or her own control, between-patient variability disappears from the comparison, and the required sample shrinks considerably. The BPOC file (COPD data) used in the practical sessions of our discipline comes from a study of this type, with measurements at the start of the study and after medication. The analysis uses tests for paired data: the paired t-test or the Wilcoxon test. The design works only for stable chronic diseases and for treatments with a reversible effect; it cannot be used for an intervention that cures the disease or for an endpoint such as death. The specific risk is the carry-over effect: if the washout period was too short, the treatment from the first period contaminates the second.

Factorial design. It tests two (or more) interventions simultaneously. In a 2 x 2 design, patients are randomized into four groups: nothing, A only, B only, A and B. With the same

number of patients, we answer two questions. The mg2026 study is not factorial: arm C adds teleconsultation on top of the app, but there is no "teleconsultation only" arm. This is why we cannot estimate the effect of teleconsultation without the app. A factorial design would have made this possible and would have allowed testing for an interaction (whether or not the app and teleconsultation enhance each other).

Cluster randomization. Sometimes we cannot randomize individual patients. If the family physician learns to carry out teleconsultations, he or she will tend to apply them to all of his or her patients, including those allocated to standard counseling; contamination between arms occurs. The solution is to randomize the practices (clusters): all patients in a practice receive the same intervention. The mg2026 study could have been designed this way, with the 12 practices randomized 4 to each arm. The price is statistical. Patients in the same practice resemble one another (same physician, same area, same counseling style), so they do not provide independent information. The intraclass correlation (ICC) measures this similarity, and the design effect shows by how much the sample must be increased:

design effect = 1 + (m - 1) x ICC

where m is the average number of patients per cluster. With 13 patients per practice and ICC = 0.05, the design effect is 1 + 12 x 0.05 = 1.6: we need 60% more patients than in an individually randomized trial. The analysis must take the clusters into account; if we ignore them, the standard errors come out too small and the p values too optimistic.

Pragmatic or explanatory. An explanatory trial asks "can the intervention work under ideal conditions?": strict inclusion criteria, motivated patients, closely monitored adherence, often a single specialized center. A pragmatic trial asks "does the intervention work in routine practice?": broad criteria, ordinary practices, adherence as close to reality as possible, patient-relevant outcomes. The mg2026 study is pragmatic, which is why its effects are probably smaller than they would have been in an explanatory trial, but closer to what the family physician will see. Most real trials fall somewhere between the two extremes.

The single-arm pilot study. Before an expensive RCT, researchers check feasibility: can the patients be recruited, do they accept the intervention, are the measurement instruments applicable, how large does the effect appear to be? A single-arm pilot study applies the intervention to everyone and compares the values over time. The analysis uses tests for paired data or for repeated measurements (paired t-test, Wilcoxon, repeated-measures ANOVA, Friedman). The limitation is that of any before-after study: without a control group, we cannot attribute the improvement to the intervention with certainty.

Case study: a mobile app in patients with asthma and a single-arm pilot study in atopic dermatitis

Two studies by colleagues in our discipline illustrate the extremes described above. Munteanu and colleagues tested a mobile app for treatment adherence in 93 patients with asthma, in a controlled pilot study with two groups (standard treatment versus treatment plus the app) and repeated measurements over one year [94]. The design bears a striking resemblance to arms A and B of mg2026. The analysis respects the structure of the data: the independent groups were compared with the Mann-Whitney test ($p < 0.001$), change over time was analyzed with the Friedman test for dependent samples, with a Bonferroni correction for multiple comparisons ($p < 0.001$), and the proportion of exacerbations was compared with the chi-square test: 10.30% versus 46.30%, $\chi^2(1) = 13.707$, $p < 0.001$. The question we ask when reading is how the patients were allocated to the two groups; if they were not randomized, the difference may also reflect the motivation of those who accepted the app.

> At the other end, Miulescu and colleagues followed 21 children with atopic dermatitis and laboratory-confirmed dysbiosis in a prospective single-arm pilot study, with assessments at enrollment, at 30 and at 90 days [91]. The POEM score fell from 14.67 ± 5.22 to 7.10 ± 3.82, and SCORAD from 41.66 ± 15.28 to 17.93 ± 11.68 (both $p < 0.001$), analyzed with repeated-measures ANOVA, the appropriate test for three time points in the same children. About 76% of the children achieved a reduction of at least 50% in POEM. The result is promising, but the design does not allow a causal conclusion: atopic dermatitis has a fluctuating course, children are often enrolled at a peak of the disease (regression to the mean), and parents who know that a new therapy has been applied may report symptoms more favorably. The correct conclusion is the one stated in the title itself: a pilot study that justifies an RCT.

13.3 The Methodological Principles of a Randomized Trial

13.3.1 Randomization

Randomization means allocating patients to arms by a random mechanism that neither the patient nor the physician can influence. Its effect is statistical: in the long run, all patient characteristics, measured or not, known or not, are distributed on average in the same way across the arms. The differences that remain between the groups at enrollment are due to chance, and chance is exactly what statistical tests know how to quantify. This is why a significant effect in a well-conducted RCT can be attributed to the intervention.

Simple randomization allocates each patient independently, as when tossing a coin (or rolling a die, for three arms). In large samples it works well. In small samples it can produce size imbalances: with 30 patients and two arms, the probability that one arm receives 10 patients or fewer is not negligible. Block randomization corrects this. We choose blocks of, for example, 6 patients for three arms and, within each block, randomly permute the sequence AABBCC. One block might be B A C C A B, the next C B A B A C. At the end of each block, the arms have exactly the same number of patients. The disadvantage: if the physician knows the block size, he or she can guess the last allocation in the block. This is why blocks of variable size are used (3, 6 or 9, themselves chosen at random).

Stratified randomization is used when an important prognostic factor must be balanced reliably, not just on average. A separate randomization list (usually in blocks) is created for each stratum: for example, for each practice, or for confirmed diabetes yes/no. In mg2026, stratification by practice would guarantee that each physician has patients in all three arms, so the "physician effect" is not confounded with the effect of the intervention. We do not stratify by too many factors: with 12 practices, two sexes and three smoking categories, we get 72 strata, many with a single patient, and stratification loses its point.

13.3.2 Allocation Concealment

The randomization sequence can be perfect and the study can still be compromised. If the physician enrolling patients sees the list and knows that the next patient will go to arm C, he or she may postpone enrolling a "difficult" patient until an allocation that suits him or her comes up. Allocation concealment means that a patient's allocation becomes known only after the patient has been definitively enrolled. Secure methods are centralized randomization, by telephone or through a web platform, and opaque, sealed, sequentially numbered envelopes (more vulnerable, but acceptable if managed correctly). It should not be confused with blinding: allocation concealment protects the moment of enrollment and is always possible; blinding protects the period after allocation and is not always possible.

13.3.3 Blinding

Blinding means that one or more of those involved in the study do not know which intervention each patient received: the patient, the treating physician, the person assessing the outcome, the statistical analyst. The terms "single-blind" and "double-blind" are used inconsistently in the literature, which is why the current recommendation is to state explicitly who was blinded, not just the label [142]. A patient who knows that he or she is receiving the active treatment may report better symptoms; a physician who knows may interpret an ambiguous clinical result more favorably; an assessor may measure with different degrees of attention.

In the mg2026 study, patients cannot be blinded: they know whether they received the app and the teleconsultations. Nor can the family physicians. What can be done is to have blood glucose measured by a laboratory that does not know the arm, and to give the analyst the data with the arms coded (X, Y, Z). Objective endpoints (laboratory blood glucose, weight on a calibrated scale) are less vulnerable to lack of blinding than subjective endpoints (satisfaction, pain, quality of life). Satisfaction in mg2026 must therefore be interpreted with caution: patients in arm C received more attention and are, naturally, more satisfied.

13.3.4 The Control Group and Placebo

Without a control group, we cannot separate the effect of the intervention from the natural course of the disease, regression to the mean, the placebo effect and the effect of simply taking part in a study (patients who are observed change their behavior, a phenomenon called the Hawthorne effect). The control can be placebo (an inactive substance identical in appearance), standard treatment, no intervention, or a waiting list. Ethics limits the use of placebo: when a proven effective treatment exists, patients in the control group must receive it, and the new treatment is added on top of it or compared with it. In mg2026, arm A receives standard counseling; it would be unethical to leave patients with prediabetes without any lifestyle advice.

13.3.5 Endpoints

The primary endpoint is the variable on which the conclusion of the study is based and for which the sample size was calculated. There must be only one, precisely defined (what is measured, how, when) and chosen before the study begins. In mg2026 it is the reduction in fasting plasma glucose between enrollment and 6 months (glicemie0 - glicemie6). Secondary endpoints provide additional information: weight loss of at least 5%, reduction in blood pressure, satisfaction. Their results are considered supportive or hypothesis-generating, because testing many endpoints increases the probability that one of them will come out significant purely by chance. With 10 independent endpoints, each tested at $\alpha = 0.05$, the probability of obtaining at least one false-positive result is $1 - 0.95^{10} = 0.40$.

A surrogate endpoint is a biological marker that stands in for an important clinical outcome: blood glucose or HbA1c instead of diabetic complications, LDL cholesterol instead of myocardial infarction, blood pressure instead of stroke. Surrogates allow smaller and shorter studies, but the history of medicine has plenty of examples of drugs that improved the surrogate and worsened the clinical outcome. "Hard" endpoints (death, myocardial infarction, stroke, hospitalization) are the ones that matter to the patient, but they require large samples and long follow-up. A composite endpoint combines several events (for example, cardiovascular death, myocardial infarction or stroke) to increase the number of events; it must be read carefully, because the effect may be driven almost entirely by the least serious component.

13.3.6 Sample Size and Power

We learned in Chapter 4 that sample size depends on four elements [42,53]: the significance level α (usually 0.05), the desired power $1 - \beta$ (usually 80% or 90%), the effect size we want to detect

and the variability of the endpoint. For comparing two means, the approximate formula per group is

$$n = \frac{2 \times \left(z\left(1 - \frac{\alpha}{2}\right) + z(1 - \beta) \right)^2 \times \sigma^2}{\Delta^2}$$

where Δ is the minimal clinically relevant difference and σ the standard deviation. Expressed in terms of the standardized effect size $d = \frac{\Delta}{\sigma}$, we have $n = \frac{2 \times (1.96 + 0.84)^2}{d^2}$. For d = 0.5, the formula gives approximately 63, and the exact calculation with the t distribution, performed in programs such as G*Power [77], gives 64 patients per group; for d = 0.3, 176 per group. For two proportions p1 and p2:

$$n = \frac{\left[z\left(1 - \frac{\alpha}{2}\right) \times \sqrt{2 \times \bar{p} \times (1 - \bar{p})} + z(1 - \beta) \times \sqrt{p_1 \times (1 - p_1) + p_2 \times (1 - p_2)} \right]^2}{(p_1 - p_2)^2}$$

with $\bar{p} = \frac{p_1 + p_2}{2}$. To the result obtained we add an allowance for dropouts: if we estimate that 10% of patients will be lost, we divide n by 0.9. Worked Problem 13.1 shows the complete calculation for mg2026.

A study that is too small is useless and, worse, unethical: it exposes patients to the risks of research without a reasonable chance of answering the question. A nonsignificant result in a low-powered study tells us almost nothing about the absence of an effect.

13.3.7 Intention to Treat and Per-Protocol Analysis

In any real study, some patients do not follow the protocol: they do not install the app, do not attend the teleconsultations, switch arms, drop out. We have two main analysis options. Intention-to-treat (ITT) analysis analyzes each patient in the arm to which he or she was randomized, regardless of what was actually received. Per-protocol (PP) analysis keeps only the patients who followed the protocol.

ITT preserves the benefit of randomization. Patients who do not adhere to treatment are not a random subset: they are usually sicker, less motivated or have adverse effects. If we remove them from the active arm but not the comparable patients from the control arm, the remaining groups are no longer comparable. ITT answers the practical question "what happens if I recommend this strategy?", and it tends to underestimate the biological effect, because it dilutes the effect with non-adherent patients. In superiority studies, this underestimation is conservative, so ITT is the primary analysis. In non-inferiority studies, the dilution would favor a conclusion of non-inferiority, which is why both analyses are required, in the hope that they agree. In mg2026 there are no losses to follow-up (Figure 13.3), so ITT coincides with the analysis of all 160 patients; with real data, things are never like this.

13.3.8 Missing Data and Losses to Follow-up

A patient who does not come to the 6-month visit leaves an empty cell in the table. The most convenient solution, complete-case analysis (deleting the patient), is correct only if the data are missing completely at random, that is, if the missingness is related neither to the patient's characteristics nor to the outcome. This is rarely the case. A patient who has not lost weight and feels guilty stops coming to follow-up visits; if we delete him or her, that patient's arm appears more effective than it is.

Three mechanisms are described. Missing completely at random: data are missing for reasons unrelated to the study (a broken test tube). Missing at random conditional on observed variables: the probability of missingness depends on things we have measured (for example, older people

are missing more often), and we can correct for it using those variables. Missing not at random: the probability of missingness depends on the missing value itself (those with high blood glucose stop coming); here no method corrects completely. Modern methods (multiple imputation, mixed models) are beyond the level of this course. What you should remember: report how much data are missing in each arm and why; do not replace a missing value with the group mean (it artificially reduces variability); replacing it with the last observed value is an old solution, which assumes that the patient remained "frozen" and can distort the result; perform a sensitivity analysis, with different assumptions about those lost, and check whether the conclusion remains the same.

13.3.9 Interim Analyses

In long studies, an independent data and safety monitoring committee analyzes the data at prespecified time points. It can stop the study early for clear benefit, for adverse effects or for futility (when it becomes unlikely that the study will ever show an effect). Each interim analysis is one more statistical test and, if we used the 0.05 threshold each time, the overall risk of a type I error would increase. This is why much stricter thresholds are used at the interim analyses (on the order of $p < 0.001$), keeping a threshold close to 0.05 for the final analysis. Studies stopped early for benefit tend to overestimate the effect, because they stopped precisely when chance had produced a large difference.

Case study: a placebo-controlled trial reported according to CONSORT and an underpowered nonrandomized cohort

Gherbon and colleagues compared, over 3 months, 20 people with obesity who received diet and Aloe ferox (2 capsules per day, 460 mg) with 20 matched people who received diet and placebo [85]. The study is explicitly reported according to CONSORT, which means that the reader finds in the article the flow of patients, the allocation method and the definition of the endpoints. The abstract reports two series of p values: for BMI ($p = 0.03$), total cholesterol ($p = 0.032$), LDL ($p = 0.01$) and fasting plasma glucose ($p = 0.018$), and for BMI ($p = 0.002$), LDL ($p = 0.039$), fasting plasma glucose ($p < 0.001$) and diastolic blood pressure ($p = 0.006$). When reading, the first thing to clarify is which series corresponds to the between-group comparison at 3 months and which to the before-after comparison in the treated group. Only the between-group comparison tests the effect of Aloe ferox; the before-after comparison in the active group also includes the effect of the diet, which both groups received. Second: with 20 patients per group, the study is powered for large effects; many endpoints tested simultaneously increase the risk of a false-positive result, so the primary endpoint must be identified in the protocol.

Popovici and colleagues compared, in a prospective cohort of 100 patients undergoing ERCP, sedation guided by entropy monitoring with standard monitoring [78]. The groups were not randomized. Hypertensive episodes were less frequent in the entropy group (6% versus 26%, $p = 0.007$), and Aldrete scores were better at 5 minutes ($p = 0.003$) and at 15 minutes ($p < 0.001$). The authors honestly report that statistical power was limited (29.3%). The lesson is twofold: a power of 29.3% means that, if the effect sought did exist, the study would have only about a one-in-three chance of detecting it, so a nonsignificant result for the primary endpoint does not rule out the effect; and the absence of randomization leaves room for confounders (who decided which monitoring the patient received?). The results are, correctly stated, hypothesis-generating for an RCT.

13.4 The Protocol, Registration and the Ethical Framework

13.4.1 The Protocol

The protocol is the document that describes the study before it begins. It works as a contract between the researchers, the patients, the ethics committee and future readers: everything done in the study must have been announced there, and any deviation must be justified. For clinical trial protocols there is an international content guideline, SPIRIT, built as a mirror image of the CONSORT reporting guideline: what is promised in the protocol is then reported in the article. Colleagues in our discipline have also published study protocols, for example the protocol for translating three health literacy questionnaires into Romanian in the ALSATION study [26], which shows that protocols are not reserved for drug trials.

A complete protocol contains, in essence, the following elements: the title and the registration identifier; the rationale (what is known, what is not known, why the study is needed); the objectives and hypotheses, formulated in PICO format (population, intervention, comparator, outcome) [24]; the design (parallel, crossover, cluster; superiority or non-inferiority); the setting; the inclusion and exclusion criteria; the interventions, described precisely enough for another center to reproduce them; the primary and secondary endpoints, with the time of measurement; the visit schedule; the sample size calculation, with all its assumptions; the method of randomization, allocation concealment and blinding; the methods of data collection and data management; the statistical analysis plan (the ITT and PP analysis populations, the tests, the handling of missing data, the prespecified subgroup analyses, the interim analyses); safety monitoring; ethical aspects, consent and confidentiality; the publication and data-sharing policy.

For a first-year student, the most useful part to practice is translating the clinical question into statistical elements. In mg2026: P, adults aged 35-79 years at cardiometabolic risk in family practices; I, counseling plus the app (B) or counseling plus the app plus teleconsultation (C); C, standard counseling (A); primary O, the reduction in fasting plasma glucose at 6 months. The main statistical hypothesis becomes H0: $\mu A = \mu B = \mu C$ for the reduction in blood glucose, and the planned test is one-way ANOVA with Tukey comparisons, or Kruskal-Wallis if the ANOVA assumptions are not met.

13.4.2 Study Registration

A clinical study is registered in a public registry before the first patient is enrolled. The best known are ClinicalTrials.gov (the US registry, also used by many European researchers) and, for drug trials in the European Union, the CTIS system (Clinical Trials Information System), introduced by EU Regulation 536/2014 [143]. Journals that follow the ICMJE recommendations do not publish the results of a clinical trial that was not prospectively registered [36]. The reason is statistical, not bureaucratic: registration publicly fixes the primary endpoint. If, at the end, the authors report as "primary" an endpoint that was initially secondary, because it came out significant, comparison with the registry reveals the change. Studies with negative results, which would otherwise remain unpublished, also become visible, and systematic syntheses are no longer distorted by publication bias.

13.4.3 The Ethics of Research in Human Subjects

The ethical foundation of medical research is the Declaration of Helsinki of the World Medical Association, last revised in 2024 [35]. Its principles can be summarized as follows: the well-being and rights of the participant take precedence over the interests of science and society; the risks must be proportionate to the expected benefits; research must be scientifically justified and conducted by qualified persons; the protocol is approved by an independent ethics committee before the study begins; participation is voluntary and based on informed consent; vulnerable

groups receive additional protection; the use of placebo is permitted only under strict conditions; the results, including negative ones, must be made public.

Informed consent is a process, not just a signature. The patient receives information about the purpose of the study, the procedures, the possible risks and benefits, the alternatives, the right to withdraw at any time without consequences for his or her medical care, and how his or her data will be protected. The information is given in language the patient understands, with time to think and the opportunity to ask questions. For minors, parental consent is required and, depending on age, the child's assent.

The research ethics committee (in Romanian hospitals and universities, the institution's ethics committee) evaluates the protocol, the documents for patients, the competence of the team and the scientific justification. From a statistical point of view, the committee also checks the sample size: a study without sufficient power exposes patients to risks without the prospect of a useful result.

Good Clinical Practice (GCP) is the international standard for the design, conduct, recording and reporting of clinical trials; the current version of the ICH E6 guideline, revision 3, emphasizes quality built in from the design stage and a risk-based approach [144]. In the European Union, clinical trials of medicinal products are governed by Regulation 536/2014, which harmonized the authorization procedure and introduced the single CTIS portal [143]. The processing of personal health data in research falls under the General Data Protection Regulation (GDPR), which treats data concerning health as a special category that may be processed for research purposes only with appropriate safeguards, such as pseudonymization and data minimization [32], supplemented in Romania by Law 190/2018 [33].

For the analyst, the GDPR has practical consequences. The analysis file does not contain names, personal numeric codes (CNP), addresses or telephone numbers, but a study code (in mg2026, P001-P160). The code-identity correspondence table is kept separately, by the principal investigator, with restricted access. Only the variables needed for the research question are collected: if age is sufficient, we do not ask for the date of birth. Files are not sent by unencrypted e-mail and are not kept on USB sticks forgotten in pockets. A Google Forms form used in a real study with patients must be checked with regard to where the data are stored and to the institution's approval; for the exercises in the practical sessions, we use fictitious data.

13.5 Data Collection Instruments

13.5.1 The Case Report Form

The case report form (CRF) is the document, on paper or electronic, in which each patient's data are recorded at each visit. A good CRF is designed starting from the analysis plan: each field corresponds to a variable in the data dictionary, and each variable in the analysis plan has a field in the CRF. Nothing extra, nothing missing.

A few design rules prevent most of the errors we will see in Section 13.6. Each numerical field has its unit of measurement printed next to it ("height: ___ cm"), so that we do not receive 1.68 instead of 168. Categorical fields have predefined options that can be ticked, not free text; otherwise we will end up with "f", "F", "Feminin" and "femeie" for the same value. Dates have a single format (dd.mm.yyyy). For each field, a way of recording why the value is missing is provided (not measured, patient refused, not applicable), so that an empty cell is not ambiguous. Derived variables (BMI, percentage weight loss, the FINDRISC score) are not calculated by hand in the CRF, but in the analysis file, from the primary variables. The first field on each page is the patient code.

13.5.2 The Questionnaire: From Question to Variable

Many studies in general medicine use questionnaires: symptoms, quality of life, adherence, satisfaction, knowledge, eating behaviors. Designing a questionnaire is a craft, and its first lesson is that the type of question determines the type of variable and, through it, the statistical test (Chapters 2 and 8). Table 13.1 makes this connection.

Table 13.1. Types of questions and the variables they produce.

Type of question	Example	Resulting variable	Description and usual tests
Free numerical answer	"How old are you? ___ years"	numerical (continuous or discrete)	mean and SD or median and IQR; t-test, ANOVA, correlation
Single choice, two options	"Do you currently smoke? Yes / No"	nominal dichotomous	frequencies, %; chi-square, Fisher, RR, OR
Single choice, unordered options	"Area of residence: urban / rural"	nominal	frequencies; chi-square
Single choice, ordered options	"Education: lower secondary / upper secondary / university / postgraduate"	ordinal	median, IQR; Mann-Whitney, Kruskal-Wallis, Spearman
Likert item	"I am satisfied with the program: 1 not at all... 5 very"	ordinal	median, frequencies; nonparametric tests
Score from several items	sum of 10 Likert items	approximately numerical	mean or median, depending on the distribution
Multiple choice	"Which chronic diseases do you have? (tick all that apply)"	one dichotomous variable per option	frequency of each option
Free text	"Other comments"	qualitative	thematic analysis; not included in statistical tests

Two mistakes appear year after year in student projects. The first: age or weight collected in classes ("18-25 years, 26-35 years..."). We have turned a numerical variable into an ordinal one and lost information; we can no longer calculate the mean or the correlation. The rule: collect the exact value and group it, if necessary, at the analysis stage. The second: double-barreled questions ("Do you exercise and eat healthily?"); a patient who exercises but eats poorly has no answer to give.

The wording of questions also requires care. Questions should be short and neutral (not "Isn't it true that the app helped you?"), without double negatives, with a clear reference period ("in the last 7 days"), and with response options that are exhaustive and do not overlap ("0-1 hours, 1-2 hours" overlap at the value 1). Sensitive questions (income, alcohol consumption, sexual life) are asked toward the end, once trust has been established.

13.5.3 Likert Scales

A Likert scale presents a statement and an odd number (usually 5 or 7) of ordered options expressing agreement or disagreement, with a neutral point in the middle. The variable satisfactie (satisfaction) in mg2026 is a 5-point Likert item. A single Likert item is an ordinal variable: the distance between "1, not at all satisfied" and "2" is not necessarily equal to the distance between "4" and "5, very satisfied". This is why we describe it with the median and frequencies and compare it with nonparametric tests, as we did in Chapter 7 (Mann-Whitney C versus A: medians 4 versus 3, $U = 2265.5$, $p < 0.001$). A score obtained by summing several Likert items that

measure the same construct has more possible values and a distribution often close to normal; in practice it is treated as a numerical variable, after checking the distribution.

Reverse-worded items ("I find it hard to use the app", in a scale measuring ease of use) are recoded before summing: 1 becomes 5, 2 becomes 4, and so on. Forgetting to recode is one of the most common errors, and it is easy to spot, because the item in question has a negative correlation with the total score.

13.5.4 The Reliability of an Instrument

A measurement instrument is reliable if it gives consistent results. Two types of reliability are used for questionnaires.

Internal consistency shows whether the items of a scale measure the same thing. The classic indicator is Cronbach's alpha coefficient:

$$\alpha = \frac{k}{k-1} \times \left(1 - \frac{\sum s^2(\text{item})}{s^2(\text{ total })}\right)$$

where k is the number of items, s²(item) the variance of each item, and s²(total) the variance of the total score. If the items are strongly correlated with one another, the variance of the total is much larger than the sum of the item variances, and α approaches 1. As a rough guide, α ≥ 0.70 is considered acceptable for research, ≥ 0.80 good, and ≥ 0.90 very good for individual decisions; values above 0.95 may signal redundant items, the same question asked several times in different words. McDonald's omega coefficient is an alternative that makes less restrictive assumptions than alpha and is increasingly reported alongside it. The corrected item-total correlation (the correlation of each item with the sum of the others) identifies "weak" items; a common threshold is 0.30.

Test-retest reliability shows whether the same patient, in the same condition, gives the same answers on two administrations separated by an interval (usually 1-3 weeks). For numerical scores, the intraclass correlation coefficient (ICC) is used, not the Pearson coefficient. The reason: Pearson's r measures only whether the values increase together; if on the second administration all patients score 5 points higher, r can be 1 even though agreement is poor. The ICC also penalizes such systematic differences. As a guide, an ICC below 0.50 indicates poor reliability, 0.50-0.75 moderate, 0.75-0.90 good and above 0.90 excellent. For categorical variables, agreement between two administrations or between two raters is measured with the kappa coefficient [103].

13.5.5 The Validity of an Instrument

A reliable instrument is not automatically valid. A scale that consistently shows 5 kg too much is reliable, but not valid. Validity shows whether the instrument measures what it claims to measure.

Content validity is judged by experts and patients: do the items cover all the important aspects of the construct, and are they worded intelligibly? Construct validity checks whether the scores behave as theory predicts, in several ways. Convergent validity: the score correlates strongly with other measures of the same construct. Discriminant validity: it correlates weakly with different constructs. Known-groups validity: groups that we know differ (for example, dietitians versus the general population, for a nutrition knowledge questionnaire) obtain different scores. Criterion validity compares the score with a reference standard, as in Chapter 12.

The internal structure of a questionnaire is studied by factor analysis. Exploratory factor analysis (EFA) looks for groups of items that correlate with one another and suggests how many factors (dimensions) lie behind the responses. Before applying it, we check whether the data are suitable: the Kaiser-Meyer-Olkin (KMO) index should preferably be above 0.80 (below 0.50 the analysis makes no sense), and Bartlett's test of sphericity should be significant (we reject the

hypothesis that the items are uncorrelated). The number of factors is chosen with parallel analysis, which compares the observed eigenvalues with those obtained from random data, or with other criteria. Confirmatory factor analysis (CFA) tests a structure established in advance and reports fit indices: CFI and TLI (preferably above 0.90, ideally above 0.95), RMSEA (preferably below 0.08, ideally below 0.06) and SRMR (below 0.08). Structural equation modeling (SEM) goes further and estimates the relationships between factors. These methods are beyond what a first-year student can calculate by hand, but you must be able to recognize them and read them in an article.

When a questionnaire validated in another language is used in Romania, it must be translated and culturally adapted, not just translated. The standard procedure includes two independent translations into Romanian, their synthesis, back-translation into the original language by translators who do not know the original, comparison with the original, testing on a small group of patients (cognitive interviews: "what did you understand by this question?") and then psychometric validation in the Romanian population. The ALSATION protocol describes such a process for three health literacy questionnaires [26].

Case study: validation of questionnaires in the research of colleagues in our discipline

Marin and colleagues developed a questionnaire based on the technology acceptance model (TAM) for home telemonitoring of patients with liver cirrhosis and validated it in 130 patients [10]. Internal consistency was very good (Cronbach's $\alpha = 0.934$; McDonald's $\omega = 0.950$). EFA identified a two-factor structure explaining 77.9% of the variance, and CFA confirmed it with acceptable indices (CFI = 0.927; TLI = 0.908, both above 0.90 but below 0.95). The structural model showed that perceived usefulness strongly predicts intention to use ($\beta = 0.823$, $p < 0.001$), and perceived ease of use influences usefulness ($\beta = 0.343$, $p < 0.001$); the model explained 69.5% of the variance in intention. The critical question for the reader is whether 130 patients are enough for EFA and CFA on the same sample; ideally, a structure found in an exploratory analysis is confirmed in an independent sample.

Dragomir and colleagues constructed the MMBQ-44 questionnaire on the burden of menopause, using a sample of 308 women aged 45-65 years [60]. The steps follow Section 13.5.5 exactly: the corrected item-total correlations all exceeded 0.30; the data were suitable for factor analysis (KMO = 0.931; Bartlett $\chi^2 = 9006.2$, $p < 0.001$); parallel analysis identified four factors explaining 48.6% of the total variance; internal consistency was α 0.813-0.941 and ω 0.864-0.953. The authors also report floor effects (many responses at the minimum value) for two subscales, 37.0% and 23.4%. This information matters: a subscale with a floor effect cannot detect improvement in patients who already start at the minimum.

Dicu and colleagues validated a multidimensional instrument for evaluating bakery products with alternative ingredients in 504 consumers, with a subsample of 30 people reassessed after 14 days [73]. In addition to EFA (KMO = 0.933) and CFA (CFI = 0.965; TLI = 0.959; RMSEA = 0.061; SRMR = 0.033, all in the good range), the study reports test-retest reliability with an ICC of 0.82-0.91, that is, good to excellent reliability, and a Cronbach's α of 0.898-0.938. The retest, however, was carried out on only 30 people, and the confidence intervals of the ICC in such a small subsample are wide; the point estimates should be read with caution.

Four other studies complete the picture. Frandeș and colleagues validated an instrument for the acceptability of mobile technology in diabetes self-management in 103 patients, with a Cronbach's α of 0.9 for attitude toward the use of mobile technology and 0.97 for another dimension, and used rank correlations (Spearman, Kendall), which are appropriate for ordinal data [54]. Putnoky and colleagues adapted a general nutrition knowledge questionnaire in four

stages: internal reliability in 412 people (0.878), test-retest reliability in 46 people (above 0.880), known-groups validity in 96 and convergent validity in 508 [145]. Şerban and colleagues validated the mindful eating questionnaire in Romanian adults, with an internal reliability of 0.72 and an external reliability of 0.83 [146]; the value of 0.72 is at the limit of acceptability and reminds us that alpha also depends on the number of items, a short subscale usually having a lower alpha. Finally, Cîrstea and colleagues translated and validated the Romanian version of the PSPSQ 2.0 questionnaire in 220 patients with chronic dermatological conditions [11]: α = 0.978 (subscales 0.961-0.969), a structure explaining 84.0% of the variance, CFI = TLI = 0.999, but RMSEA = 0.109, a value that indicates some model misfit. When fit indices contradict each other, the authors must discuss why; an alpha of 0.978 suggests, in addition, that some items may be redundant.

13.5.6 Electronic Data Capture

Entering data from paper into the computer is an important source of errors. Electronic data capture (EDC) eliminates this step and allows data to be checked at the very moment of entry. In the practical sessions of our discipline we use Google Forms, linked to a Google Sheets spreadsheet in which the responses are collected [46,147]. For real clinical trials, specialized platforms are used, such as REDCap (widespread in academic settings, with an audit trail and management of access rights) or Epi Info, the free CDC program that has both a form-building module and analysis modules [51].

A well-built Google Forms form uses four features. The first is the correct question type: "Short answer" for numerical values, "Multiple choice" for single choice, "Checkboxes" for multiple choice, "Linear scale" for Likert items, "Dropdown" for long lists (for example, the county), "Date" for calendar dates. The second is response validation rules [45]: for age, "Number", "Between 18 and 99"; for height, "Between 120 and 220" (which would have stopped the value 1.68 in mg2026_brut); for blood glucose, "Number", with no text (which would have stopped "112 mg/dl"); each with an explicit error message ("enter height in centimeters, for example 168"). The third is making questions required for indispensable variables (patient code, arm, primary endpoint), but not for all of them: a form in which everything is required forces the respondent to invent answers. The fourth is sections and branching [148]. The first section is the information and consent page, with the question "I agree to take part", and the option "I do not agree" sends the respondent straight to the end of the form; the question "Do you smoke?" with the answer "Yes" opens a section on the number of cigarettes per day, and the answer "No" skips it.

After the responses have been collected, the spreadsheet is downloaded and checked: one row per patient, one column per variable, short column names without spaces or diacritics (glicemie0, not "Blood glucose at enrollment (mg/dL)"), consistent codes for categories. The pivot table in Google Sheets [44] is useful for quick frequency checks.

13.5.7 The Data Dictionary and Pilot Testing

The data dictionary (or codebook) is the table that describes each variable: the short name, the full label, the type, the unit, the permitted values, the codes for categories and how missing values are coded. Table 13.2 reproduces an excerpt from the dictionary of the mg2026 dataset.

Table 13.2. Excerpt from the data dictionary of the mg2026 dataset.

Variable	Description	Type	Permitted values / unit
id	patient code	identifier	P001-P160, unique
brat	randomization arm	nominal	A, B, C
sex	sex	nominal dichotomous	F, M
varsta	age at enrollment	numerical	years, 18-99

Variable	Description	Type	Permitted values / unit
educatie	education level	ordinal	1 lower secondary, 2 upper secondary/vocational, 3 university, 4 postgraduate
inaltime	height	numerical	cm, 120-220
glicemie0	fasting plasma glucose at enrollment	numerical	mg/dL, 40-400
hba1c0	glycated hemoglobin at enrollment	numerical	%, 3.5-15, with one decimal place
diabet_ogtt	OGTT-confirmed diabetes	nominal dichotomous	0 no, 1 yes
satisfactie	satisfaction at 6 months	ordinal	1-5 (Likert)

Before the study is launched, the form is tested on a few colleagues and a few patients (pilot testing). The time needed to complete it is measured, and we check whether everyone understands the questions in the same way, whether the branching works, whether the data arrive correctly in the spreadsheet, and whether the downloaded file can be imported directly into JASP or MedCalc. Five test forms filled in with extreme values (age 18 and 99, fields left blank, the answer "I do not agree") uncover more problems than an hour of rereading.

13.6 Data Management and Data Cleaning

However good the form, raw data contain errors. The file mg2026_brut.csv, posted on Moodle alongside the clean version, was built precisely for this exercise: it has 161 rows instead of 160 and a series of typical errors. Table 13.3 presents all of them, together with the method by which each is detected.

Table 13.3. The errors in the file mg2026_brut.csv and the detection method.

Patient	Variable	Value in the raw file	Type of error	How it is detected
P005	sex	"f"	inconsistent coding	frequency table: more than 2 categories appear
P021	sex	"Feminin"	inconsistent coding	same as above
P034	tas0	1540	impossible value (an extra zero)	minimum/maximum, box plot
P042	varsta	missing	missing value	number of valid values below 160
P057	inaltime	1,72	wrong unit (m instead of cm)	minimum, histogram, aberrant recalculated BMI
P064	glicemie0	"87 mg/dl"	text in a numerical column	the column is not recognized as numerical
P084	fumat	"Fumator "	capital letter and extra space	frequency table: an extra category
P089	hba1c0	"5,2" stored as text	different decimal separator	column type, misaligned value
P102	greutate0	7,26	shifted decimal point	minimum, aberrant recalculated BMI
P121	mediu	"Urban"	capital letter	frequency table
P134	glicemie6	missing	missing value at follow-up	number of valid values
P091	entire row	duplicate (rows 91 and 161)	duplicate record	searching for duplicate codes

The cleaning procedure follows a logical order. First, an untouched copy of the raw file is kept; one never works on the original. Then duplicates are searched for by patient code. The type

of each column (numerical or text) is checked. Frequency tables are produced for all categorical variables and, for numerical ones, the minimum, the maximum, the number of valid values and a box plot. Logical consistency is checked: BMI recalculated from weight and height, weight loss consistent with the variable scadere5, blood glucose at 6 months plausible relative to the baseline value.

Correction is the delicate part. An error of form ("Urban", "Fumator " with a space, "5,2" stored as text) can be corrected without doubt, if the meaning is obvious. An impossible value is not "guessed". The temptation is to believe that 1540 mmHg was 154, that 7.26 kg was 72.6 and that 1.72 m means 172 cm. In the mg2026 file these assumptions are confirmed, but only because we checked each value against the source document (the patient's record) and, for height and weight, recalculated the BMI stored in the table: $\sqrt{\frac{77.7}{26.3}} = 1.72m$ and 27.0 x 1.64² = 72.6 kg. In real data, the same distorted figure can arise from several different mistakes; 1540 could just as well have been 150 or 154, and a blood pressure of 154 mmHg changes whether the patient is classified as hypertensive. Corrections are made only from the source; if the source is no longer available, the value is declared missing. Genuinely missing values (the age of P042, the 6-month blood glucose of P134) are retrieved from the source documents if possible; otherwise, they remain missing and are reported.

Every change is recorded in a change log: the date, who made it, the row, the variable, the old value, the new value, the reason and the source document. In professional platforms this record is kept automatically and is called an audit trail; the GCP guideline requires every change to the data to be traceable and not to erase the original value [144]. For a student project, a separate sheet in the Google Sheets file, named "modificari" (changes), is enough. Statistical analysis begins only after the database has been declared final ("locked"), and any subsequent change is justified in writing.

13.7 Reporting: CONSORT 2025

The results of an RCT are reported according to the CONSORT guideline, updated in 2025 [56]. CONSORT is a checklist of the elements that must appear in the article (a title stating that the study is randomized, registration, access to the protocol, design, participants, interventions, endpoints, sample size, randomization and allocation concealment, blinding, statistical methods, participant flow, baseline characteristics, results with the effect size and its precision, adverse effects, limitations, generalizability), accompanied by a flow diagram. For observational studies the corresponding guideline is STROBE [25], and for diagnostic accuracy studies it is STARD [140].

The CONSORT flow diagram shows what happened to each patient, from the assessment of eligibility to the analysis. Figure 13.3 presents the diagram for mg2026: 212 people assessed, 52 excluded (31 did not meet the criteria, 21 declined), 160 randomized (53 in arm A, 53 in B, 54 in C), all followed up at 3 and 6 months and analyzed by intention to treat.

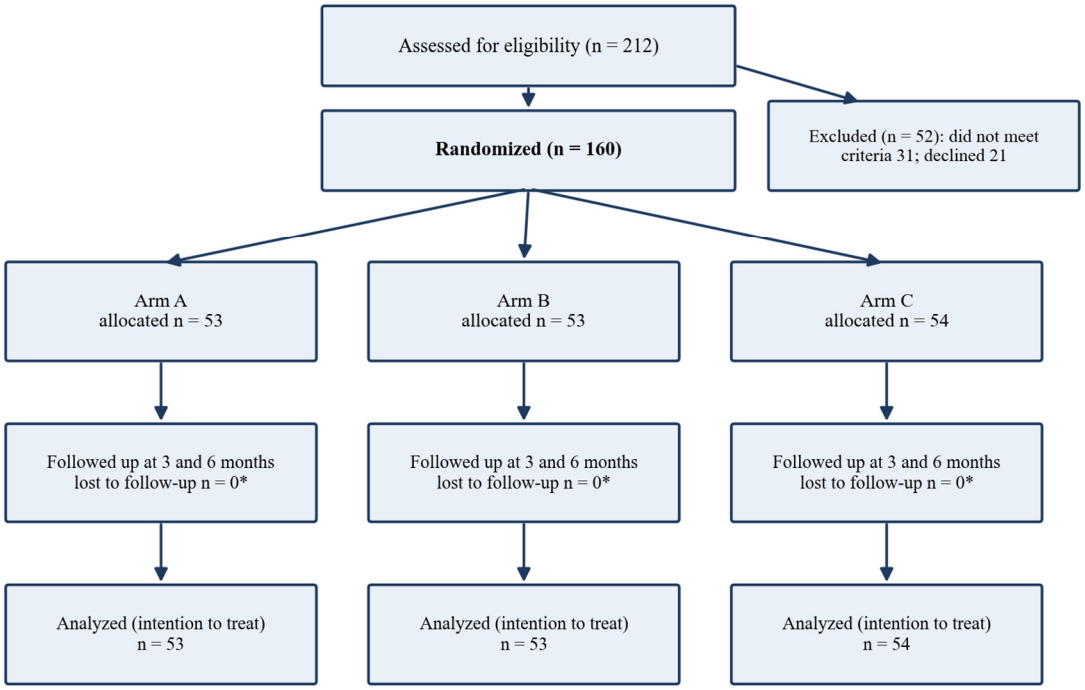


Figure 13.3. CONSORT flow diagram of the simulated mg2026 study: 212 people assessed for eligibility, 52 excluded (31 do not meet the criteria, 21 decline), 160 randomized into three arms (A 53, B 53, C 54), no losses to follow-up in the teaching dataset, all analyzed by intention to treat.

The diagram answers questions the reader should be asking. How many eligible patients declined? If refusals were massive, the sample might not be representative: those who accept a study with a mobile app may be younger and more familiar with technology. How many were lost in each arm, and for what reasons? Unequal losses between arms are a warning sign. How many were analyzed, compared with how many were randomized? If the numbers differ, the analysis is no longer ITT. In a real study, the "lost to follow-up" box almost never contains zero; in the teaching dataset we chose this simplification so as not to complicate the exercises, and the note in the figure says so explicitly.

13.8 Integrated Analysis of the mg2026 Study: A Complete Example

We have reached the part where all the previous chapters work together. We will go through the analysis of the mg2026 study in the order in which it would appear in an article: the plan, the population, the baseline characteristics, the primary endpoint, the secondary endpoints, the effect sizes, the conclusion and limitations, and then the observational analyses that can be performed on the same data. All the figures come from the mg2026_curat.csv dataset and can be reproduced in JASP 0.98.1 [47] or MedCalc 24.0 [50], as you will do in Practical Session 13, following the steps also described in the collections of practical applications of the biostatistics disciplines [14,149].

13.8.1 The Statistical Analysis Plan, Written before the Data

The study's analysis plan, drafted before the arm codes were unblinded, specifies the following. The primary analysis population is the ITT population (all 160 randomized patients, in the allocated arm). The primary endpoint is the reduction in fasting plasma glucose at 6 months (glicemie0 - glicemie6, in mg/dL), compared across the three arms by one-way ANOVA, after checking normality in each arm (Shapiro-Wilk) and homogeneity of variances (Levene); pairwise comparisons are made with the Tukey test; if the assumptions are not met, Kruskal-Wallis with

corrected Mann-Whitney comparisons is used. The significance level is $\alpha = 0.05$, two-sided. The secondary endpoints are: the proportion of patients with weight loss of at least 5% (chi-square, relative risk, risk difference, NNT, with 95% CI); weight loss in kg (ANOVA); the reduction in SBP at 6 months (ANOVA); satisfaction at 6 months (Kruskal-Wallis); the course of blood glucose at 0, 3 and 6 months (Friedman and repeated-measures ANOVA). The secondary results are interpreted as supportive, without formal correction, but with caution regarding multiple comparisons. Additional observational analyses, at baseline: the obesity-diabetes association and the performance of the FINDRISC score against the OGTT.

Such a paragraph contains nothing spectacular, but it is what gives the results credibility. Note that the alternative test (Kruskal-Wallis) is also announced in advance, together with the rule by which it is chosen. Choosing the test after seeing which one gives the smaller p is one of the most common forms of "fishing" for significance.

13.8.2 Baseline Characteristics: Table 1

The first table of any article reporting an RCT describes the patients at enrollment, by arm. Numerical variables with an approximately normal distribution are described by the mean and standard deviation, skewed ones by the median and interquartile range, and categorical ones by absolute frequency and percentage (Chapter 2). Table 13.4 is the "Table 1" of the mg2026 study.

Table 13.4. Patient characteristics at enrollment, by arm (mg2026 study, n = 160). Mean (SD), median [IQR] or n (%).

Characteristic	Arm A (n = 53)	Arm B (n = 53)	Arm C (n = 54)	Total (n = 160)
Age, years	57,5 (9,2)	56,9 (9,4)	54,5 (7,5)	56,3 (8,8)
Female sex	32 (60,4)	26 (49,1)	28 (51,9)	86 (53,8)
Urban residence	25 (47,2)	36 (67,9)	34 (63,0)	95 (59,4)
University or postgraduate education	26 (49,1)	29 (54,7)	28 (51,9)	83 (51,9)
Current smoker	18 (34,0)	18 (34,0)	17 (31,5)	53 (33,1)
Former smoker	5 (9,4)	11 (20,8)	17 (31,5)	33 (20,6)
Never smoker	30 (56,6)	24 (45,3)	20 (37,0)	74 (46,3)
Physical activity (1-5)	3 [2-4]	3 [2-3]	3 [2-3]	3 [2-3]
BMI, 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]
Obesity (BMI ≥ 30)	18 (34,0)	20 (37,7)	16 (29,6)	54 (33,8)
Waist circumference, cm	89,4 (12,3)	87,8 (12,1)	88,4 (12,1)	88,5 (12,1)
SBP, mmHg	131,5 (13,8)	132,8 (13,8)	129,1 (14,0)	131,1 (13,9)
DBP, mmHg	80,5 (8,7)	80,1 (8,3)	77,1 (8,6)	79,2 (8,6)
Total cholesterol, mg/dL	209,3 (39,0)	202,8 (37,5)	215,9 (33,6)	209,4 (36,9)
HDL cholesterol, mg/dL	50,7 (11,5)	51,1 (12,4)	46,9 (11,2)	49,5 (11,8)
Triglycerides, mg/dL	143 [95-204]	144 [112-185]	166,5 [107-213]	151,5 [107-203,25]
Fasting plasma glucose, 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)
FINDRISC score	10,7 (4,4)	10,1 (4,2)	10,3 (4,2)	10,4 (4,3)
OGTT-confirmed diabetes	10 (18,9)	9 (17,0)	10 (18,5)	29 (18,1)

The table has no column of p values, and this is intentional. CONSORT recommends against statistically testing baseline differences in an RCT [56]. The reasoning is simple: if randomization

was carried out correctly, any difference is, by definition, due to chance, and a statistical test would be testing a hypothesis we already know to be true. With 20 rows tested at α = 0.05, we expect, on average, one "significant" result purely by chance. What matters is the size of the difference and the prognostic importance of the variable, not the p value.

What, then, do we read in the table? The arms are well balanced for age, BMI, blood pressure, HbA1c, FINDRISC and the prevalence of diabetes (18.9%, 17.0% and 18.5%). A few differences stand out: arm C has more former smokers (31.5% versus 9.4% in A), arm A has more patients from rural areas (52.8%), and baseline blood glucose is about 5 mg/dL lower in C than in A. The last difference matters for interpretation: a patient who starts from a higher blood glucose has, in principle, more room to decrease. The fact that arm C, which starts lower, nevertheless shows the greatest reduction makes the result more convincing, not less. If we believed that a baseline difference could influence the result, a sensitivity analysis with adjustment (linear regression with baseline blood glucose and arm as predictors) should be specified in the plan, not decided after seeing Table 1.

13.8.3 The Primary Endpoint: Reduction in Blood Glucose at 6 Months

Description by arm always comes before the test. The mean reduction in blood glucose was 3.0 mg/dL (SD 8.8) in arm A, 11.1 mg/dL (SD 11.2) in B and 16.3 mg/dL (SD 10.6) in C. The 95% confidence intervals for the mean of each arm, calculated with the t distribution (Chapter 4), are 0.6-5.4 mg/dL for A, 8.0-14.2 for B and 13.4-19.2 for C. Already here we can see that the interval for arm A does not overlap with either of the others.

We check the ANOVA assumptions. The Shapiro-Wilk test applied to the reduction in blood glucose in each arm gives p = 0.575 (A), p = 0.052 (B) and p = 0.903 (C); none of them rejects normality, and with more than 50 patients per arm ANOVA is robust to moderate departures anyway. Levene's test gives p = 0.163, so we do not reject homogeneity of variances. ANOVA is therefore the appropriate test, as the plan specified.

The result: F(2, 157) = 22.93, p < 0.001. We reject H0 that the mean reduction in blood glucose is the same in the three arms. The effect size is $\eta^2 = 0.226$: about 23% of the variability in the reduction in blood glucose is due to the treatment arm. Omega², a less optimistic variant calculated from the same sums of squares, is 0.215. Tukey comparisons show that all three pairs differ significantly: B versus A, difference 8.1 mg/dL(CI 95% 3.4 − 12.8; $p = 0.0002$); C versus A, 13.3 mg/dL (95% CI 8.6-18.0; p < 0.001); C versus B, 5.2 mg/dL(CI 95% 0.6 − 9.9; $p = 0.024$). The Kruskal-Wallis test, specified as the alternative, confirms the result (H = 36.63, p < 0.001). The left panel of Figure 13.4 shows the Tukey differences graphically.

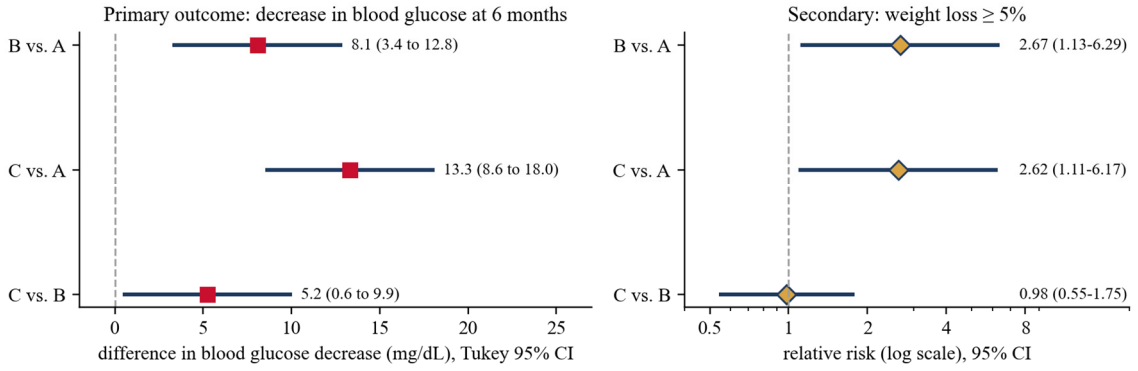


Figure 13.4. Effects of the interventions in the mg2026 study. Left: differences between arms in the reduction in blood glucose at 6 months (mg/dL), with Tukey 95% confidence intervals; no interval includes 0. Right: relative risk of achieving a weight loss of at least 5%, with 95% CI, on a logarithmic scale; the intervals for B and C versus A do not include 1, while the one for C versus B includes 1.

How do we read these figures clinically? An additional reduction of 13.3 mg/dL in fasting plasma glucose for arm C compared with standard counseling is considerable in a population with mean baseline values of 105 mg/dL, that is, right in the prediabetes range. The standardized effect size (Cohen's d, Chapter 6) is d = 1.37 for C versus A, d = 0.80 for B versus A and d = 0.48 for C versus B; by the usual benchmarks, these are large effects for the first two and a moderate effect for the third [76]. The C versus B difference, however, has a 95% CI that goes down to 0.6 mg/dL, that is, very close to zero: the data are compatible both with a substantial additional benefit of teleconsultation and with an almost negligible one. It is the interval, not the p value of 0.024, that tells us how precise the estimate is.

13.8.4 Secondary Endpoints

Weight loss of at least 5%. This threshold was reached by 6 of 53 patients in arm A (11.3%), 16 of 53 in B (30.2%) and 16 of 54 in C (29.6%). The chi-square test for the 3 x 2 table gives χ^2 = 6.77, df = 2, p = 0.034; the smallest expected frequency is 12.6, so the condition for applying the test (all expected frequencies at least 5) is met. For the C versus A comparison, the relative risk is 2.62 (95% CI 1.11-6.17): the probability of losing at least 5% of body weight was 2.6 times higher in arm C than in arm A. The absolute risk difference is 18.3 percentage points (95% CI 3.4-33.2), and the number needed to treat is $\text{NNT} = \frac{1}{0.183} = 5.5$, rounded up to 6 (95% CI 3.0-29.1): we need to enroll 6 patients in program C instead of standard counseling for one more patient to reach a 5% weight loss. For B versus A, RR = 2.67 (95% CI 1.13-6.29), with a risk difference of 18.9 percentage points. For C versus B, RR = 0.98 (95% CI 0.55-1.75): teleconsultation brought no additional gain on this endpoint. The right panel of Figure 13.4 shows these three ratios.

Note the width of the intervals. The NNT could be anywhere between 3 and 29, that is, between "highly effective" and "modest". With only 22 events in the two arms being compared, precision is limited, and an honest report says so.

Weight loss in kilograms. A: 0.71 ± 2.48 kg; B: 2.84 ± 2.36 kg; C: 3.37 ± 2.39 kg; ANOVA F = 18.25, p < 0.001. The continuous outcome and the dichotomized one (≥ 5%) tell the same story, and the continuous version retains more information. Dichotomizing a numerical variable loses statistical power, which is why the primary endpoint of a study is preferably the continuous variable.

Reduction in systolic blood pressure. Here we have an instructive example. The mean reduction in SBP at 6 months was 2.8 mmHg (SD 9.0) in A, 4.5 mmHg (SD 9.6) in B and 7.0 mmHg (SD 8.2) in C. The test specified in the plan, ANOVA, gives F = 2.94, p = 0.056. Kruskal-Wallis, applied to the same data, would give p = 0.046. What do we report? The test in the plan: the difference does not reach statistical significance (p = 0.056). Changing the test after seeing the result, in order to "get" below 0.05, would be exactly the practice described above as fishing for significance. The correct statement is that there is a trend toward a greater reduction in SBP in the arms with the app, but the study was not sufficiently powered for this secondary endpoint and the result needs to be confirmed. Across all 160 patients, SBP fell on average by 4.8 mmHg (SD 9.0; paired t = 6.68, p < 0.001), but this overall reduction cannot be attributed to the intervention, because it has no comparator: it includes regression to the mean and the effect of taking part in the study.

Satisfaction. The medians were 3 (IQR 2-3) in A, 4 (IQR 3-4) in B and 4 (IQR 3-5) in C; Kruskal-Wallis H = 32.16, p < 0.001; the C versus A comparison with the Mann-Whitney test: U = 2265.5, p < 0.001, with a rank-biserial correlation of 0.58 (a large effect). This is a subjective endpoint, reported by patients who know which intervention they received. The difference may reflect both the real value of the program and the simple fact that they received more attention.

Change in blood glucose over time. In the whole sample, mean blood glucose was 105.4, 99.4 and 95.2 mg/dL at 0, 3 and 6 months (medians 104.5, 99.0 and 92.5); the Friedman test gives $\chi^2 = 74.67$, $p < 0.001$, and repeated-measures ANOVA gives $F = 64.97$, $p < 0.001$. By arm, the Friedman test is borderline in arm A ($\chi^2 = 6.04$, $p = 0.049$; means 107.8, 105.7, 104.8 mg/dL) and highly significant in B ($\chi^2 = 21.07$, $p < 0.001$; means 106.3, 101.3, 95.2) and in C ($\chi^2 = 64.63$, $p < 0.001$; means 102.3, 91.2, 86.0). Beware of a trap: the fact that the test is significant in one arm and "more significant" in another does not prove that the arms differ from each other. The question "does blood glucose change differently across the arms?" is answered directly by the primary endpoint (the comparison of reductions between arms) or, more elegantly, by the time x arm interaction term in a mixed ANOVA, a method beyond the level of this course.

13.8.5 Multiple Comparisons and the Hierarchy of Results

We performed a single primary test and at least five secondary tests, plus the post hoc comparisons. If all the H0s were true, the probability that at least one test out of ten would come out significant purely by chance would be about 40%. The mg2026 study protects itself through a hierarchy: the conclusion rests on the primary endpoint, chosen in advance and tested only once, with the post hoc comparisons controlled by the Tukey test; the secondary endpoints are consistent with the primary one and are presented as supportive. Had we wanted a secondary endpoint to be conclusive, we could have used the Bonferroni correction (the 0.05 threshold divided by the number of tests) or a prespecified hierarchical testing procedure, in which the second endpoint is tested only if the first was significant.

13.8.6 Conclusion and Limitations

A concluding paragraph, written in the style of an article, would read as follows. "In a pragmatic randomized trial in 12 family practices, adding a self-monitoring app to standard counseling increased the reduction in fasting plasma glucose at 6 months by 8.1 mg/dL (95% CI 3.4-12.8), and additionally adding monthly teleconsultation increased it by 13.3 mg/dL (95% CI 8.6-18.0) compared with standard counseling. The proportion of patients who lost at least 5% of their body weight doubled in the arms with the app (RR 2.62 for arm C versus A, 95% CI 1.11-6.17; NNT 6). The additional benefit of teleconsultation over the app alone was smaller and imprecisely estimated."

The limitations must be stated just as clearly. Patients and physicians were not blinded, which mainly affects the subjective endpoints (satisfaction). The primary endpoint is a surrogate (blood glucose), not a hard clinical outcome (incidence of diabetes, cardiovascular events). Follow-up of 6 months tells us nothing about whether the effect is maintained. Randomization was individual, within practices in which the physician applied all three strategies, so contamination between arms is possible and would have tended to reduce the differences. The data are complete, an unrealistic situation for a real study. And the patients who agreed to take part (160 of 181 eligible) may be more motivated than the general population of patients at cardiometabolic risk.

13.8.7 Observational Analyses on the Same Data

A dataset from an RCT can also be used for observational questions, with one condition that many people forget: randomization protects only the comparison between arms. Any other association in the same dataset (obesity and diabetes, age and blood pressure) is an observational, cross-sectional one, exposed to confounding as in any study of this type (Chapter 10).

At baseline, the prevalence of OGTT-confirmed diabetes was 33.3% in obese patients (18 of 54) and 10.4% in non-obese patients (11 of 106). The prevalence ratio is 3.21 (95% CI 1.64-6.31), and the odds ratio is OR = 4.32 (95% CI 1.86-10.03); $\chi^2 = 12.70$, $p < 0.001$. Since this is a cross-sectional design, the natural measure is the prevalence ratio; the OR overestimates it, as expected

when the disease is not rare. In the logistic regression adjusted for age and sex, the OR for obesity becomes 6.01 (95% CI 2.35-15.39; p = 0.0002), and each year of age increases the odds of diabetes by 11% (OR 1.11; 95% CI 1.05-1.18). We cannot say that obesity "causes" diabetes on the basis of these data: we do not know which came first, and other uncontrolled factors may explain the association.

The second observational question: how well does the FINDRISC score identify patients with OGTT-confirmed diabetes [125]? The area under the ROC curve (Chapter 12) is 0.852 (95% CI 0.762-0.943), close to that of fasting plasma glucose (0.844). The cutoff that maximizes the Youden index is FINDRISC ≥ 11 (Se 93.1%, Sp 66.4%, J = 0.595). At the cutoff ≥ 12, the 2 x 2 table has 22 true positives, 33 false positives, 7 false negatives and 98 true negatives: Se 75.9%, Sp 74.8%, PPV 40.0%, NPV 93.3%. With a prevalence of 18.1% in the sample, a positive score raises the probability of diabetes only to 40%, but a negative score lowers it below 7%. FINDRISC therefore works better for exclusion than for confirmation, and the OGTT remains necessary in those who test positive. Figure 13.5 reproduces the ROC curve from Chapter 12.

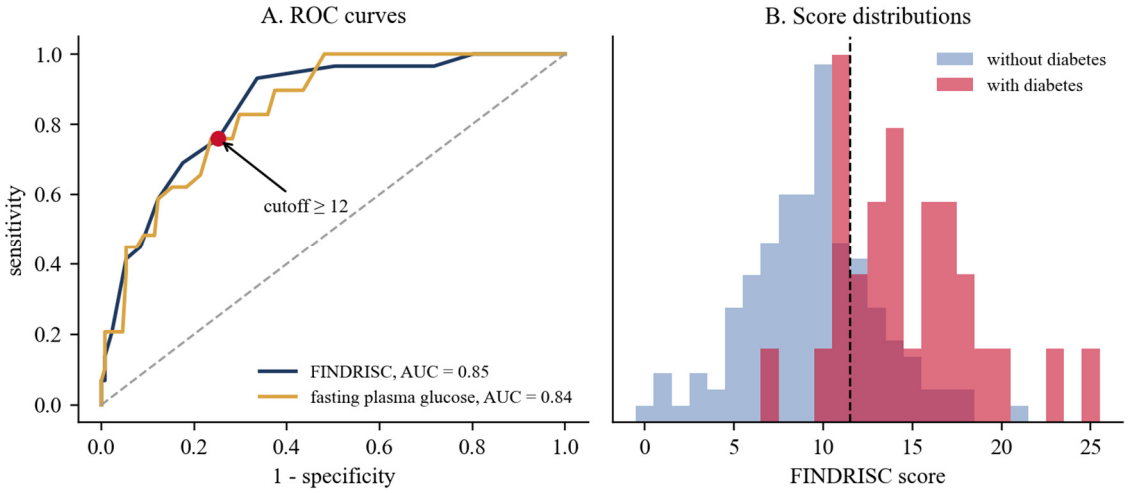


Figure 13.5. ROC curve of the FINDRISC score for OGTT-confirmed diabetes at enrollment in the mg2026 study (area under the curve 0.852, 95% CI 0.762-0.943), a cross-sectional observational analysis of data from a randomized trial.

Case study: confounding, nonrandomized parallel groups and the before-after design

Christodorescu and colleagues followed 278 patients hospitalized with decompensated heart failure for 12 months to see whether the hemoglobin to red cell distribution width ratio (HRR) predicts events [23]. The median HRR was 0.89, and the primary endpoint occurred in 167 patients (60.1%). In the Kaplan-Meier analysis, patients with low HRR had lower event-free survival (log-rank p = 0.027), but in the Cox model adjusted for age, sex, NT-proBNP, ejection fraction and glomerular filtration rate, the association was no longer significant (p = 0.240). This is the classic example of confounding: low HRR accompanies older patients, with poorer renal function and higher NT-proBNP, and these patients have a poorer prognosis anyway. In a cohort, adjustment is the only defense; in an RCT, randomization would have balanced these factors. The authors correctly report both results.

Uivarosan and colleagues compared, in 70 patients after stroke, two groups receiving combined therapy (modern rehabilitation techniques and neurotrophic medication, n = 22 and n = 18) with a control group receiving standard treatment (n = 30), followed up for 12 months [89]. The abstract reports "significantly better" results (p < 0.05) in the combined therapy groups and improvements at 12 months compared with baseline (p < 0.05), without specifying the tests. The parallel-group design with a control group is sound in principle, but the reader

must find out from the full text how the groups were formed (randomization or allocation by some other criterion?), why they differ in size and which test was applied. For three independent groups, ANOVA or Kruskal-Wallis would be appropriate, and for change over time a test for repeated measurements.

Buda and colleagues measured serum thrombospondin-1 in hypertensive patients with endothelial dysfunction, before and after one year of treatment with perindopril [86]. Values rose from 8061.9 ± 3684.80 to 15380 ± 5887 ng/mL($p < 0.001$), compared with 9221.03 ± 6510.21 ng/mL in the non-hypertensive controls, and another marker fell from 0.91 ± 0.51 to 0.50 ± 0.24 ng/mL($p < 0.001$). The appropriate analysis is one for paired data (paired t-test or Wilcoxon, noting that the large standard deviations relative to the means suggest skewness). The control group of healthy people serves as a reference, not as a treatment control group: it received neither perindopril nor placebo. This is why the before-after change cannot be attributed with certainty to the drug, and the number of patients, not mentioned in the abstract, is the first piece of information the reader should look for.

13.9 Worked Problems

Worked Problem 13.1. How Many Patients Should Be Enrolled in the mg2026 Study?

Statement. The team designing the mg2026 study considers a difference of 6 mg/dL in the reduction in blood glucose between an intervention arm and the control arm to be clinically relevant and, from previous studies, estimates the standard deviation of the reduction at 10 mg/dL. For the secondary endpoint, 10% of the patients in arm A and 30% of those in arm C are expected to lose at least 5% of their body weight. The calculation uses α = 0.05 (two-sided) and 80% power. Losses to follow-up are estimated at 15%. How many patients should be enrolled per arm?

Solution. For the primary endpoint we use the formula for two means, with $z\left(1-\frac{\alpha}{2}\right) = 1.96$ and z(1-β) = 0.84 (for 80% power):

$$n = \frac{2 \times (1.96 + 0.84)^2 \times \sigma^2}{\Delta^2} = \frac{2 \times 7.84 \times 100}{36} = 43.6$$

so 44 patients per arm. The standardized effect size is $d = \frac{6}{10} = 0.6$. Had the team been more cautious and assumed d = 0.5, it would have obtained 64 patients per arm (exact calculation with the t distribution).

For the secondary endpoint, p1 = 0.10, p2 = 0.30, $\bar{p} = 0.20$:

$\sqrt{2 \times 0.20 \times 0.80} = \sqrt{0.32} = 0.566$; $1.96 \times 0.566 = 1.109\sqrt{0.10 \times 0.90 + 0.30 \times 0.70} = \sqrt{0.30} = 0.548$; $0.84 \times 0.548 = 0.461n = \frac{(1.109+0.461)^2}{(0.30-0.10)^2} = \frac{1.570^2}{0.04} = \frac{2.464}{0.04} = 61.6$

so 62 patients per arm. This is more than for the primary endpoint: binary endpoints generally require larger samples.

Correction for losses: if we want 62 patients to remain evaluable in each arm and we expect 15% losses, we enroll 62/0.85 = 72.9, that is, 73 patients per arm, about 219 in total.

Interpretation. With 53-54 patients per arm, the mg2026 study has sufficient power for the primary endpoint (44 needed), but falls slightly short of what is needed for the binary secondary endpoint. This is exactly what we see in the results: the primary endpoint is very clear, whereas the intervals for the RR are wide (1.11-6.17). The formula does not take into account that we have three arms and three comparisons; a correction for multiple comparisons would increase the sample slightly.

Worked Problem 13.2. Building a Block Randomization List

Statement. Build the first 12 allocations of a randomization list with blocks of 6 for three arms (A, B, C), using the following random numbers (generated, for example, with the =RAND() function in Google Sheets): block 1: 0.71; 0.15; 0.88; 0.42; 0.07; 0.56; block 2: 0.33; 0.91; 0.26; 0.64; 0.48; 0.12. How many distinct sequences are possible within a block? Why is a fixed block size dangerous?

Solution. Each block contains the letters A, A, B, B, C, C. We assign a random number to each letter, in order, and sort in ascending order by number.

Block 1: A 0.71; A 0.15; B 0.88; B 0.42; C 0.07; C 0.56. Sorted in ascending order: C (0.07), A (0.15), B (0.42), C (0.56), A (0.71), B (0.88). The sequence: C A B C A B.

Block 2: A 0.33; A 0.91; B 0.26; B 0.64; C 0.48; C 0.12. Sorted: C (0.12), B (0.26), A (0.33), C (0.48), B (0.64), A (0.91). The sequence: C B A C B A.

The first 12 patients: C, A, B, C, A, B, C, B, A, C, B, A. After every 6 patients, each arm has exactly 2.

The number of distinct sequences in a block is the number of permutations of the multiset AABBCC: $\frac{6!}{2!\times 2!\times 2!} = \frac{720}{8} = 90$.

The danger of a fixed size: if the physician enrolling patients knows that the blocks have 6 places and has already seen 5 allocations in the block, the sixth is completely predictable; after 4 allocations, he or she already knows in many cases which two letters remain. The physician can then influence the enrollment of the next patient. The solutions are allocation concealment (the physician does not see the list) and blocks of variable size (3, 6 or 9, in random order). For stratification by practice, a separate list of this kind is generated for each of the 12 practices.

Worked Problem 13.3. Why Do We Not Test the Differences in Table 1?

Statement. A student proposes adding to Table 13.4 a column of p values for each of the 20 rows. He notes that for smoking the chi-square test gives p = 0.072, and for area of residence p = 0.075, and concludes that "randomization narrowly failed". What do you tell him?

Solution. If randomization was carried out correctly, H0 "the arms come from the same population" is true by construction for all baseline characteristics. Any observed difference is a sampling error. A statistical test in this situation tells us nothing new: it will come out "significant" with probability α, that is, in about 5% of cases. With 15 independent tests at α = 0.05, the probability that at least one comes out significant purely by chance is

$$1 - 0.95^{15} = 1 - 0.463 = 0.537$$

that is, more than half. A p of 0.072 does not indicate a "narrow failure" of randomization; it indicates an ordinary fluctuation. What must be judged is whether the difference (for example, 31.5% former smokers in C versus 9.4% in A) is large enough, and whether the variable is strongly enough related to the outcome, to be able to influence the conclusion. If so, a sensitivity analysis with adjustment is specified in the plan. The correct conclusion: Table 1 is read clinically, not through tests.

Worked Problem 13.4. ANOVA Calculated by Hand from Group Statistics

Statement. Starting only from the descriptive statistics of the reduction in blood glucose (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), calculate F, eta² and Cohen's d for C versus A.

Solution. The overall mean, weighted by group size:

$$m = \frac{53 \times 3.02 + 53 \times 11.09 + 54 \times 16.33}{160} = \frac{160.06 + 587.77 + 881.82}{160} = \frac{1629.65}{160} = 10.19$$

The between-group sum of squares:

$$\text{SSB} = 53 \times (3.02 - 10.19)^2 + 53 \times (11.09 - 10.19)^2 + 54 \times (16.33 - 10.19)^2$$
$$= 53 \times 51.34 + 53 \times 0.82 + 54 \times 37.76 = 2721.1 + 43.4 + 2038.9 = 4803.4$$

The within-group sum of squares, from the standard deviations:

$$\text{SSW} = 52 \times 8.76^2 + 52 \times 11.21^2 + 53 \times 10.59^2 = 3990.4 + 6534.5 + 5943.8 = 16468.7$$

Degrees of freedom: between groups k - 1 = 2; within groups N - k = 157. Mean squares: $\text{MSB} = \frac{4803.4}{2} = 2401.7$; $\text{MSW} = \frac{16468.7}{157} = 104.9$. The ratio $F = \frac{2401.7}{104.9} = 22.90$. The software, which works with unrounded values, gives 22.93; the difference comes only from rounding.

$\eta^2 = \frac{\text{SSB}}{\text{SSB+SSW}} = \frac{4803.4}{21272.1} = 0.226.$

For Cohen's d for C versus A, the pooled standard deviation is

$$\text{sp} = \sqrt{\left[\frac{53 \times 10.59^2 + 52 \times 8.76^2}{54 + 53 - 2}\right]} = \sqrt{\frac{9934.2}{105}} = \sqrt{94.6} = 9.73$$

$$d = \frac{16.33 - 3.02}{9.73} = 1.37$$

Interpretation. The critical value of F for 2 and 157 degrees of freedom at α = 0.05 is approximately 3.05; F = 22.9 is far above it, so p < 0.001. The arm explains 22.6% of the variability in the reduction in blood glucose, and the C-A difference is almost 1.4 standard deviations, a large effect.

Worked Problem 13.5. Relative Risk, Risk Difference and NNT with Confidence Intervals

Statement. Calculate the RR, the risk difference and the NNT, with 95% CI, for weight loss of at least 5% in arm C (16 of 54) versus arm A (6 of 53).

Solution. The risks: $p_C = \frac{16}{54} = 0.296$; $p_A = \frac{6}{53} = 0.113$. $\text{RR} = \frac{0.296}{0.113} = 2.62$.

The interval for the RR is calculated on the logarithmic scale (Chapter 10):

$\text{SE}(\ln \text{RR}) = \sqrt{\frac{1}{16} - \frac{1}{54} + \frac{1}{6} - \frac{1}{53}} = \sqrt{0.0625 - 0.0185 + 0.1667 - 0.0189} = \sqrt{0.1918} =$ $0.438 \ln 2.62 = 0.963$; $0.963 \pm 1.96 \times 0.438 = 0.963 \pm 0.858$, that is, 0.105 to 1.821; $\text{CI}\,95\%\,\text{RR} = e^{0.105}$ to $e^{1.821} = 1.11$ to 6.18 (the software, with unrounded values, gives 1.11-6.17)

Risk difference: 0.296 - 0.113 = 0.183 (18.3 percentage points).

$\text{SE(RD)} = \sqrt{\frac{0.296 \times 0.704}{54} + \frac{0.113 \times 0.887}{53}} = \sqrt{0.00386 + 0.00189} = 0.0759;\ \text{CI}\,95\% = 0.183 \pm$ $1.96 \times 0.0759 = 0.183 \pm 0.149$, that is, 0.034 to 0.332

$\text{NNT} = \frac{1}{0.183} = 5.5$, rounded up to 6. The interval for the NNT is obtained by inverting the limits of the interval for the risk difference: 1/0.332 = 3.0 and 1/0.034 = 29 (the software, with unrounded values, gives 29.1).

Statistical interpretation. The 95% CI for the RR does not include 1, and the one for the risk difference does not include 0, so the difference is significant at α = 0.05 (consistent with Fisher's exact test, p = 0.030). Clinical interpretation. For every 6 patients treated in program C instead of

standard counseling, one more reaches a 5% weight loss; the estimate, however, is imprecise, compatible with an NNT between 3 and 29.

Worked Problem 13.6. Cronbach's Alpha Calculated by Hand

Statement. A scale of 4 Likert items (1-5) measuring confidence in using the mobile app was administered to 40 patients from arms B and C. The item variances are 1.2; 1.0; 1.4 and 1.1, and the variance of the total score (the sum of the 4 items) is 14.2. Calculate Cronbach's α and interpret it.

Solution. The sum of the item variances: 1.2 + 1.0 + 1.4 + 1.1 = 4.7.

$$\alpha = \frac{4}{3} \times \left(1 - \frac{4.7}{14.2}\right) = 1.333 \times (1 - 0.331) = 1.333 \times 0.669 = 0.892$$

Interpretation. α = 0.89 indicates good internal consistency: the items appear to measure the same construct. Think about what would have happened if the items had been uncorrelated: the variance of the total would have been approximately equal to the sum of the variances (4.7), and α would have been close to 0. The more the items covary, the more the variance of the total rises above the sum of the variances and the closer α gets to 1. A high α does not, however, demonstrate that the scale measures confidence and not something else (validity), nor that patients would answer in the same way two weeks later (test-retest reliability, measured with the ICC).

Worked Problem 13.7. Intention to Treat versus Per-Protocol Analysis

Statement. Suppose, for the sake of the exercise, that in arm C of the study 10 of the 54 patients had never used the app and had never attended the teleconsultations. Of the 44 adherent patients, 15 would have reached the 5% reduction, and of the 10 non-adherent patients, 1. In arm A, all patients would have received counseling (6 of 53 with a 5% reduction). Calculate the RR in the ITT analysis and in the per-protocol analysis, and comment.

Solution. ITT: all 54 randomized patients remain in arm C, with 15 + 1 = 16 events; $p_C = \frac{16}{54} = 29.6\%$; $\text{RR} = \frac{0.296}{0.113} = 2.62$. Per protocol: we keep only the 44 adherent patients; $p_C = \frac{15}{44} = 34.1\%$; $\text{RR} = \frac{0.341}{0.113} = 3.01$.

Comment. The per-protocol analysis gives a larger effect, but it loses the comparability of the arms. The 10 non-adherent patients are not chosen at random: they are probably less motivated, less familiar with smartphones, perhaps older. Arm A certainly contains similar patients, but we did not remove them, because there was no intervention there for them not to adhere to. We are therefore comparing "the motivated patients in C" with "all patients in A", and part of the difference is due to motivation, not to the intervention. ITT answers the family physician's question: "if I recommend this program to my patients, with their real adherence, what do I gain?". The per-protocol analysis approximates the effect in the patient who adheres, but at the risk of systematic error. When the two analyses agree in direction, the conclusion is robust.

Worked Problem 13.8. What a Single Wrong Value Does

Statement. In the raw file, patient P034 has tas0 = 1540 mmHg (the real value, from the source record, is 154). Calculate the mean, the standard deviation and the median of tas0 in the two versions, and comment.

Solution. With the correct data: mean 131.13 mmHg, SD 13.88, median 131. Replacing the value 154 with 1540 adds 1540 - 154 = 1386 to the sum of the values, so the mean increases by 1386/160 = 8.66 mmHg, to 139.79 mmHg. The standard deviation explodes to 112.24 mmHg, because the squared deviation of this single value from the mean dominates the entire sum. The median remains 131: the extreme value has moved a single element from below the median to

above it, which can shift the median by at most one position, and around the center of this dataset several values are equal to 131.

Interpretation. A single typing error shifts the mean by more than 8 mmHg and makes the confidence interval absurdly wide; any t-test applied to these data would be compromised. The median is robust, exactly as we learned in Chapter 2. The lesson for data management: the minimum, the maximum and a box plot for every numerical variable, before any analysis. And once again: the correction is made from the source record, not by guesswork, even though here the source confirms the value of 154 mmHg.

13.10 Proposed Exercises

1. A study compares two exercise programs for lowering SBP. The minimal relevant difference is 8 mmHg, and the estimated standard deviation of the reduction is 12 mmHg. Calculate the number of patients per group for $\alpha = 0.05$ and 90% power ($z(1-\beta) = 1.28$). What does it become if 10% losses are expected?
2. In an influenza vaccination study in older adults, the expected proportion of patients with severe respiratory infection is 35% without intervention and 20% with a home vaccination program. Calculate the sample size per group for $\alpha = 0.05$ and 80% power.
3. A cluster randomized trial includes practices with 20 patients each. The estimated intraclass correlation is 0.02. Calculate the design effect and the number of patients per arm, if an individually randomized trial would have needed 64 per arm.
4. A 5-item scale has a sum of item variances of 5.5 and a total score variance of 18.0. Calculate Cronbach's α and interpret it.
5. In an RCT with 150 patients per arm, the adverse event (hospitalization within 12 months) occurs in 30 patients treated with the new program and in 45 in the control group. Calculate the relative risk, the absolute risk reduction and the NNT.
6. In addition to the primary endpoint, a study has 8 secondary endpoints, each tested at $\alpha = 0.05$. What is the probability of obtaining at least one false-positive result if all the H0s are true and the tests are independent? What threshold would the Bonferroni correction impose?
7. A non-inferiority study compares teleconsultation with in-person visits, with a non-inferiority margin $\Delta = 3$ mg/dL for the reduction in blood glucose. What conclusion do you draw if the 95% CI of the difference (teleconsultation minus in-person visits) is: a) -2.4 to 1.8 mg/dL; b) -4.1 to 0.9 mg/dL; c) 0.5 to 4.2 mg/dL?
8. Classify the design of each study: a) 60 patients with COPD receive, in random order, a new bronchodilator and a standard one, for 4 weeks each, with a 2-week washout between them; b) 24 schools are randomized to a nutrition education program; c) 21 children receive a new therapy and are assessed at 0, 30 and 90 days; d) 400 patients are randomized into four groups: placebo, vitamin D, omega-3, vitamin D and omega-3; e) patients with myocardial infarction are followed for 5 years according to whether or not they received statins at discharge, as decided by their attending physician.
9. The column "inaltime" (height) of a raw file contains the values: 168, 172, 1.75, 159, 1680, 165, "1 m 70", 181. Identify the suspicious values, the type of error and what you do with each.
10. In a phase III study with 300 patients, no case of hepatotoxicity occurred. Using the rule of 3, estimate the upper limit of the 95% CI for the frequency of this adverse effect. What does this mean for phase IV?

11. In arm B of the mg2026 study, 16 of 53 patients lost at least 5% of their body weight. Calculate the proportion and the Wald 95% CI. Compare it with the Wilson interval.
12. For the B versus A comparison in the mg2026 study (16/53 versus 6/53), calculate the RR, the risk difference and the NNT. Compare them with the results for C versus A.

Common Pitfalls

- Confusing allocation concealment with blinding: the former protects the moment of enrollment and is always possible; the latter protects the assessment and is not always possible.
- Statistically testing baseline differences in an RCT and declaring randomization "failed" because of a $p < 0.05$ that occurred by chance.
- Changing the test or the primary endpoint after seeing the data in order to obtain $p < 0.05$ (for example, Kruskal-Wallis instead of ANOVA for SBP, $p = 0.046$ instead of 0.056).
- Interpreting a nonsignificant result from a superiority study as evidence of equivalence.
- Removing non-adherent patients and calling the resulting analysis "intention to treat".
- Replacing missing values with the group mean, or deleting patients with missing data without reporting how many and why.
- Collecting age or weight in classes instead of the exact value.
- "Correcting" an impossible value by guesswork (1540 becomes 154) without checking the source document.
- Treating a single Likert item as a numerical variable and calculating its mean without discussion.
- Assuming that a high Cronbach's α proves the validity of a questionnaire.
- Analyzing a cluster randomized trial as if the patients were independent.
- Reading observational associations in an RCT (obesity and diabetes) as if they were protected by randomization.

Key Points

- An interventional study assigns the exposure; randomization is the only method that also balances unknown confounders.
- The design determines the test: parallel groups, tests for independent samples; crossover or before-after, tests for paired data; three time points, Friedman or repeated-measures ANOVA; clusters, methods that take the intraclass correlation into account.
- The primary endpoint, the test, the significance level and the rule for choosing the alternative test are fixed in the protocol and in the public registry, before the data.
- Sample size depends on α, power, the minimal relevant effect and variability; an allowance for losses is added.
- The primary analysis in a superiority study is ITT; the per-protocol analysis is a sensitivity analysis.
- Ethics (Helsinki 2024, informed consent, the ethics committee, GCP, Regulation 536/2014, GDPR) is not a formality but a precondition of the study.
- The type of question in a questionnaire determines the type of variable; reliability (alpha, ICC) and validity (content, construct, EFA/CFA) are reported for any new or translated instrument.
- Data cleaning is done on a copy, with a change log, and corrections are made only from the source document.
- An RCT is reported according to CONSORT 2025, with a flow diagram, a Table 1 without p values, and effects expressed by their size and confidence interval.

Self-Assessment Questions

1. (SC) Which of the following characteristics defines an interventional study?

   a) following patients over time
   b) the existence of a control group
   c) assignment of the exposure by the investigator, according to the protocol
   d) prospective data collection
   e) the use of a validated questionnaire

2. (SC) In a crossover study with two treatments, the comparison of the values obtained by the same patient under the two treatments, for a numerical variable with normally distributed differences, is made with:
   a) the independent-samples t-test
   b) the paired-samples t-test
   c) the chi-square test
   d) one-way ANOVA
   e) the Mann-Whitney test

3. (SC) The role of allocation concealment is to:
   a) prevent the patient from finding out which treatment he or she is receiving
   b) prevent the assessor from finding out the patient's arm at the final visit
   c) prevent anticipation of the allocation at the time the patient is enrolled
   d) balance the number of patients in the arms
   e) reduce the number of patients lost to follow-up

4. (SC) For weight loss of at least 5% in the mg2026 study, the C versus A comparison gives an absolute risk difference of 18.3 percentage points. The NNT is:
   a) 2
   b) 3
   c) 6
   d) 18
   e) 29

5. (SC) For a questionnaire, the ICC calculated between two administrations 14 days apart in the same patients assesses:
   a) internal consistency
   b) test-retest reliability
   c) content validity
   d) convergent validity
   e) the factor structure

6. (MC) Regarding intention-to-treat (ITT) analysis in a superiority study, the following are true:
   a) it analyzes each patient in the arm to which he or she was randomized
   b) it preserves the benefit of randomization
   c) it tends to overestimate the biological effect of the treatment
   d) it is the recommended primary analysis
   e) it excludes patients who did not follow the protocol

7. (MC) Which of the following are surrogate endpoints?
   a) HbA1c in antidiabetic drug trials
   b) cardiovascular death
   c) LDL cholesterol in lipid-lowering drug trials

   d) blood pressure in antihypertensive drug trials
   e) stroke

8. (MC) In the "Table 1" of a randomized trial:
   a) numerical variables with a normal distribution are described by the mean and standard deviation
   b) skewed variables are described by the median and interquartile range
   c) it is recommended to test each difference between arms with a statistical test
   d) categorical variables are described by absolute frequency and percentage
   e) a large difference in an important prognostic variable may justify a sensitivity analysis specified in the plan
9. (MC) In the file mg2026_brut.csv, which of the errors can be detected with a frequency table of the categorical variables?
   a) sex coded as "f" and "Feminin"
   b) "Fumator " with a trailing space
   c) "Urban" with a capital letter
   d) $\mathrm{SBP}_0 = 1540$ mmHg
   e) a height of 1.72 entered in meters for one patient
10. (MC) Regarding the validation of a questionnaire, the following are true:

a) a KMO index of 0.93 indicates data that are very well suited to factor analysis
b) a nonsignificant Bartlett test is a necessary condition for factor analysis
c) parallel analysis helps choose the number of factors
d) a Cronbach's α above 0.95 may suggest redundant items
e) a high Cronbach's α demonstrates construct validity

Chapter 14. Review Examples: Integrated Interpretation of Statistical Results

Objectives

By the end of this chapter, the student will be able to:

- place any method learned on the map of the discipline and explain the connection between the clinical question, design, type of data, description, estimation, testing, association and decision;
- read the methods and results sections of a medical article using a step-by-step method;
- recognize the appropriate test for a given question and detect mismatches between design, data and analysis;
- interpret the p value, the effect size and the confidence interval together, distinguishing between statistical and clinical significance;
- recalculate simple indicators from published figures (confidence interval for a proportion, 2 x 2 table, PPV at a different prevalence, R^2 from r);
- write a results paragraph correctly and recognize incorrect wording;
- use the summary table of formulas from the lectures and solve multiple-choice questions in the format of the Romanian residency examination that combine several chapters.

14.1 The Map of the Discipline

Fourteen chapters look like fourteen separate subjects. They are not. All of them answer, from different angles, the same sequence of questions that every researcher, and every physician reading an article, asks: what do I want to find out, how did I obtain the data, what kind of data do I have, what do they look like, how precisely can I estimate what interests me, is the difference or association real, how large is it, what clinical decision do I make? Figure 14.1 draws this sequence.

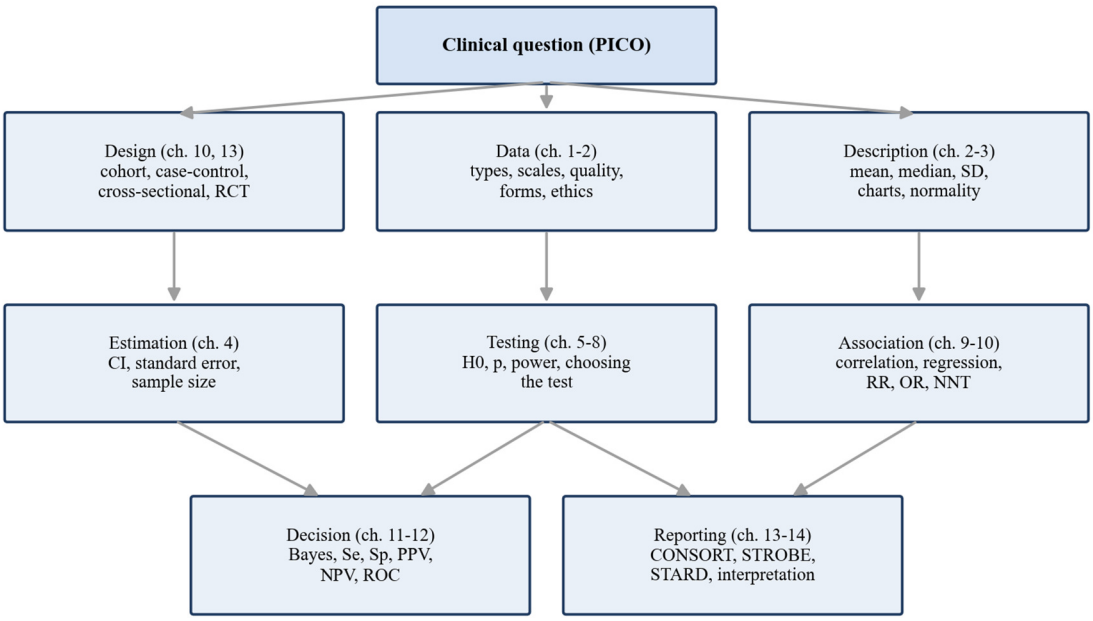


Figure 14.1. The map of the discipline: from the clinical question (PICO) to design (Chapters 10 and 13), data (1-2), description (2-3), estimation (4), testing (5-8), association (9-10), diagnostic decision (11-12) and reporting (13-14).

Everything starts from the clinical question, formulated as precisely as possible, preferably in PICO format [24]. The question determines the design. If I want to know whether an intervention works, the ideal design is the randomized trial (Chapter 13). If I want to find out whether an exposure is a risk factor, I use a cohort or, for rare diseases, a case-control study (Chapter 10). If I want to know how common a disease is, a cross-sectional study. If I want to judge a diagnostic test, an accuracy study against a reference standard (Chapter 12).

The design and the data collection instruments produce data of a certain type: numerical, ordinal, nominal (Chapter 2). The type of data determines the description: mean and standard deviation for symmetric numerical variables, median and interquartile range for skewed or ordinal ones, frequencies and percentages for categorical ones. The shape of the distribution (Chapter 3) determines whether we can use methods based on the normal distribution or whether we transform the data or switch to nonparametric methods.

From the sample, we want to say something about the population. Estimation (Chapter 4) gives the most likely value and the confidence interval, that is, the precision. Hypothesis testing (Chapter 5) tells us whether a difference or an association is compatible with chance. The choice of test (Chapters 6-8) depends on three things: the type of outcome variable, the number of groups and whether the groups are independent or paired. Associations between numerical variables are measured by correlation and regression (Chapter 9), and those between an exposure and a disease by the relative risk and the odds ratio (Chapter 10). For decisions at the bedside, Bayes' theorem (Chapter 11) and the indicators of diagnostic tests (Chapter 12) translate a test result into the probability that the patient has the disease. Finally, the results are reported according to the guideline appropriate to the design: CONSORT for randomized trials [56], STROBE for observational studies [25], STARD for diagnostic accuracy studies [140].

Table 14.1 condenses the map into a form you can use in the exam.

Table 14.1. From question to method: a synthesis of the discipline.

Question	Typical design	Outcome variable	Effect measure	Test	Chapter

Question	Typical design	Outcome variable	Effect measure	Test	Chapter
How common is the disease?	cross-sectional	dichotomous	prevalence with 95% CI	CI for a proportion	4, 10
Does the mean differ between two groups?	cross-sectional, cohort, RCT	numerical	difference in means, Cohen's d	independent t / Mann-Whitney	6, 7
Has the value changed after treatment?	before-after, crossover	numerical	mean paired difference	paired t / Wilcoxon	6, 7
Does the mean differ among 3 or more groups?	multi-arm RCT, cross-sectional	numerical	eta², Tukey differences	ANOVA / Kruskal-Wallis	6, 7
Does the value change across 3 time points?	repeated measurements	numerical or ordinal	differences between time points	repeated-measures ANOVA / Friedman	6, 7
Are two categorical variables associated?	any	nominal	difference in proportions, RR, OR	chi-square / Fisher	7, 10
Do two numerical variables vary together?	cross-sectional	numerical	r, rho, R^2, slope	Pearson / Spearman, regression	9
Is the exposure a risk factor?	cohort	dichotomous	RR, RD, NNT/NNH	chi-square, CI for RR	10
Is the exposure associated with the disease (rare disease)?	case-control	dichotomous	OR	chi-square, CI for OR	10
How good is the diagnostic test?	diagnostic accuracy	dichotomous	Se, Sp, PPV, NPV, LR, AUC	ROC, CI for proportions	11, 12
Does the intervention work?	RCT	any	difference, RR, NNT, with CI	depending on the type of variable	13

Two observations about the table. First, no column can be read on its own; the test depends on the design as well (paired or independent), not only on the type of variable. Second, the "effect measure" column is at least as important as the test column. An article that reports only p values, without the effect size and without a confidence interval, tells only half the story [70,74].

14.2 How to Read the Statistics of a Medical Article

In seminars, students often tell me that they read the abstract, skip the methods and go straight to the conclusion. This is exactly the reverse of the correct order. The conclusion is the least verifiable part of an article; the methods and the tables are what allow you to judge whether it is justified. The core textbooks of our discipline discuss the critical reading of the medical literature at length [18,19,64,150,151]; here we reduce it to a short method, to be used with every article. I propose a method based on twelve questions, which we will then apply in the case studies.

1. The title and the question. What question does the study ask? Is it a question of frequency, difference, association, prognosis, diagnosis or efficacy? Can it be written in PICO format?
2. The design. Who decided the exposure? Is it a cross-sectional, case-control, cohort (prospective or retrospective), RCT, before-after or diagnostic accuracy study? Does the label in the title correspond to what was actually done?

3. Population and sample. Who was included, from where, how were they selected, and how many refused or were lost? Is the sample representative of the patients to whom I want to apply the result?
4. Variables. What is the primary outcome and what type of variable is it? What are the exposures or groups? Were numerical variables dichotomized?
5. Description. Are the variables described correctly (mean ± SD for symmetric variables, median and IQR for skewed ones)? A standard deviation nearly equal to or larger than the mean of a positive variable (length of hospital stay, a biomarker) signals skewness.
6. Tests. Do the tests match the design and the type of data (independent or paired, two or more groups, parametric or nonparametric)? Are the tests named?
7. Effect size. How large is the difference or the association: difference in means, RR, OR, HR, r, AUC? Is it clinically relevant?
8. Confidence interval. How precise is the estimate? Does it include the "null" value (0 for differences, 1 for ratios)? Is it narrow enough to support a decision?
9. p value. Which hypothesis does it test? How many tests were performed? Were multiple comparisons taken into account?
10. Confounding and bias. Could a third factor explain the association? Was the analysis adjusted? For diagnostic studies, was there spectrum bias (severely ill patients compared with healthy volunteers)?
11. Limitations. Do the authors acknowledge them? Are obvious limitations missing (power, no control group, no blinding, missing data)?
12. Reporting guideline. Was CONSORT, STROBE or STARD followed? Are the mandatory elements (flow diagram, Table 1, CI for the main effect) present?

The method does not require sophisticated computers. It requires patience and three skills you have practiced all semester: recognizing the type of variable, matching the design with the test, and reading a confidence interval. In many cases you can check a figure from the article yourselves: from a percentage and n you compute the confidence interval of a proportion; from Se, Sp and prevalence you reconstruct the 2 x 2 table; from r you compute R^2. The case studies below do exactly this exercise.

The studies selected come from the research of colleagues in our discipline and from their clinical collaborations. We use only the figures reported in the published abstracts. When we point out a methodological problem, we do so as an exercise in critical reading: every article has limitations, and recognizing them is a form of respect for the authors' work, not a criticism of them.

Case study: the twelve-question method applied to a hemodialysis study

Marc and colleagues compared two cohorts of hemodialysis patients from 10 centers in Romania, assessed in 2015 (n = 1401) and in 2019 (n = 1698), before and after the national treatment program with direct-acting antivirals [39]. Let us go quickly through the method. The question: did the prevalence of hepatitis B and C virus infections change? The design: repeated cross-sectional, two independent samples, not the same patients followed over time. The description: mean age 59.7 ± 12.92 years in 2015 and 61 ± 12.93 years in 2019. The tests: comparison of means (t-test) and of proportions (chi-square) between cohorts. The results: age differs significantly ($p = 0.005$), although the difference is only 1.3 years, which is clinically irrelevant; with more than 1,400 patients in each cohort, even small differences become significant. One-year mortality (9.9% vs. 10.7%, $p = 0.46$) and HBV prevalence (4.7% vs. 4.8%, $p = 0.604$) do not differ, while HCV prevalence falls from 16.9% to 10.5% ($p < 0.0001$), a difference that is clinically relevant as well. Confounding: the cohorts also differ in dialysis vintage ($p < 0.0001$) and in the frequency of vascular calcifications ($p = 0.015$), so the

> populations are not identical; the fall in HCV is compatible with an effect of antiviral treatment, but the design cannot attribute it with certainty. What else would be useful: the differences in prevalence with 95% CIs. The exercise shows that the method can be applied in a few minutes and that it separates the clinically important result (HCV) from the merely statistical one (age).

14.3 Integrated Case Studies

14.3.1 Hepatology: A Noninvasive Diagnostic Test for Cirrhosis

Lupușoru and colleagues evaluated contrast-enhanced ultrasound with parametric imaging of arrival time (CEUS-PAT) for the noninvasive diagnosis of liver cirrhosis [2]. They included 64 subjects: 37 patients with cirrhosis and 27 healthy volunteers, mean age 58.98 ± 8.90 years, 56.3% men. The CEUS-PAT index was 0.83 ± 0.09 in patients with cirrhosis and 0.49 ± 0.11 in healthy volunteers ($p < 0.0001$) and correlated with vibration-controlled transient elastography (VCTE) with $r = 0.81$. At the cutoff > 0.7, the area under the ROC curve was 0.98 ($p < 0.001$), with Se = 89.19%, Sp = 100%, PPV = 100% and NPV = 87.1%.

For comparison, Moga and colleagues evaluated time-harmonic elastography in 165 patients, with VCTE as the reference method, and proposed two cutoffs [3]: a rule-out cutoff for cirrhosis (< 1.61 m/s), with Se = 90.7%, Sp = 66.6%, PPV = 55.7%, NPV = 93.9%, and a rule-in cutoff (> 1.83 m/s), with Se = 65.1%, Sp = 96.7%, PPV = 90.3%, NPV = 85.7%, both with an AUROC of 0.90 (95% CI 0.82-0.93); in the sample, 28.1% had cirrhosis (F4). Interobserver agreement was assessed by the ICC (0.94; 95% CI 0.89-0.97).

Interpretation questions, with model answers:

1. What design does the CEUS-PAT study have, and what bias threatens it? Model answer: it is a diagnostic accuracy study with a case-control design: patients with confirmed disease were compared with healthy volunteers. This design is exposed to spectrum bias: the difference between established cirrhosis and a healthy liver is much larger than the difference between early cirrhosis and advanced fibrosis, which is the real situation in which the physician needs the test. The performance (AUC 0.98, Sp 100%) is almost certainly better than what will be obtained in patients with chronic liver disease of varying severity.
2. Reconstruct the 2 x 2 table from the published figures. Model answer: 37 patients with the disease, Se 89.19%, hence 33 true positives and 4 false negatives (33/37 = 0.8919). 27 healthy subjects, Sp 100%, hence 27 true negatives and 0 false positives. PPV $= \frac{33}{33} =$ 100%; NPV $= \frac{27}{27+4} = \frac{27}{31} = 87.1\%$, exactly the reported value. The reconstruction confirms that the figures are consistent.
3. Does a specificity of 100% mean that the test never gives false positive results? Model answer: no. It means that no false positive occurred among the 27 volunteers. The rule of 3 gives the upper limit of the 95% CI for the proportion of false positives as approximately 3/27 = 11%, so the true specificity could be as low as about 89%. With 27 subjects, "0 out of 27" is compatible with a specificity far from perfect.
4. The prevalence of cirrhosis in the sample is 37/64 = 57.8%. How would the predictive values change at a prevalence of 20%? Model answer: assuming Se 89.2% and a more cautious specificity of 89% (the limit above), at a prevalence of 20% the PPV would be approximately 67% and the NPV approximately 97%; at the sample prevalence, the same Se and Sp would give a PPV of 92% and an NPV of 86%. Predictive values depend on

prevalence (Chapters 11-12), so the PPV = 100% in the article does not transfer automatically to the population of a gastroenterology practice.

5. Which test was probably used to compare CEUS-PAT between the two groups, and what additional information would have been useful? Model answer: two independent groups, a numerical variable with small standard deviations relative to the means: the independent-samples t-test (or Welch's test, if the variances differ). The difference in means with its 95% CI and confidence intervals for the AUC, Se and Sp would have been useful.
6. What does r = 0.81 with VCTE mean? Model answer: a strong positive correlation; $R^2 = 0.66$, that is, approximately two thirds of the variability of one method is linearly "explained" by the other. Correlation does not mean agreement: two methods can correlate perfectly and still give systematically different values. Agreement is assessed with the ICC or the Bland-Altman method [102].
7. Why does the time-harmonic elastography study propose two different cutoffs? Model answer: a single cutoff cannot have maximal sensitivity and specificity at the same time. The rule-out cutoff has high Se and NPV (90.7% and 93.9%): a result below the cutoff makes cirrhosis unlikely. The rule-in cutoff has high Sp and PPV (96.7% and 90.3%): a result above the cutoff makes cirrhosis very likely. Between the cutoffs lies a "gray zone", in which further investigations are needed. The LR+ at the rule-in cutoff is 0.651/(1 - 0.967) = 19.7, and the LR- at the rule-out cutoff is (1 - 0.907)/0.666 = 0.14.

14.3.2 Nutrition: Eating Habits and Weight in Students

Three studies by colleagues in our discipline followed students in Timișoara at the start of university life. Moleriu and colleagues conducted a cross-sectional analysis of 921 students aged 18-24 years [4]: 24.1% were overweight or obese (19.4% overweight and 4.7% obese). For two of the behavioral factors analyzed, the abstract reports adjusted odds ratios of 7.20 (95% CI 4.26-12.17) and 1.98 (95% CI 1.26-3.10) for overweight. A composite dietary risk score (DRS, 0-4) showed a dose-response relationship: the proportion with overweight rose from 19.8% at DRS = 0 to 50.0% at DRS = 4, with an adjusted OR of 1.43 per point (95% CI 1.23-1.66; p < 0.001).

Piroș and colleagues studied 693 students cross-sectionally for prediabetes risk [6]. The prevalence of prediabetes was 21.1% (that of diabetes 1.4%). Across dietary patterns derived by cluster analysis, prediabetes was 15.4% among students with a healthy diet, 20.0% among those with a mixed diet and 27.3% among those oriented toward fast food (χ^2, p = 0.003), with an OR of 1.78 per category (95% CI 1.38-2.30; p < 0.001) and a p for trend < 0.001.

Muntean and colleagues prospectively followed a cohort of 921 students during their first university year [5]. The mean change in BMI was -0.45 ± 1.07 kg/m² (95% CI -0.52 to -0.38; p < 0.001): 48.2% had a decrease in BMI, 38.7% remained stable and 13.1% gained weight; a clinically significant gain occurred in 7.2% of students. One of the chrononutrition behaviors was associated with an increase a BMI $\left(\beta = +\frac{0.22\ \text{kg}}{m^2};\ p = 0.004\right)$.

Interpretation questions, with model answers:

1. Compute the 95% CI for the prevalence of overweight (24.1% of 921). Model answer: $\text{SE} = \sqrt{\frac{0.241 \times 0.759}{921}} = 0.0141$; $\text{CI}\,95\% = 24.1\% \pm 1.96 \times 1.41\% = 21.3\%$ to 26.9%. With almost 1000 students, the estimate is precise (±2.8 percentage points).
2. How do you interpret aOR = 7.20 (95% CI 4.26-12.17)? Model answer: after adjustment for the other factors in the model, the odds of overweight are about 7 times higher among students exposed to that factor. The interval is wide (from 4 to 12), but it lies far from 1, so the association is strong and significant. Caution: with an outcome prevalence of 24%,

the OR overestimates the prevalence ratio; an OR of 7 does not mean that overweight is 7 times more frequent.

3. What does OR = 1.43 per DRS point mean, and what is the OR between DRS = 4 and DRS = 0? Model answer: each additional point on the risk score increases the odds of overweight by 43%. The effect is multiplicative, not additive: between 4 points and 0 points, $\mathrm{OR} = 1.43^4 = 4.18$ (assuming a log-linear relationship, as the logistic model imposes).
4. Can the authors conclude that a convenience-oriented diet causes obesity? Model answer: not on the basis of a cross-sectional design. Exposure and outcome are measured at the same time, so we do not know which came first (perhaps overweight students change their diet); there may be unmeasured confounders (sleep, stress, income). The dose-response relationship and the strength of the association support, according to the Bradford Hill criteria [108], a possible causal relationship, but they do not prove it.
5. Check the 95% CI of the mean change in BMI from the prospective study. Model answer: $\mathrm{SE} = \frac{1.07}{\sqrt{921}} = 0.0353$; $\mathrm{CI}\,95\% = -0.45 \pm 1.96 \times 0.0353 = -0.45 \pm 0.069$, that is, -0.52 to -0.38, exactly what the authors report. The appropriate test is the paired t-test (the same student at the beginning and at the end of the year).
6. How can a small mean decrease (-0.45 kg/m²) be reconciled with the fact that 13.1% of students gained weight? Model answer: the mean describes the center of the distribution of changes, not each student. The standard deviation (1.07) is more than twice the mean change, so individual changes vary considerably in both directions. That is why the authors supplemented the linear analysis of BMI with a logistic analysis of clinically significant gain: the same variable, viewed as continuous and as dichotomized, answers different questions.
7. In the prediabetes study, why is a "p for trend" also reported? Model answer: the three dietary patterns have a logical order (healthy, mixed, fast food), and the proportions increase monotonically (15.4%, 20.0%, 27.3%). The ordinary chi-square test (p = 0.003) tests only whether the proportions differ, without taking the order into account; the trend test specifically tests a monotonic increase and usually has more power for such a hypothesis.

14.3.3 COVID-19 and Diabetes: Three Studies, Three Types of Analysis

Cioca and colleagues retrospectively compared the course of COVID-19 in 486 patients: 200 without diabetes, 62 with type 1 diabetes and 224 with type 2 diabetes [8]. Severe forms occurred in 33.5% with type 2 diabetes vs. 25.8% with type 1, mortality was 11.6% vs. 8.1%, length of stay 11.6 ± 7.0 days vs. 9.1 ± 5.8 days, and blood glucose at admission 178.3 ± 34.7 vs. 164.8 ± 39.6 mg/dL; the odds ratios were 2.24 for intensive care admission and 2.46 for mechanical ventilation.

Albai and colleagues validated the ISARIC-4C score for severity and mortality in 159 patients with type 2 diabetes and COVID-19 [75]. Mortality was 25.15%. The area under the ROC curve was 0.875 (95% CI 0.820-0.930), and the score correctly classified 77.36% of patients, with Se 87.80% (95% CI 87.66-87.93), Sp 73.72% (95% CI 73.48-73.96), PPV 53.73% and NPV 94.56%.

Dănilă and colleagues compared 168 patients with COVID-19 and diabetes with 168 patients without diabetes, matched for sex, BMI and severity [105]. For the dNLR index, at the cutoff 2.685: Se 70.312%, Sp 65.978%, AUC 0.624 ($p < 0.001$); for ALRI, at the cutoff 0.812: Se 76.429%, Sp 69.541%, AUC 0.752. For the SII index at 7 days, a hazard ratio of 2.62 (95% CI 1.29-5.04) was reported.

Interpretation questions, with model answers:

1. What does a length of stay of 11.6 ± 7.0 days tell us about its distribution, and how should it have been described? Model answer: the standard deviation is 60% of the mean, and the mean minus two standard deviations gives a negative value (11.6 - 14.0 = -2.4 days), which is impossible. The distribution is right-skewed (a few patients with very long stays). The appropriate description is the median with the IQR, and the appropriate comparison between groups is the Mann-Whitney test (or Kruskal-Wallis for three groups).
2. What is missing from the reporting of OR = 2.24 for intensive care admission? Model answer: the confidence interval (and the p value). Without a CI we do not know whether the association is precise or compatible with no effect. Neither the reference group nor whether the OR is crude or adjusted is stated.
3. With three groups and a numerical variable such as blood glucose at admission, which test would you use? Model answer: one-way ANOVA if the distributions are approximately normal and the variances homogeneous (Levene), followed by post hoc comparisons (Tukey); otherwise Kruskal-Wallis with corrected Mann-Whitney comparisons. Comparing only type 1 with type 2 is done with the independent-samples t-test or Mann-Whitney.
4. Is the confidence interval reported for the sensitivity of the ISARIC-4C score (87.66-87.93) plausible? Model answer: no. With a mortality of 25.15% among 159 patients, approximately 40 patients died. The Wald interval for a proportion of 87.8% based on 40 patients is $87.8\% \pm 1.96 \times \sqrt{\frac{0.878 \times 0.122}{40}} = 87.8\% \pm 10.1\%$, that is, approximately 77.7-97.9%. The published interval, 0.27 percentage points wide, would correspond to a sample of tens of thousands of patients. It was probably computed some other way (for example, by incorrectly reported resampling methods). The exercise shows why it pays to check intervals.
5. Check the reported PPV for ISARIC-4C from Se, Sp and prevalence. Model answer: $\text{PPV} = \frac{0.2515 \times 0.878}{0.2515 \times 0.878 + 0.7485 \times 0.2628} = \frac{0.2208}{0.4175} = 0.529$, that is, approximately 53%, close to the reported 53.73%. $\text{NPV} = \frac{0.7485 \times 0.7372}{0.7485 \times 0.7372 + 0.2515 \times 0.122} = 0.947$, close to 94.56%. The score is useful mainly for identifying low-risk patients (high NPV).
6. Compare the AUC of 0.624 (dNLR) with 0.752 (ALRI) and 0.875 (ISARIC-4C). Model answer: an AUC of 0.624 indicates poor discrimination (0.5 would be chance); even though it is significantly different from 0.5 ($p < 0.001$), dNLR is not clinically useful as a stand-alone test. ALRI has acceptable discrimination, ISARIC-4C good discrimination. The statistical significance of an AUC says nothing about its usefulness.
7. What problem is raised by reporting a hazard ratio in a study described as cross-sectional, and by reporting Se with three decimals? Model answer: the hazard ratio comes from time-to-event analysis and presupposes following patients over time, hence a longitudinal design; a purely cross-sectional design cannot produce an HR. The figure 70.312% suggests illusory precision: with 168 patients per group, the third decimal is meaningless; one reports 70.3% or even 70%, together with the 95% CI.

14.3.4 Emergency Medicine: Patient Sex, Thrombolysis Times and Stroke Biomarkers

Popa and colleagues retrospectively analyzed 213 patients (100 women, 113 men) treated in the emergency department for suspected bacterial infection [9]. The women were older (median 77.0 vs. 67.0 years, $p < 0.001$), and a urinary source of infection was more frequent in women (51.0% vs. 34.5%, $p = 0.015$; OR 2.14, 95% CI 1.20-3.81, $p = 0.010$; Benjamini-Hochberg adjusted p =

0.039). A severe outcome occurred in 14 of 100 women (14.0%) and 21 of 113 men (18.6%; OR 0.71, 95% CI 0.34-1.49); in-hospital mortality in 7 women (7.0%) and 9 men (8.0%; OR 0.87, 95% CI 0.31-2.43); after adjustment, OR 0.80 (95% CI 0.27-2.35; p = 0.690).

In the same department, a study with the same first author followed patients with ischemic stroke arriving within the first 4.5 hours [62]: of 355 initial patients, 258 were eligible, and 46 received intravenous thrombolysis (rate 17.83%). In the between-group comparisons, the abstract reports a median of 100 minutes for the time from onset to arrival at the emergency department ($p < 0.001$) and 15.5 minutes for the time from arrival to CT ($p = 0.009$), as well as a clinical variable with values of 9 ± 4.94 vs. $13.85 \pm 2.41 (p < 0.001)$. A third study evaluated the biomarkers GFAP and UCH-L1 in 85 patients with a stroke alert (69 with stroke, 16 without; 37 ischemic, 32 hemorrhagic) [61]: GFAP 142.91 ± 102.19 pg/mL vs. 37.76 ± 19.92 pg/mL($p < 0.001$), UCH-L1 1307.68 ± 967.54 vs. 189.81 ± 92.69 pg/mL($p < 0.001$); GFAP distinguished hemorrhagic from ischemic stroke with an accuracy of 94.2% at the cutoff of 77.15 pg/mL.

Interpretation questions, with model answers:

1. Check the OR for a severe outcome (14/100 women, 21/113 men). Model answer: the odds are 14/86 = 0.163 in women and 21/92 = 0.228 in men; OR $= \frac{0.163}{0.228} = 0.71$, exactly the reported value. Similarly, for mortality, (7/93)/(9/104) = 0.075/0.087 = 0.87.
2. How do you interpret OR 0.71 (95% CI 0.34-1.49)? Model answer: the point estimate suggests lower odds of a severe outcome in women, but the interval includes 1 and ranges from a two-thirds reduction to an increase of almost 50%. The data are compatible both with an important protective effect and with an increased risk. The correct conclusion is not "sex does not influence prognosis" but "the study lacks the precision to answer"; with only 35 severe events, power is low.
3. Why is a Benjamini-Hochberg adjusted p of 0.039 reported in addition to p = 0.015? Model answer: the authors compared women and men on many characteristics. The correction for multiple comparisons (here, control of the false discovery rate) raises the p value to take the number of tests into account. The fact that the difference in urinary source remains below 0.05 after correction makes it more credible.
4. Why were the times reported as medians? Model answer: times to an event are almost always right-skewed (a few patients arrive very late). The median is not influenced by these extreme values. Two independent groups are then compared with the Mann-Whitney test.
5. Compute the 95% CI for the thrombolysis rate. Model answer: $p = \frac{46}{258} = 0.178$; SE $= \sqrt{\frac{0.178 \times 0.822}{258}} = 0.0238$; CI 95% $= 17.8\% \pm 4.7\%$, that is, 13.2% to 22.5%.
6. What do the standard deviations of the biomarkers GFAP and UCH-L1 in stroke patients suggest? Model answer: the standard deviation is almost as large as the mean (102.19 vs. 142.91 for GFAP; 967.54 vs. 1307.68 for UCH-L1). For a variable that cannot be negative, this indicates a strongly skewed distribution. The t-test may be inappropriate; the Mann-Whitney test or log transformation of the values (Chapter 3) would be more suitable, and the description should include the median and IQR.
7. An accuracy of 94.2% for distinguishing hemorrhagic from ischemic stroke: what else should be reported? Model answer: 94.2% of 69 patients means 65 correctly classified. Accuracy alone is not enough: Se and Sp are needed (how well hemorrhages and ischemic strokes are each identified), with 95% CIs, and the cutoff of 77.15 pg/mL, derived in the

same sample, must be validated in an independent sample, because cutoffs chosen as "optimal" on one's own data overestimate performance.

14.3.5 Validating a Questionnaire and Analyzing Scores

In Chapter 13 we discussed the validation of the TAM questionnaire for telemonitoring in patients with cirrhosis [10] and the validation of the Romanian version of the PSPSQ 2.0 in dermatology [11]. We return to them here together with a third study, which uses an already validated questionnaire to answer a clinical question. Frandeş and colleagues studied digital engagement in diabetes care in 304 adults with diabetes [7]. The scale used had a Cronbach's $\alpha = 0.89$; 35.5% of participants had high digital engagement. Scores were reported as medians and quartiles, for example 1.71 [1.71-2.39] vs. 3.86 [3.86-4.43] between groups, and a multivariable beta regression, with linear regression as a sensitivity analysis, yielded coefficients such as $\beta = -0.147$ (95% CI -0.219 to -0.075; $p < 0.001$), $\beta = 0.996$ (95% CI 0.859-1.134; $p < 0.001$) and $\beta = 0.069$ (95% CI 0.006-0.132; $p = 0.033$).

Interpretation questions, with model answers:

1. In the TAM study, $\alpha = 0.934$ and $\omega = 0.950$. What do these figures say, and what do they not say? Model answer: they say that the items of the scales are strongly intercorrelated (very good internal consistency). They do not say whether the scale measures acceptance of telemonitoring (validity), nor whether patients would answer the same way on a second administration (test-retest reliability).
2. A CFA model has CFI = 0.999 and RMSEA = 0.109 (PSPSQ 2.0). How is that possible? Model answer: the indices measure different aspects of fit. The CFI compares the model with a model of complete independence of the items; when the items are highly correlated with one another ($\alpha = 0.978$), almost any structured model looks much better than independence, so the CFI comes out very high. The RMSEA measures approximation error per degree of freedom and penalizes models with few degrees of freedom and moderate samples; 0.109 exceeds the usual threshold of 0.08. Reporting both indices and discussing the contradiction is the correct practice.
3. In the TAM model, "the model explained 69.5% of the variance in intention to use". Which concept from Chapter 9 does this figure relate to? Model answer: the coefficient of determination R^2. In simple regression, R^2 is the square of the correlation; here, in a structural model with several predictors, $R^2 = 0.695$ means that the predictors together explain approximately 70% of the variability in intention. The remaining 30% is due to factors not included and to measurement error.
4. What does a median equal to the first quartile, as in 1.71 [1.71-2.39], mean? Model answer: all values between the 25th and the 50th percentile, that is, at least a quarter of the participants in that group, are equal to 1.71, very probably a boundary value of the score. The distribution has a floor effect and is strongly skewed. This is exactly the situation in which the mean and standard deviation would describe the data poorly, and the authors' choice to report medians and to use a regression suited to bounded scores (beta regression) is justified.
5. How do you interpret $\beta = 0.069$ (95% CI 0.006-0.132; $p = 0.033$)? Model answer: the association is positive and significant at the 0.05 level, and the interval does not include 0. The lower limit, 0.006, is however very close to 0: the effect could be negligible. By comparison, $\beta = -0.147$ (CI -0.219 to -0.075) is estimated more precisely and lies far from 0. Magnitude and precision, not just p, distinguish the two results.

6. Why is a "sensitivity analysis" performed with linear regression? Model answer: to check whether the conclusion depends on the choice of model. If beta regression and linear regression give coefficients with the same direction and significance, the result is robust to this methodological choice.

14.3.6 Heart Failure: Three-Group Comparisons, Correlations and Confounding

Stoichescu-Hogea and colleagues compared global longitudinal strain (GLS) in three groups: 62 patients with heart failure with preserved ejection fraction (HFpEF), 46 hypertensive patients and 40 age-matched controls [90]. GLS differed among the three groups ($p < 0.001$): controls had -20.2 ± 1.4%, hypertensive patients -18.4 ± 3.0% ($p = 0.031$ vs. controls) and patients with HFpEF -17.6 ± 2.3% ($p < 0.001$ vs. controls). GLS correlated with an ordinal clinical variable with Spearman's rho = 0.423 ($p < 0.001$), and the optimal cutoff of -19.35% had an AUC of 0.833 ($p < 0.001$).

Buleu and colleagues compared 60 patients with ischemic heart failure with reduced ejection fraction (< 40%) with 60 age-matched healthy controls [87]. Osteoprotegerin (OPG) was 4.7 ± 0.25 vs. 1.3 ± 0.67 ng/mL, and vitamin 25-OH-D3 20.49 ± 7.31 vs. 37.09 ± 4.59 ng/mL (both $p < 0.001$). OPG correlated with left ventricular end-diastolic volume ($r = 0.862$), with ejection fraction ($r = -0.832$), with pulse wave velocity ($r = 0.833$), with vitamin D ($r = -0.636$) and with glomerular filtration rate ($r = -0.488$), all with $p < 0.001$.

Finally, the HRR study discussed in Chapter 13 [23] showed an association with prognosis in the Kaplan-Meier analysis (log-rank $p = 0.027$), which disappeared after Cox adjustment for age, sex, NT-proBNP, ejection fraction and filtration rate ($p = 0.240$).

Interpretation questions, with model answers:

1. Which test was probably used to compare GLS across three groups, and what should be checked? Model answer: one-way ANOVA, with post hoc pairwise comparisons. The standard deviations differ considerably (1.4 in controls, 3.0 in hypertensive patients), so homogeneity of variances (Levene's test) is doubtful; a robust alternative (Welch's ANOVA with Games-Howell post hoc tests) or Kruskal-Wallis would be preferable if Levene's test rejects homogeneity.
2. Why are the groups compared two at a time with a post hoc test and not with three t-tests? Model answer: three t-tests, each at $\alpha = 0.05$, raise the overall probability of a type I error to approximately $1 - 0.95^3 = 14\%$. Post hoc tests (Tukey, Bonferroni) control this error across the whole set of comparisons.
3. How much of the variability does rho = 0.423 explain, and how does the interpretation differ from r = 0.862 in the OPG study? Model answer: $\text{rho}^2 \approx 0.18$ (approximately 18%; for Spearman, interpreting the square is only approximate); r = 0.862 gives $R^2 = 0.74$. The first is a moderate correlation, the second a strong one. Both are significant, but the strength of the association differs greatly.
4. OPG correlates strongly with ejection fraction ($r = -0.832$). Can we say that OPG lowers ejection fraction? Model answer: no. Correlation measures linear association, not causal direction. It is possible that ventricular dysfunction raises OPG, that both are determined by a third factor (atherosclerosis, inflammation, renal function, itself correlated with OPG), or that the relationship is bidirectional.
5. The OPG study has a case-control design with healthy controls. What are the implications for correlations computed on the whole sample? Model answer: if the correlations are computed on all 120 subjects, they combine two very different groups (OPG 4.7 vs. 1.3). The difference between groups alone can create a strong correlation in the combined

sample, even if the correlation within each group is weak (Chapter 9, the effect of distinct groups on r). It is worth checking whether the correlations were computed in patients only.

6. What common lesson do the HRR study and the OPG study teach? Model answer: a crude association (log-rank p = 0.027; strong correlations) can be produced or amplified by confounders. Multivariable adjustment (Cox, multiple regression) is the step that separates the marker's own association from the association it inherits from other factors. In the HRR study, after adjustment, the association was no longer significant.

14.3.7 Pharmacology in Older Adults: Potentially Inappropriate Prescriptions

Baneu and colleagues conducted a cross-sectional analysis of 1498 electronic prescriptions issued for older adults in urban areas of Timiș and Arad counties, using the STOPP/START criteria, version 2 [38]. Of the prescriptions, 57% were for women; the mean age of the patients was 74.1 ± 6.95 years, and the mean number of medicines per prescription was 4.7 ± 1.51. Among drug classes, the abstract mentions statins, at 47.4%. The prevalence of potentially inappropriate medications was 18.58%, and that of potential prescribing omissions 49.2%.

As a point of comparison for paired data in pharmacology, Gherbon and colleagues retrospectively followed 200 patients with type 2 diabetes for one year after the initiation of an SGLT-2 inhibitor [59]: blood glucose fell from 180.00 mg% (IQR 154.50-207.00) to 130.00 mg% (IQR 117.50-150.00), HbA1c from 8.40% (IQR 7.98-9.15) to 7.30% (IQR 6.90-7.95), both p < 0.001, and waist circumference from 107.05 (± 16.39) to $102.50(\pm 15.11)$ cm$(p = 0.042)$.

Interpretation questions, with model answers:

1. What is the unit of analysis in the STOPP/START study, and why does it matter? Model answer: the prescription, not the patient. If a patient has several prescriptions, the observations are not independent, and confidence intervals computed as if for 1498 independent patients will be too narrow. The correct interpretation is "18.58% of prescriptions", not necessarily "18.58% of older adults".
2. Compute the 95% CI for the prevalence of potentially inappropriate medications. Model answer: $\text{SE} = \sqrt{\frac{0.1858 \times 0.8142}{1498}} = 0.0100$; $\text{CI}\,95\% = 18.6\% \pm 2.0\%$, that is, 16.6% to 20.5% (with the caveat about independence from question 1).
3. Is the description "4.7 ± 1.51 medicines per prescription" appropriate? Model answer: the number of medicines is a discrete numerical variable (0, 1, 2...). With a mean of 4.7 and an SD of 1.5, the distribution is probably not very skewed, so the mean and SD are acceptable; the median and IQR would usefully complete the description. The mean minus two SDs (1.7) remains plausible.
4. What type of study is it, and what conclusions does it allow? Model answer: a descriptive cross-sectional study. It allows prevalences to be estimated and prescribing practice to be described. It does not allow conclusions about the clinical consequences of inappropriate prescriptions (hospitalizations, falls), because it does not follow patients over time.
5. In the SGLT-2 study, why are medians with IQRs reported, and which test is appropriate? Model answer: paired data (the same patient before and after one year) with a skewed distribution are described by medians and compared with the Wilcoxon signed-rank test for paired samples. The abstract uses the word "mean" for values that are in fact medians; when reading, figures given with IQRs should be read as medians.
6. The SGLT-2 study has no control group: what alternative explanations are there for the fall in blood glucose? Model answer: regression to the mean (patients start a new

treatment precisely when their blood glucose is highest), other changes in treatment or lifestyle during the year, and better adherence among those who remained under follow-up. A before-after design shows a change but cannot attribute it with certainty to the drug.

14.3.8 Oncology: The Pandemic and Colorectal Cancer, BRCA Mutations and Survival

Feier and colleagues compared colon cancer care before and during the COVID-19 pandemic in 867 patients operated on in a university hospital over four years [55]. During the pandemic, elective procedures decreased significantly ($p < 0.001$), emergency procedures increased ($p < 0.001$), and in the second pandemic year the proportion of stage IV cases rose ($p = 0.045$). An inverse correlation was reported between patient age and duration of surgery ($r = -0.088$, $p = 0.01$).

Petruşan and colleagues prospectively followed 133 patients with high-grade serous ovarian carcinoma for 24 months [117]. Pathogenic BRCA mutations were identified in 39.1% of the patients. The abstract reports differences in proportions between groups (50% vs. 77.77%, $p = 0.001$; 32.69% vs. 51.85%, $p = 0.017$). In univariable Cox analysis, BRCA status was associated with progression-free survival (HR 0.52, 95% CI 0.27-0.99, $p = 0.048$), but in the multivariable model the association was no longer significant (HR 0.57, $p = 0.124$).

Interpretation questions, with model answers:

1. The correlation $r = -0.088$ is significant ($p = 0.01$). Is it also important? Model answer: $r^2 = 0.0077$, that is, age explains less than 1% of the variability in the duration of surgery. The significance comes from the large sample (867 patients): with a large n, even negligible correlations become significant. This is the classic example of the difference between statistical and clinical significance.
2. Which test was probably used to compare the proportion of emergency procedures between periods? Model answer: the chi-square test for contingency tables (period x type of procedure), or Fisher's test if some expected frequencies are small. For a comparison across more than two periods, a chi-square test on a k x 2 table.
3. The result for stage IV ($p = 0.045$) should be read with caution. Why? Model answer: it is one of many tests (several periods, several variables), and the value is close to the threshold. Without correction for multiple comparisons, such a result may be a false positive. The size of the difference and its CI would help interpretation.
4. How do you read HR = 0.52 (95% CI 0.27-0.99)? Model answer: at any point during follow-up, the instantaneous rate of progression in patients with a BRCA mutation is approximately half that in patients without the mutation, that is, a relative reduction of 48%. The upper limit of the interval, 0.99, is however very close to 1: the data are also compatible with a reduction of only 1%. And $p = 0.048$ is borderline.
5. Why does significance disappear in the multivariable model (HR 0.57, $p = 0.124$)? Model answer: there are two explanations, which are not mutually exclusive. The first is confounding: part of the crude association is due to other factors (stage, type of surgery, chemotherapy) related both to BRCA status and to prognosis. The second is precision: the adjusted HR (0.57) is very close to the crude one (0.52), so the effect has not disappeared, but adding covariates widened the confidence interval. With 133 patients, the multivariable model has limited power. The correct wording: the association does not remain significant after adjustment, and its magnitude remains uncertain.
6. Which descriptive method and which test correspond to the analysis of progression-free survival? Model answer: Kaplan-Meier curves for description, the log-rank test for comparing two curves and the Cox model for estimating the HR with adjustment. All of

them take into account censored patients (who had not progressed by the end of follow-up or who left the study), which a simple comparison of proportions does not do.

14.3.9 An RCT-Type Study: Aloe ferox and an Asthma App

In Chapter 13 we presented the placebo-controlled study of Aloe ferox in people with obesity, reported according to CONSORT [85], and the controlled pilot study of the mobile app for asthma [94]. We now use them to review the principles of clinical trials. Let us recall the main data: in the first study, 20 people with obesity received diet plus Aloe ferox (2 capsules a day, 460 mg), and 20 matched controls received diet plus placebo, for 3 months; the abstract reports p values for BMI (0.03 and 0.002), total cholesterol (0.032), LDL (0.01 and 0.039), fasting plasma glucose (0.018 and < 0.001) and diastolic blood pressure (0.006). In the second, 93 patients with asthma were compared using Mann-Whitney ($p < 0.001$), Friedman with Bonferroni correction ($p < 0.001$) and chi-square for the proportion of exacerbations (10.30% vs. 46.30%, $\chi^2(1) = 13.707$, $p < 0.001$).

Interpretation questions, with model answers:

1. Which CONSORT elements should you find in the article on Aloe ferox? Model answer: how the randomization sequence was generated, allocation concealment, who was blinded (patients, assessors), the definition of the primary outcome, the sample size calculation, the flow diagram with losses, Table 1 by group, the main effect with its size and 95% CI, adverse effects and trial registration [56].
2. What does "matched controls" mean in a placebo study, and how does it affect the analysis? Model answer: the controls were chosen to resemble the treated patients on certain characteristics (age, sex, BMI). If allocation was done by matching rather than by randomization, comparability is guaranteed only for the matched variables. If matching was done in pairs, the analysis should take the paired structure into account.
3. Why are there two series of p values for the same variables (BMI 0.03 and 0.002)? Model answer: one series probably corresponds to the between-group comparison and the other to the before-after comparison within a group. Only the between-group comparison at 3 months (or the comparison of changes between groups) tests the effect of Aloe ferox, because the diet was common to both. A significant decrease in the treated group, without comparison with placebo, may be the effect of the diet.
4. With 20 patients per group, what effect size could the study detect with 80% power? Model answer: from the formula for two means, $d = (1.96 + 0.84) \times \sqrt{\frac{2}{20}} = 2.8 \times 0.316 = 0.89$. The study had sufficient power only for large effects (almost one standard deviation). Moderate, clinically relevant effects could go undetected.
5. Compute the relative risk of exacerbations from the app study, if 10.30% is the proportion in the app group. Model answer: $\text{RR} = \frac{10.30}{46.30} = 0.22$, that is, a relative risk reduction of 78% and an absolute reduction of 36 percentage points (NNT ≈ 3). The interpretation must, however, be accompanied by the 95% CI and by a check of how allocation was done: in a nonrandomized pilot, patients who accept the app may be more adherent from the outset.
6. Why was the Bonferroni correction needed for the Friedman test? Model answer: the global Friedman test says only that at least one time point differs from the others. To find out which one, comparisons are made between pairs of time points (for example, Wilcoxon); with several comparisons, the threshold for each is divided by their number, to keep the overall type I error at 0.05.

14.3.10 A Complete Reanalysis of the mg2026 Dataset

We end the case studies with the dataset you have used all semester. The questions below go through the whole discipline on the same data; the answers repeat the figures computed in the previous chapters, so that you can see how they connect.

1. Describe age, BMI and triglycerides, justifying the choice of indicators. Model answer: age has a symmetric distribution (skewness 0.01; Shapiro-Wilk p = 0.687): mean 56.29 years, SD 8.77. BMI is right-skewed (skewness 1.42; excess kurtosis 3.68; Shapiro-Wilk W = 0.904, p < 0.001): median 28.25 kg/m², IQR 26.2-31.4. Triglycerides are skewed (skewness 0.95; p < 0.001): median 151.5 mg/dL, IQR 107-203.25; after log transformation the distribution becomes nearly normal (skewness -0.28; p = 0.200), and the geometric mean is 144.0 mg/dL (95% CI 132.95-156.02).
2. Estimate the population mean SBP and interpret it. Model answer: mean 131.13 mmHg, SD 13.88, SE 1.097; 95% CI 128.96-133.30 mmHg. With 95% confidence, the mean SBP of the population from which the sample comes lies between these limits. The interval for individual values is something else: approximately 95% of patients have an SBP between 103.9 and 158.3 mmHg.
3. Does SBP differ between women and men? Model answer: women 131.37 (SD 14.47; n 86), men 130.85 (SD 13.25; n 74); Levene p = 0.529; Welch's t-test t = -0.24, df 157.4, p = 0.81; difference M-F -0.5 mmHg (95% CI -4.9 to 3.8); d = -0.04. We do not reject H0. The interval shows that a difference larger than about 5 mmHg in either direction is unlikely, which is more useful information than a mere "not significant".
4. Did blood glucose decrease over the 6 months? Model answer: from 105.4 to 95.2 mg/dL; mean difference 10.19 mg/dL (SD 11.57; 95% CI 8.38-11.99); the differences are normally distributed (Shapiro-Wilk p = 0.725), so a paired t-test: t = 11.14, df 159, p < 0.001; dz = 0.88. Caution: the overall decrease cannot be attributed to the intervention; to assess the effect of the intervention we compare the arms.
5. Does the decrease in blood glucose differ between arms? Model answer: ANOVA F(2, 157) = 22.93, p < 0.001, $\eta^2 = 0.226$; Tukey: B-A 8.1 mg/dL (CI 3.4-12.8), C-A 13.3 (8.6-18.0), C-B 5.2 (0.6-9.9). All arms differ from one another; the largest effect is seen with the app plus teleconsultation combination.
6. Is BMI associated with SBP? Model answer: r = 0.405, p < 0.001 (Spearman's rho 0.372, preferable given the skewness of BMI); $R^2 = 0.164$. Regression line: SBP = 89.47 + 1.435 x BMI (95% CI for the slope 0.925-1.944): each additional kg/m² is associated with an SBP higher by 1.4 mmHg; at a BMI of 30, the predicted SBP is 132.5 mmHg. In multiple regression with age, sex and smoking, the BMI coefficient remains almost unchanged (1.389, p < 0.001), and the model explains 30.8% of the variability in SBP (adjusted R^2 0.286); at the same BMI, age and sex, current smokers have an SBP 6.7 mmHg higher than former smokers (p = 0.012).
7. Is obesity associated with diabetes? Model answer: prevalence of diabetes 33.3% among obese patients vs. 10.4% among non-obese patients; prevalence ratio 3.21 (95% CI 1.64-6.31); OR 4.32 (95% CI 1.86-10.03); $\chi^2 = 12.70$, p < 0.001. Adjusted for age and sex (logistic regression), OR 6.01 (95% CI 2.35-15.39). Cross-sectional design: association, not causation.
8. How useful is FINDRISC, and how does the probability of diabetes change according to its result? Model answer: AUC 0.852 (95% CI 0.762-0.943). At the cutoff ≥ 12: Se 75.9%, Sp 74.8%, LR+ 3.01, LR- 0.32. Starting from a prevalence of 18.1% (pre-test odds 0.181/0.819 = 0.221), a positive score gives post-test odds of 0.221 x 3.01 = 0.665, that is,

a probability of 0.665/1.665 = 40%, identical to the PPV; a negative score gives odds of 0.221 x 0.32 = 0.071, a probability of 6.6%, that is, 1 - NPV. Bayes' theorem (Chapter 11) and the 2 x 2 table (Chapter 12) say the same thing. In a population with a prevalence of 5%, the same test would have a PPV of only 13.7%.

14.4 How to Write the Results Paragraph

A well-reported statistical result answers four questions: what was compared (groups, variable, unit), how large the difference or association is, how precise the estimate is, and how compatible it is with chance. The natural order is: the description in each group, the effect size with its 95% CI, the test with its statistic, degrees of freedom and p value. The clinical conclusion comes afterward, in the discussion, not mixed in with the figures.

Here are a few pairs of examples, all based on the mg2026 data.

Example 1, weak: "Blood glucose decreased significantly in arm C ($p < 0.05$)." The problems: it does not say compared with what (with baseline or with another arm?), it gives no figure for the size of the decrease, "$p < 0.05$" hides a much smaller value, and the test is not named.

Example 1, correct: "At 6 months, the mean decrease in fasting plasma glucose was 3.0 mg/dL (SD 8.8) in arm A, 11.1 mg/dL (SD 11.2) in arm B and 16.3 mg/dL (SD 10.6) in arm C (ANOVA, $F(2;\ 157) = 22.93$, $p < 0.001$; $\eta^2 = 0.23$). Compared with arm A, the decrease was greater by 8.1 mg/dL in arm B (95% CI 3.4-12.8) and by 13.3 mg/dL in arm C (95% CI 8.6-18.0) (Tukey's test)."

Example 2, weak: "There is no difference in SBP between the sexes ($p = 0.81$), so sex does not influence blood pressure." The problems: a nonsignificant result does not prove the absence of a difference, and the second part of the sentence makes a causal claim that a cross-sectional study cannot support.

Example 2, correct: "Mean SBP was 131.4 mmHg (SD 14.5) in women and 130.9 mmHg (SD 13.3) in men; the difference of -0.5 mmHg (95% CI -4.9 to 3.8; Welch's t-test, $p = 0.81$) was not statistically significant. The data are incompatible with sex differences larger than about 5 mmHg."

Example 3, weak: "Obesity increases the risk of diabetes 4.32-fold ($p < 0.001$)." The problems: the OR is not a relative risk, "increases" suggests causation, and the CI is missing.

Example 3, correct: "The prevalence of OGTT-confirmed diabetes was 33.3% (18/54) in obese patients and 10.4% (11/106) in non-obese patients (prevalence ratio 3.21; 95% CI 1.64-6.31; $\chi^2 = 12.70$, $p < 0.001$). The association persisted after adjustment for age and sex (OR 6.01; 95% CI 2.35-15.39)."

Example 4, weak: "The correlation between BMI and SBP was highly significant ($p < 0.001$), so BMI determines SBP." The problems: "highly significant" speaks about p, not about the strength of the correlation; correlation does not show causation.

Example 4, correct: "BMI correlated positively and moderately with SBP($r = 0.41$, $p < 0.001$; $R^2 = 0.16$). Each 1 kg/m² increase in BMI was associated with an SBP higher by 1.4 mmHg (95% CI 0.9-1.9)."

Example 5, weak: "FINDRISC has an accuracy of 75%, so it is a good test." The problems: overall accuracy depends on prevalence and hides the trade-off between Se and Sp; "good" is meaningless without the purpose of the test.

Example 5, correct: "The area under the ROC curve of the FINDRISC score for OGTT-confirmed diabetes was 0.852 (95% CI 0.762-0.943). At the cutoff ≥ 12, sensitivity was 75.9%

and specificity 74.8%; at the sample prevalence (18.1%), the PPV was 40.0% and the NPV 93.3%."

A few rules of form complete the examples. Report exact p values (p = 0.034), except for very small ones (p < 0.001); do not write "p = 0.000", which some programs display. Use a reasonable number of decimals: a percentage from a sample of 50 is not written with three decimals. The name of the test appears once, at first use. The unit of measurement appears every time a value appears. The mean is accompanied by the SD when you describe the data and by the CI when you estimate a parameter.

> **Case study: an abstract with many odds ratios and no confidence intervals**
>
> Ursoiu and colleagues analyzed the association between the clinical severity of HIV infection and psychiatric disorders in 101 patients on antiretroviral treatment in western Romania [84]. The abstract reports seven odds ratios, all significant: 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) and 4.66 (p = 0.0013). As a writing exercise, note what is missing and how it could be completed. The confidence intervals are missing: with 101 patients, an OR of 8.72 almost certainly has a very wide interval, and the reader does not know whether the lower limit is 1.2 or 5. It is not stated whether the ORs are crude or adjusted. Many associations are tested on the same sample, so the risk that one of them is a false positive is not negligible; several p values (0.024; 0.038) would not survive a Bonferroni correction for seven tests (threshold 0.05/7 = 0.007). And the cross-sectional design does not allow us to say whether advanced disease leads to psychiatric disorders or the other way around. A rewritten results paragraph would give, for each association, the frequencies in the two groups, the OR with its 95% CI, the p value and, at the end, a statement about multiple comparisons. This is not a criticism of the authors: the space limits of an abstract force compromises, and the full text may contain the missing information. For the reader, however, the rule remains: an OR without a confidence interval is incomplete information.

14.5 Forty Statements: True or False

The statements below cover the whole course. Try to decide on your own before reading the justification; in the multiple-choice exam, many wrong options are built on exactly these confusions.

1. The p value is the probability that H0 is true. False. The p value is the probability of obtaining a result at least as extreme as the one observed, assuming that H0 is true; it is not a probability of the hypothesis [70].
2. A 95% CI for a difference in means that does not include 0 corresponds to a two-sided test that is significant at α = 0.05. True. The interval and the test use the same standard error and the same critical value.
3. The median is less influenced by extreme values than the mean. True. In mg2026, a single SBP value of 1540 mmHg (instead of 154) shifts the mean by 8.66 mmHg, while the median stays at 131.
4. The standard deviation decreases as the sample size increases. False. The standard deviation estimates the variability in the population and stays, on average, the same; it is the standard error that decreases.
5. The standard error of the mean decreases in proportion to $\sqrt{n}$. True. SE $= \frac{s}{\sqrt{n}}$; to halve it, n must increase 4 times.
6. If we double the sample, the width of the confidence interval is halved. False. The width decreases by a factor of √2, that is, by approximately 29%; halving it requires a sample 4 times larger.

7. The paired t-test is appropriate for comparing SBP between women and men. False. Women and men are independent groups; the independent-samples t-test is used (or Welch, Mann-Whitney).
8. A significant ANOVA shows which groups differ from one another. False. ANOVA says only that at least one mean differs; the pairs are identified by post hoc tests (Tukey, Bonferroni).
9. The chi-square test requires sufficiently large expected frequencies, as a rule at least 5 in each cell. True. Otherwise Fisher's exact test is used (for 2 x 2 tables).
10. A correlation coefficient r = 0.9 shows that X causes Y. False. Correlation measures linear association; a causal direction or a third factor cannot be ruled out.
11. $R^2 = 0.164$ for the regression of SBP on BMI means that 16.4% of the variability in SBP is linearly explained by BMI. True. The remaining 83.6% is due to other factors and to individual variability.
12. The Spearman coefficient is appropriate for ordinal variables. True. It works on ranks and does not require normality.
13. In a case-control study we can compute the relative risk directly. False. The proportion of diseased subjects is fixed by the investigator (the number of cases and controls), so risks cannot be estimated; the OR is computed.
14. When the disease is rare, the OR approximates the RR well. True. When a and c are small relative to b and d, the odds are close to the risks.
15. The NNT is the inverse of the relative risk reduction. False. NNT = 1 / absolute risk reduction.
16. An RR = 1.02 with 95% CI 0.80-1.30 proves that the exposure has no effect. False. The data are compatible with a 20% reduction or with a 30% increase; we have no evidence of an effect, but neither do we have evidence of its absence [79].
17. With constant sensitivity and specificity, the PPV increases as prevalence increases. True. For FINDRISC ≥ 12, the PPV rises from 5.8% at a prevalence of 2% to 75.1% at a prevalence of 50%.
18. The sensitivity of a test changes directly with disease prevalence. False. Se and Sp are properties of the test with respect to diseased and healthy subjects, respectively; they can change with the disease spectrum, not with prevalence as such.
19. A negative result on a test with very high sensitivity helps rule out the disease. True. The test misses few diseased subjects, so a negative result has a high NPV.
20. $\mathrm{LR}^+ = \frac{\mathrm{Se}}{1-\mathrm{Sp}}$. True. For FINDRISC ≥ 12: 0.759 / 0.252 = 3.01.
21. The classification error rate is (FP + FN) / N. True. For FINDRISC ≥ 12: (33 + 7)/160 = 25%, the complement of the 75% accuracy.
22. An area under the ROC curve of 0.5 corresponds to a test with no discriminative value. True. The test does no better than tossing a coin.
23. Randomization guarantees that the arms are identical at enrollment. False. It balances them on average; chance differences remain (in mg2026, 9.4% former smokers in A vs. 31.5% in C).
24. Blinding and allocation concealment are the same thing. False. Allocation concealment protects enrollment; blinding protects treatment and assessment.
25. Intention-to-treat analysis includes all randomized patients, in the arm to which they were allocated. True. This is how the comparability provided by randomization is preserved.
26. A cross-sectional study can establish which came first, the exposure or the disease. False. Exposure and disease are measured at the same time.
27. The power of a test increases with sample size. True. The standard error decreases, so real effects are detected more easily.
28. A type I error means rejecting a true H0. True. Its probability is α.
29. If p = 0.06, H0 is true. False. We say only that we do not reject H0 at the 0.05 level; H0 may be false and the study underpowered.

30. A nonsignificant Shapiro-Wilk test proves that the data are normal. False. We do not reject normality; in small samples the test has low power and can miss important departures.
31. Log transformation can make a positively skewed distribution symmetric. True. The triglycerides in mg2026 go from a skewness of 0.95 to -0.28.
32. Mean ± 1.96 SD is the confidence interval for the mean. False. It is the interval containing approximately 95% of individual values; the CI for the mean uses the standard error.
33. The Kruskal-Wallis test is the nonparametric alternative to one-way ANOVA. True. It compares three or more independent groups on the basis of ranks.
34. The Friedman test is used for independent groups. False. It is for three or more measurements on the same subjects (dependent data).
35. In a very large sample, a clinically negligible difference can turn out statistically significant. True. The correlation r = -0.088 between age and duration of surgery, in 867 patients, has p = 0.01 [55].
36. A Cronbach's α of 0.93 proves the validity of a questionnaire. False. Alpha measures internal consistency (reliability), not validity.
37. The slope of a regression line has the same unit of measurement as the dependent variable. False. Its unit is the unit of Y per unit of X (for example, mmHg per kg/m²).
38. The coefficient r does not change if we express height in meters instead of centimeters. True. r has no unit of measurement and is invariant to linear transformations of the variables.
39. In Table 1 of an RCT, differences between arms must always be tested with p values. False. CONSORT recommends not testing them; the differences are, by construction, due to chance [56].
40. The significant chi-square test for a decrease ≥ 5% across the three arms of mg2026 (p = 0.034) shows that all arms differ from one another. False. It shows that at least one differs; B and C have almost identical proportions (RR C vs. B 0.98, 95% CI 0.55-1.75).

14.6 Summary Table of Formulas

Table 14.2 brings together the working formulas of the course, with a reference to the chapter in which each was explained. Notation: n sample size, m mean, s standard deviation, p proportion, a, b, c, d the cells of the 2 x 2 table (a exposed and diseased or test-positive and diseased, b exposed and healthy or test-positive and healthy, c unexposed and diseased or test-negative and diseased, d unexposed and healthy or test-negative and healthy).

Table 14.2. The main formulas of the course.

Indicator	Formula	Chapter
Mean	$m = \frac{\sum x}{n}$	2
Standard deviation	$s = \sqrt{\left[\frac{\sum (x-m)^2}{n-1}\right]}$	2
Coefficient of variation	CV = 100 x s / m (%)	2
z-score	$z = \frac{x-m}{s}$	3
Interval for individual values (95%)	$m \pm 1.96 \times s$	3
Standard error of the mean	$\mathrm{SE} = \frac{s}{\sqrt{n}}$	4
95% CI for the mean	$m \pm \frac{t(n-1;\ 0.975) \times s}{\sqrt{n}}$	4
95% CI for a proportion (Wald)	$p_i \pm 1.96 \times \sqrt{\left[\frac{p_i(1-p_i)}{n}\right]}$	4

Indicator	Formula	Chapter
Sample size for a mean with margin E	$n = \left(\frac{1.96 \times \sigma}{E}\right)^2$	4
Sample size for a proportion with margin E	$n = \frac{1.96^2 \times p_i(1-p_i)}{E^2}$	4
Sample size per group, two means	$n = \frac{2 \times \left(z\left(1-\frac{\alpha}{2}\right) + z(1-\beta)\right)^2 \times \sigma^2}{\Delta^2}$	4, 13
Independent-samples t-test	$t = \frac{m_1 - m_2}{\sqrt{s^2p \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	6
Paired t-test	$t = \frac{\text{md}}{\frac{\text{sd}}{\sqrt{n}}}$, df = n - 1	6
Cohen's d	$d = \frac{m_1 - m_2}{\text{sp}}$	6
ANOVA	$F = \frac{\frac{\text{SSB}}{k-1}}{\frac{\text{SSW}}{N-k}}$	6
Eta squared	$\eta^2 = \frac{\text{SSB}}{\text{SSB} + \text{SSW}}$	6
Expected frequency (chi-square)	E = row total x column total / N	7
Chi-square	$\chi^2 = \frac{\sum(O-E)^2}{E}$, df = (r - 1)(c - 1)	7
Test for r	$t = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}}$	9
Coefficient of determination	$R^2 = r^2$ (simple regression)	9
Regression line	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
Relative risk (cohort, RCT)	$\text{RR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$	10
Odds ratio	$\text{OR} = \frac{a \times d}{b \times c}$	10
95% CI for the RR	$\exp\left[\ln \text{RR} \pm 1.96 \times \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}\right]$	10
95% CI for the OR	$\exp\left[\ln \text{OR} \pm 1.96 \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right]$	10
Risk difference	$\text{RD} = \frac{a}{a+b} - \frac{c}{c+d}$	10
NNT (NNH)	$\text{NNT} = \frac{1}{\text{RD}}$, rounded up	10, 13
Attributable fraction among the exposed	$\text{AF} = \frac{\text{RR} - 1}{\text{RR}}$	10
Odds and probability	odds = P / (1 - P); P = odds / (1 + odds)	11
Bayes' theorem (PPV)	$\text{PPV} = \frac{\text{Se} \times P}{\text{Se} \times P + (1 - \text{Sp}) \times (1 - P)}$	11
Post-test odds	post-test odds = pre-test odds x LR	11
Sensitivity	$\text{Se} = \frac{a}{a+c}$	12
Specificity	$\text{Sp} = \frac{d}{b+d}$	12

Indicator	Formula	Chapter
Positive predictive value	$\text{PPV} = \frac{a}{a+b}$	12
Negative predictive value	$\text{NPV} = \frac{d}{c+d}$	12
Accuracy	$\frac{a+d}{N}$	12
Error rate	$\frac{b+c}{N}$	12
Likelihood ratios	$\text{LR}^+ = \frac{\text{Se}}{1-\text{Sp}}$; $\text{LR}^- = \frac{1-\text{Se}}{\text{Sp}}$	12
Youden index	$J = \text{Se} + \text{Sp} - 1$	12
Cronbach's alpha	$\alpha = \frac{k}{k-1} \times \left[1 - \frac{\sum s^2(\text{item})}{s^2(\text{ total })}\right]$	13
Design effect (clusters)	1 + (m - 1) x ICC	13

The table is not meant to be memorized as a list. Each formula has a logic: the standard error is the variability divided by the square root of n; a test statistic is an effect divided by its standard error; a confidence interval is the estimate plus or minus a critical value multiplied by the standard error; the indicators of diagnostic tests are proportions computed on the rows or on the columns of the same 2 x 2 table. Anyone who understands these four ideas can reconstruct almost the entire table.

14.7 Worked Problems

Worked Problem 14.1. Frequency Table with Missing Values

Statement. In a questionnaire administered to 200 patients in a family practice, the question about educational level received the following answers: lower secondary school 30, high school or vocational school 74, university 62, postgraduate 22, no answer 12. Construct the frequency table, with percentages of the total and valid percentages, and state the median category.

Solution. The number of valid answers is 200 - 12 = 188. The percentage of the total is computed out of 200, the valid percentage out of 188, and the cumulative percentage from the valid percentages.

Table 14.3. Distribution of educational level (n = 200, 12 missing values).

Education	Frequency	% of total	Valid %	Cumulative valid %
Lower secondary school	30	15,0	16,0	16,0
High school or vocational school	74	37,0	39,4	55,3
University	62	31,0	33,0	88,3
Postgraduate	22	11,0	11,7	100,0
Missing	12	6,0	-	-
Total	200	100,0	100,0	

The median category is the one in which the cumulative valid percentage first exceeds 50%: high school or vocational school (55.3%). The variable is ordinal, so a median makes sense, not a mean. Valid percentages describe the respondents; if nonresponse is not random (for example, people with less education avoid the question), the valid percentages may be distorted, and the number of missing values must always be reported.

Worked Problem 14.2. A Case-Control Study: Oral Contraceptives and Rheumatoid Arthritis

Statement. A case-control study includes 120 women with rheumatoid arthritis and 240 women without the disease, of comparable age. Past use of oral contraceptives was reported by 42 of the cases and 54 of the controls. Compute the appropriate measure of association, with its 95% CI, and test the association.

Solution. (a) Research hypothesis: oral contraceptive use is associated with rheumatoid arthritis. (b) H0: OR = 1 (no association); H1: OR ≠ 1. (c) Case-control design, so the measure is the OR; the test is the chi-square test for a 2 x 2 table. The table: a = 42 (exposed cases), b = 54 (exposed controls), c = 78 (unexposed cases), d = 186 (unexposed controls).

$$\text{OR} = \frac{42\times186}{54\times78} = \frac{7812}{4212} = 1.85;\ \text{SE}(\ln \text{OR}) = \sqrt{\frac{1}{42}+\frac{1}{54}+\frac{1}{78}+\frac{1}{186}} =$$

$$\sqrt{0.0238+0.0185+0.0128+0.0054} = \sqrt{0.0605} = 0.246 \ln 1.85 = 0.618; \quad 0.618 \pm 1.96 \times 0.246 = 0.618 \pm 0.482, \text{ that is, } 0.136 \text{ to } 1.100;\ \text{CI}\,95\% = e^{0.136} \text{ to } e^{1.100} = 1.15 \text{ to } 3.00$$

Chi-square (without correction) = 6.39, df = 1, p = 0.011. (e) Result: OR 1.85 (95% CI 1.15-3.00). (f) Statistical conclusion: we reject H0; the interval does not include 1. (g) Research conclusion: the odds of oral contraceptive use are approximately 1.9 times higher in women with rheumatoid arthritis than in controls. A case-control study does not allow the relative risk to be computed, and the association may be influenced by recall bias (cases remember past exposures differently) and by confounders.

Worked Problem 14.3. A Cohort: Screen Time and Visual Acuity

Statement. A cohort of 400 employees is followed for 3 years. Of the 150 who work at least 6 hours a day at a computer screen, 45 develop a decrease in visual acuity; of the 250 who work less, 30 do. Compute the RR, RD, NNH and the attributable fraction among the exposed.

Solution. Risk among the exposed: 45/150 = 0.30; among the unexposed: 30/250 = 0.12.

$$\text{RR} = \frac{0.30}{0.12} = 2.5;\ \text{SE}(\ln \text{RR}) = \sqrt{\frac{1}{45}-\frac{1}{150}+\frac{1}{30}-\frac{1}{250}} =$$

$$\sqrt{0.0222-0.0067+0.0333-0.0040} = \sqrt{0.0449} = 0.212;\ \text{CI}\,95\%\ \text{RR} = \exp(\ln 2.5 \pm 1.96 \times 0.212) = \exp(0.916 \pm 0.415) = 1.65 \text{ to } 3.79\ \text{RD} = 0.30 - 0.12 = 0.18;\ \text{NNH} = \frac{1}{0.18} = 5.6, \text{ that is, } 6;\ \text{AF} = \frac{2.5-1}{2.5} = 0.60$$

Chi-square = 19.94, p < 0.001. OR-ul ar fi $\frac{(45\times220)}{105\times30} = 3.14$, larger than the RR, as happens when the event is frequent. Interpretation: the risk of decreased acuity is 2.5 times higher in the exposed; for every 6 exposed employees there is one additional case compared with the unexposed; 60% of the cases occurring among the exposed could be attributed to the exposure, if the relationship is causal. Since this is an observational cohort, the exposed employees may also differ in age or other characteristics, so an adjusted analysis is necessary before reaching a causal conclusion.

Worked Problem 14.4. Reconstructing the Table of a Diagnostic Test

Statement. A screening test is applied to 500 people, 20% of whom have the disease. The sensitivity of the test is 85%, and the PPV is 68%. Reconstruct the 2 x 2 table and compute Sp, NPV, accuracy, error rate and the likelihood ratios.

Solution. Diseased: 0.20 x 500 = 100; healthy: 400. True positives: 0.85 x 100 = 85; false negatives: 15. From $\text{PPV} = \frac{\text{TP}}{\text{TP+FP}} = 0.68$ it follows that $\text{TP} + \text{FP} = \frac{85}{0.68} = 125$, hence FP = 40. True negatives: 400 - 40 = 360.

Table 14.4. The reconstructed table of the screening test.

	Diseased	**Healthy**	**Total**
Test positive	85	40	125
Test negative	15	360	375
Total	100	400	500

$\text{Sp} = \frac{360}{400} = 90\%$; $\text{NPV} = \frac{360}{375} = 96\%$; accuracy = (85 + 360)/500 = 89%; error rate = (40 + 15)/500 = 11%; $\text{LR}^+ = \frac{0.85}{0.10} = 8.5$; $\text{LR}^- = \frac{0.15}{0.90} = 0.17$. Check using Bayes: pre-test odds 0.20/0.80 = 0.25; post-test odds for a positive result 0.25 x 8.5 = 2.125; probability 2.125/3.125 = 68%, exactly the PPV given in the statement.

Worked Problem 14.5. Choosing the Test on the mg2026 Dataset

Statement. For each question, choose the test, justify your choice and give the result from the mg2026 dataset: a) does baseline blood glucose differ between men and women? b) is educational level associated with satisfaction? c) does baseline SBP differ among current smokers, former smokers and never smokers? d) did weight change in arm A? e) is smoking associated with hypertension (SBP ≥ 140) at enrollment?

Solution. a) Numerical variable, two independent groups; blood glucose shows a slight departure from normality (SW p = 0.037), but with 74 and 86 patients Welch's t-test is robust: 106.2 vs. 104.8 mg/dL, t = 0.54, p = 0.589; we do not reject H0. b) Two ordinal variables: Spearman correlation, rho = 0.125, p = 0.115; there is no significant association. c) Numerical variable, three independent groups; normality is acceptable in each group (SW p 0.37; 0.86; 0.23), Levene p = 0.114: ANOVA, F = 7.85, p < 0.001, $\eta^2 = 0.09$; Tukey: current smokers have an SBP 9.5 mmHg higher than former smokers (95% CI 2.5-16.5) and 8.5 mmHg higher than never smokers (95% CI 2.9-14.2). d) Paired data (the same patients at 0 and 6 months), a single group; Wilcoxon test: median of the differences 1.1 kg, p = 0.052, a borderline result that does not allow H0 to be rejected at α = 0.05. e) Two dichotomous variables, cross-sectional design: chi-square; 21/53 (39.6%) of current smokers vs. 20/107 (18.7%) of never and former smokers have hypertension; $\chi^2 = 8.15$, p = 0.004; prevalence ratio 2.12 (95% CI 1.26-3.55), OR 2.85 (95% CI 1.37-5.95).

14.8 Proposed Exercises

1. In a study of 250 adults, smoking status was: never smokers 120, current smokers 70, former smokers 40, no answer 20. Construct the frequency table with percentages of the total and valid percentages.
2. A case-control study includes 80 patients with COPD and 160 controls. Of the cases, 56 are smokers, and of the controls, 64. Compute the OR with its 95% CI and the chi-square statistic.
3. In a cohort, 30 of 200 people exposed to an occupational factor become ill, and 15 of 300 unexposed people. Compute the RR with its 95% CI, the RD and the NNH.

4. A test with Se 90% and Sp 95% is used for screening in 10,000 people, with a disease prevalence of 2%. Reconstruct the 2 x 2 table and compute the PPV and NPV. Comment.
5. In 36 patients, the mean cholesterol is 220 mg/dL and the SD is 18 mg/dL. Compute the standard error and the 95% CI for the mean ($t(35, 0.975) = 2.03$).
6. In 27 patients, the correlation between two numerical variables is $r = 0.6$. Compute R^2 and the t statistic, and state whether r is significant at $\alpha = 0.05$ (critical value $t(25) \approx 2.06$).
7. Using the regression line from mg2026 (SBP = 89.47 + 1.435 x BMI), compute the predicted SBP for a patient with a BMI of 27 kg/m². Why would a prediction at a BMI of 15 not be valid?
8. The pre-test probability of a disease is 30%, and the LR+ of a test is 6. Compute the post-test probability after a positive result.
9. Choose the test for: a) comparing HbA1c across the three arms of an RCT, with normal distributions; b) comparing a pain score (1-10) before and after an intervention in the same patients; c) the association between sex and the presence of diabetes; d) the relationship between age and SBP; e) comparing a skewed variable between two independent groups; f) the course of a score at 0, 3 and 6 months in the same patients, skewed distribution.
10. Write the results paragraph for the following analysis from mg2026: HDL in women 53.92 mg/dL(SD 11.70; $n = 86$), in men 44.41 mg/dL(SD 9.66; $n = 74$); Levene p = 0.034; Welch's t-test t = 5.63, df 157.7, $p < 0.001$; difference 9.5 mg/dL (95% CI 6.2-12.9).
11. Rewrite the following statement correctly: "The chi-square test showed that obesity causes diabetes ($p < 0.001$), the risk being 4.32 times higher."
12. An article reports: "length of stay 6.2 ± 5.9 days in group A and 4.1 ± 4.4 days in group B (t-test, p = 0.04)". What do you notice, and what would you have recommended?

Common Pitfalls

- Skipping the methods and reading only the article's conclusion.
- Interpreting "not significant" as "there is no effect", without looking at the confidence interval.
- Confusing statistical significance with clinical importance, especially in large samples.
- Reading an OR as a relative risk when the event is frequent.
- Transferring the PPV and NPV from a high-prevalence study to a low-prevalence screening population.
- Believing a specificity of 100% obtained in a few dozen healthy volunteers.
- Failing to check whether published confidence intervals are plausible for the reported n.
- Drawing causal conclusions from cross-sectional studies or from correlations.
- Describing visibly skewed variables by mean ± SD (an SD close to the mean for a positive variable).
- Changing the test after seeing the result, or reporting only the significant comparisons.
- Forgetting that an HR presupposes follow-up over time and cannot come from a purely cross-sectional design.

Key Points

- The map of the discipline: question, design, data, description, estimation, testing, association, decision, reporting.
- The choice of test depends on the type of outcome variable, on the number of groups and on whether they are independent.

- A complete result has three parts: the effect size, the confidence interval and the p value.
- The confidence interval shows precision; a wide interval means a study that has not settled much.
- Predictive values depend on prevalence; Se, Sp and LRs are more stable across populations.
- Confounding is countered by randomization (at the design stage) or by adjustment (at the analysis stage).
- Many published figures can be checked with the simple formulas of the course; do this exercise every time something looks too good to be true.

Self-Assessment Questions

The review quiz below has 30 questions in the style of the Romanian residency examination: questions 1-15 have a single correct answer (SC), and questions 16-30 have between two and four correct answers (MC). The answers are in the chapter's answer file.

1. (SC) To describe BMI in the mg2026 dataset (skewness 1.42, Shapiro-Wilk $p < 0.001$), the most appropriate indicators are:
    a) the mean and the standard deviation
    b) the median and the interquartile range
    c) the mode
    d) the range
    e) the coefficient of variation
2. (SC) The 95% confidence interval for mean SBP in mg2026 is 128.96-133.30 mmHg. The correct statement is:
    a) 95% of patients have an SBP between 128.96 and 133.30 mmHg
    b) the sample mean lies in the interval with a probability of 95%
    c) the procedure used produces intervals that contain the population mean in 95% of possible samples
    d) SBP is significantly different from 130 mmHg
    e) the distribution of SBP is normal
3. (SC) To compare blood glucose at enrollment and at 6 months in the same 160 patients, with normally distributed differences, one uses:
    a) the independent-samples t-test
    b) the paired-samples t-test
    c) one-way ANOVA
    d) the chi-square test
    e) the Mann-Whitney test
4. (SC) Satisfaction (a 1-5 Likert item) is compared across the three arms of the mg2026 study using:
    a) one-way ANOVA
    b) the Kruskal-Wallis test
    c) the paired t-test
    d) the Friedman test
    e) Pearson correlation
5. (SC) A study follows 500 smokers and 500 nonsmokers for 10 years, recording the occurrence of COPD. The design is:
    a) case-control
    b) cross-sectional

c) prospective cohort
d) randomized controlled trial
e) ecological study

6. (SC) In a case-control study, the appropriate measure of association between exposure and disease is:
   a) the relative risk
   b) the odds ratio
   c) the number needed to treat
   d) the prevalence
   e) the cumulative incidence
7. (SC) The correlation r = -0.832 between osteoprotegerin and ejection fraction indicates:
   a) a weak correlation
   b) a strong inverse correlation
   c) that osteoprotegerin lowers ejection fraction
   d) a coefficient of determination of 0.832
   e) a nonsignificant association
8. (SC) The absolute risk reduction for an event is 10 percentage points. The NNT is:
   a) 1
   b) 5
   c) 10
   d) 18
   e) 100
9. (SC) A test with Se 90% and Sp 80% is applied to 1000 people, with a disease prevalence of 15%. The PPV is approximately:
   a) 15%
   b) 44%
   c) 80%
   d) 90%
   e) 98%
10. (SC) The classification error rate of a diagnostic test, using the notation of the 2 x 2 table, is:

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. (SC) A three-group ANOVA gives $p < 0.001$. The correct next step is:

a) t-tests between all pairs, without correction
b) post hoc tests, for example Tukey
c) the chi-square test
d) concluding that all groups differ from one another

e) repeating the ANOVA on another sample before any interpretation

12. (SC) In a randomized superiority trial, the recommended primary analysis is:

a) per-protocol analysis

b) intention-to-treat analysis

c) analysis of only the patients who completed the study

d) analysis of only the adherent patients

e) subgroup analysis

13. (SC) For the secondary outcome SBP in mg2026, the ANOVA specified in the plan gives p = 0.056. The correct wording is:

a) H0 is accepted; the intervention does not influence SBP

b) H0 is not rejected; the data do not provide sufficient evidence of a difference between arms

c) there is a significant effect of the intervention

d) Kruskal-Wallis is used instead, which gives $p < 0.05$, and that result is reported

e) the result proves that the arms are equivalent

14. (SC) In a cohort, 18 of 150 exposed and 10 of 250 unexposed people develop the disease. The relative risk is:

a) 1,8

b) 3,0

c) 3,27

d) 0,33

e) 8,0

15. (SC) At the FINDRISC cutoff ≥ 12, the PPV is 40.0% at a prevalence of 18.1%. In a population with a prevalence of 5%, with the same Se and Sp, the PPV will be:

a) greater than 40%

b) less than 40%

c) equal to 40%

d) 95%

e) impossible to estimate

16. (MC) Regarding the mean and the median, the following are true:

a) the median is robust to extreme values

b) in a positively skewed distribution, the mean is usually larger than the median

c) in a normal distribution, the mean and the median coincide

d) the median is computed only for nominal variables

e) the mean is the recommended indicator for a single Likert item

17. (MC) Regarding the p value, the following are true:

a) it is the probability of obtaining a result at least as extreme if H0 is true

b) it is the probability that H0 is true

c) it depends on the sample size

d) a very small value guarantees a clinically important effect

e) it is compared with the significance level α

18. (MC) Regarding statistical errors, the following are true:

a) a type I error means rejecting a true H0

b) α is the probability of a type I error
c) the power of the test is 1 - β
d) increasing the sample size decreases power
e) a type II error means failing to reject a false H0

19. (MC) The conditions for applying the independent-samples t-test are:
a) normality in each group, or sufficiently large samples
b) independence of observations
c) equal variances, otherwise the Welch version is used
d) nominal variables
e) the same patients measured twice

20. (MC) Regarding the chi-square test, the following are true:
a) it applies to categorical variables
b) it requires sufficiently large expected frequencies
c) with small expected frequencies in 2 x 2 tables, Fisher's test is used
d) it compares means
e) expected frequency = row total x column total / N

21. (MC) For the regression line SBP = 89.47 + 1.435 x BMI from mg2026, the following are true:
a) SBP increases on average by 1.435 mmHg for each kg/m² of BMI
b) at a BMI of 30, the predicted SBP is approximately 132.5 mmHg
c) the intercept of 89.47 mmHg has a direct clinical interpretation
d) with $R^2 = 0.164$, BMI linearly explains 16.4% of the variability in SBP
e) the equation proves that elevated BMI causes hypertension

22. (MC) Regarding the RR and the OR, the following are true:
a) the RR is computed in cohort studies and in RCTs
b) the OR is the appropriate measure in case-control studies
c) when RR > 1 and the event is frequent, the OR is larger than the RR
d) RR = 1 indicates no association
e) a 95% CI that includes 1 indicates a significant association

23. (MC) Regarding the indicators of a diagnostic test, the following are true:
a) $\mathrm{Se} = \frac{a}{a+c}$
b) $\mathrm{Sp} = \frac{d}{b+d}$
c) the PPV depends on prevalence
d) sensitivity increases directly with prevalence
e) $\mathrm{LR}^- = \frac{1-\mathrm{Se}}{\mathrm{Sp}}$

24. (MC) Regarding the ROC curve, the following are true:
a) it plots sensitivity against 1 - specificity
b) an area of 0.5 indicates no discrimination
c) the Youden index is Se + Sp - 1
d) the cutoff that maximizes the Youden index maximizes the PPV
e) the area of 0.852 for the FINDRISC score indicates good discrimination

25. (MC) Regarding Bayesian reasoning, the following are true:

a) post-test odds = pre-test odds x LR

b) probability = odds / (1 + odds)

c) a large LR+ increases the post-test probability after a positive result

d) the pre-test probability does not influence the post-test probability

e) an LR- less than 1 decreases the probability of disease after a negative result

26. (MC) For weight loss ≥ 5% in mg2026 ($\chi^2 = 6.77$, df 2, p = 0.034), the following are true:

a) H0 is rejected at $\alpha = 0.05$

b) at least one arm differs from the others

c) all three arms differ from one another

d) the RR for C vs. B of 0.98 (95% CI 0.55-1.75) does not show a difference between C and B

e) the expected frequencies (minimum 12.6) allow the test to be applied

27. (MC) Regarding study designs, the following are true:

a) randomization also balances unknown confounders, on average

b) a cohort study allows incidence to be estimated

c) a cross-sectional study estimates prevalence

d) a case-control study directly estimates the incidence of the disease

e) multivariable adjustment can partially correct for confounding in observational studies

28. (MC) Regarding the normal distribution, the following are true:

a) approximately 68% of values lie in the interval $m \pm s$

b) approximately 95% of values lie in the interval $m \pm 1.96s$

c) it is symmetric about the mean

d) a Shapiro-Wilk test with $p < 0.05$ leads to rejection of the normality hypothesis

e) all biological variables are normally distributed

29. (MC) Regarding correlation, the following are true:

a) r takes values between -1 and 1

b) the Spearman coefficient is appropriate for ordinal or skewed variables

c) r = 0 shows that there is no relationship of any kind between the variables

d) a single extreme value can change the value of r considerably

e) in a very large sample, a weak correlation can be statistically significant

30. (MC) Regarding the reporting of studies, the following are true:

a) CONSORT is used for randomized trials

b) STROBE is used for observational studies

c) STARD is used for diagnostic accuracy studies

d) Table 1 of an RCT must contain p values for each characteristic

e) the main effect is reported with its size and its 95% confidence interval

Part II. Practical Sessions

The practical sessions follow the lecture topics week by week and use the tools listed in the course syllabus: Google Forms and Google Sheets (the university Google account), JASP version 0.98.1 or later [47], MedCalc 24.0 or its online calculators [50], and Epi Info 7.2.7 [51], which also runs offline. The CDC announced the end of development and technical support for Epi Info in September 2025; the program remains usable but no longer receives new versions, which is why each session also gives the JASP or MedCalc alternative. The working files are downloaded from the discipline's Moodle platform (Emediqual, moodle.umft.ro): the teaching dataset "mg2026" in two versions, mg2026_brut.csv (raw) and mg2026_curat.csv (clean), plus the discipline's classic files used for the additional exercises. The data dictionary of the mg2026 dataset is in Appendix C. Completed work is uploaded to Moodle, named LPx_Surname_FirstName.

With the Romania regional setting, Google Sheets uses the comma as the decimal separator and the semicolon as the function argument separator. In this English edition, the formulas in the practical sessions are written for the English (United States) setting, with a decimal point and the comma as the argument separator, for example =AVERAGE(M2:M161); with the Romania setting, replace the "," between arguments with ";". JASP and MedCalc display results with a decimal point, which is also the form used in reports.

Each session gives the expected results on the mg2026 dataset. They are not provided to be copied, but so that students can check their own steps: if the mean systolic blood pressure does not come out as 131.13 mmHg, something went wrong during import or cleaning, and it is worth finding before moving on. For assignments and for the practical exam, the a-g reporting format is used: the research hypothesis, the statistical hypotheses, the method chosen with its justification, the procedure, the result with the appropriate graph, the statistical conclusion, and the research conclusion in words.

Practical Session 1. Introduction to Biostatistics and Digital Tools: Google Sheets, JASP and MedCalc

Objective

The first practical session has a very concrete aim: by the end of it, each student should have a clean, correctly organized data file, accompanied by a data dictionary, and should know how to open this file in the programs they will use all semester. The statistical calculations themselves are simple for now. What matters at this point is working discipline: one row for each patient, one column for each variable, consistent values, no hidden errors. Experience from recent years shows that most wrong results in the practical exam come not from a wrongly chosen test but from a badly imported or uncleaned file.

By the end, the student should be able to:

- build a demonstration medical dataset in Google Sheets, with validation rules and a data dictionary;
- import a CSV file correctly into Google Sheets, taking the regional setting into account (decimal comma and the ";" separator);
- detect and correct data entry errors in a raw file, documenting each change;
- open the dataset in JASP and in MedCalc, set the variable types and obtain the basic descriptive statistics;
- recognize the role of the Epi Info program and upload the results to the Moodle platform.

Tools

Software: Google Sheets (the university Google account), JASP 0.98.1 [47], MedCalc 24.0 or its online calculators [50], Epi Info 7.2.7 [51]. Files from Moodle: mg2026_curat.csv (160 patients) and mg2026_brut.csv (161 rows, with errors left in on purpose). The data dictionary of the mg2026 dataset is in Appendix C and, in brief, in Table 1.1 of Chapter 1. Approximate time: 100 minutes.

In the mg2026 files, the columns are in this order: 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. The data occupy rows 2-161 in the clean file and 2-162 in the raw file; row 1 contains the variable names.

Procedure

L1.1 Creating a Demonstration Medical Dataset in Google Sheets

The first exercise does not use the mg2026 dataset but a small dataset that you build yourselves, so that you understand from the start how a data table is designed. Imagine that you are a resident and that you want to record a few basic data for 12 fictitious patients seen during one on-call day in a family practice.

1. Open Google Sheets from your university account and create a new spreadsheet named LP1_Surname_FirstName. Check the regional setting: File > Settings > Locale should be set to "United States", the setting assumed by the formulas in this edition (with "Romania", replace the "," between arguments with ";"). Menu names may vary slightly depending on the interface language.

2. Rename the first sheet "date" (data) and write the variable names in row 1, without spaces and without diacritics: id, sex, varsta, inaltime, greutate, imc, tas, glicemie, fumat, educatie.
3. Create a second sheet, named "dictionar" (dictionary), with the columns: name, description, type, unit or codes, plausible range. Fill in one row for each variable, following the model in Table 2.2 of Chapter 2. For example: tas; systolic blood pressure, measured in the sitting position after 5 minutes of rest; integer numeric; mmHg; 70-260. For educatie: ordinal; 1 lower secondary school, 2 high school or vocational school, 3 university, 4 postgraduate; 1-4.
4. Add validation rules that prevent absurd values from being entered. Select column B (sex), from B2 downward, then Data > Data validation > Add rule. Choose the "Drop-down" criterion with the items F and M and select the option to reject invalid values ("Reject the input"). For column G (tas) choose the criterion "Is between" 70 and 260; for H (glicemie), between 40 and 400; for D (inaltime), between 130 and 210; for E (greutate), between 30 and 250; for I (fumat), a drop-down list with nefumator, fost fumator, fumator; for J (educatie), a list with 1, 2, 3, 4. The validation rules transfer the plausible ranges from the dictionary into the spreadsheet.
5. Compute BMI with a formula, not by hand. In cell F2 write =ROUND(E2/(D2/100)^2, 1) and copy the formula down by dragging the corner of the cell. Note the separator between arguments: "," in the English (United States) setting, ";" in the Romania setting.
6. Enter the data for 12 fictitious patients, with codes P01-P12. Deliberately try to enter a blood pressure of 1540 or the sex "f": the validation rule should reject the value. If it does not, check how the rule is set up.
7. Below the data, compute a few simple statistics: the number of patients =COUNT(G2:G13), mean SBP =AVERAGE(G2:G13), the median =MEDIAN(G2:G13), the standard deviation =STDEV(G2:G13), the minimum and maximum =MIN(G2:G13) and =MAX(G2:G13), the number of women =COUNTIF(B2:B13, "F").

What you should obtain: a spreadsheet with two sheets, "date" and "dictionar", in which no value outside the plausible ranges can be entered and in which BMI is recalculated automatically when you change weight or height.

L1.2 Importing the File mg2026_curat.csv

A CSV (comma-separated values) file is a text file in which the values are separated by commas and decimals are written with a point (73.4). With the Romania regional setting, Google Sheets expects a decimal comma, and a value such as 5.1 may be misinterpreted, for example as a calendar date (January 5). That is why the import is done in a specific order.

1. Create a new spreadsheet, mg2026_curat. Before importing, temporarily set File > Settings > Locale to "United States".
2. File > Import > Upload and choose mg2026_curat.csv downloaded from Moodle. Under "Import location" choose "Replace spreadsheet" or "Insert new sheet(s)", for the separator type choose "Detect automatically" or "Comma", and leave the conversion of text to numbers checked.
3. Check the import: the sheet has 161 rows (1 header row and 160 patients) and 28 columns (A-AB). Numbers should be right-aligned; a left-aligned column contains text.
4. Switch the Locale back to "Romania" if that is your working setting (otherwise keep "United States", which the formulas in this edition assume). With "Romania", the numeric values remain numbers, but they are displayed with a decimal comma (73,4), and new

formulas use the ";" separator. Freeze the first row (View > Freeze > 1 row) so that the variable names remain visible when you scroll.

What you should obtain (self-check): =COUNTA(A2:A161) gives 160; =AVERAGE(M2:M161) gives 131.13 (mean SBP); =STDEV(M2:M161) gives 13.88; =MEDIAN(M2:M161) gives 131; =AVERAGE(D2:D161) gives 56.29 (mean age); =AVERAGE(K2:K161) gives 29.04 (BMI); =AVERAGE(S2:S161) gives 105.44 (blood glucose); =COUNTIF(C2:C161, "F") gives 86; =COUNTIF(G2:G161, "fumator") gives 53; =COUNTIF(E2:E161, "urban") gives 95. If mean SBP is not 131.13, do not move on until you find the cause: most likely, some of the values were imported as text or as calendar dates.

L1.3 Cleaning the File mg2026_brut.csv

The raw file simulates what you actually receive from a colleague who has entered the data from the case record forms: the same variables, but with mistakes. It contains 11 errors left in on purpose and one duplicate row. Your task is to find them, correct them and document each correction. Always work on a copy: keep a "brut_original" sheet untouched and make the corrections in a "brut_lucru" (working copy) sheet.

1. Import mg2026_brut.csv into a new spreadsheet, using the same procedure as in L1.2 (United States locale for the import, then Romania). Check: the data occupy rows 2-162, that is, 161 data rows, one more than there should be.
2. Add a "jurnal" (log) sheet, with the columns: no., patient id, variable, value found, how it was detected, corrected value, source of the correction. Every change made to the data must appear here. This is the principle of traceability: anyone must be able to retrace the path from the raw file to the clean one.
3. Look for errors in the qualitative variables. In an empty cell write =SORT(UNIQUE(C2:C162)) for sex, then do the same for E (mediu) and G (fumat). The UNIQUE function displays each distinct value once and distinguishes between uppercase and lowercase letters. Instead of two categories for sex, you should see four: F, f, Feminin, M; for mediu, urban, Urban, rural; for fumat, in addition to the three correct categories, also "Fumator " (capitalized and with a trailing space). Caution: the COUNTIF function does not distinguish between uppercase and lowercase, so =COUNTIF(C2:C162, "F") also counts the value "f" without warning you; JASP, by contrast, would treat them as two different categories. For invisible spaces, the function =LEN(G85) shows one character more than the expected length.
4. Look for errors in the numerical variables. For each numerical column, compute =MIN(...), =MAX(...), =COUNT(...) and =COUNTBLANK(...). Compare the minimum and maximum with the plausible range in the dictionary. You will find a maximum systolic blood pressure of 1540 mmHg (column M), a minimum height of 1.72 (column I) and a minimum weight of 7.26 kg (column J). The COUNT function counts only numerical values: for baseline blood glucose (column S) it gives 160 instead of 161, and the same for HbA1c (column V), a sign that each contains one value stored as text; =ISNUMBER(S65) gives FALSE for the cell containing "87 mg/dl", and =ISNUMBER(V90) gives FALSE for the HbA1c value "5,2", written with a comma in a file in which all other decimals use a point. The COUNTBLANK function gives 1 for age (column D) and 1 for blood glucose at 6 months (column U).
5. Look for duplicates. =COUNTUNIQUE(A2:A162) gives 160, although there are 161 rows: one patient code appears twice. In a helper column, write the formula =COUNTIF(A$2:A$162, A2) in AC2 and copy it down; the rows with the value 2 are the

duplicates. You will find patient P091 twice, with identical values in all columns. Delete the second row (the last row of the file) or use Data > Data cleanup > Remove duplicates.

6. Correct the errors following the rule from Chapter 1: a suspicious value is not corrected by intuition but checked against the source document. In this session, the role of the source document is played by the file mg2026_curat.csv, in which the values were taken from the case record forms. Table L1.1 summarizes the 12 problems and their corrections.

Table L1.1. Errors in mg2026_brut.csv, how they were detected and how they were corrected.

No.	Patient	Variable	Value found	Detection	Correction and source
1	P005	sex	f	UNIQUE	F (standardization; confirmed in the source record)
2	P021	sex	Feminin	UNIQUE	F (standardization; confirmed in the source record)
3	P034	tas0	1540	MAX	154 (extra digit; confirmed in the source record)
4	P042	varsta	missing	COUNTBLANK	70 (source record)
5	P057	inaltime	1,72	MIN	172 (meters converted to centimeters; checked from BMI and weight)
6	P064	glicemie0	"87 mg/dl"	COUNT, ISNUMBER	87 (unit deleted; confirmed in the source record)
7	P084	fumat	"Fumator "	UNIQUE, LEN	fumator (TRIM and lowercase; confirmed in the source record)
8	P089	hba1c0	"5,2" (text, with a comma)	COUNT, ISNUMBER	5.2 as a number (confirmed in the source record)
9	P102	greutate0	7,26	MIN	72.6 (misplaced decimal separator; checked from BMI and height)
10	P121	mediu	Urban	UNIQUE	urban (standardization)
11	P134	glicemie6	missing	COUNTBLANK	135 (source record)
12	P091	all	duplicate row	COUNTUNIQUE, COUNTIF	the second row is deleted

7. Discuss with the colleague next to you what you would have corrected if you had not had the source document. Each error in the file is a distortion of the true value, and for most of them there is a very likely hypothesis: "f" and "Feminin" mean F, "Urban" means urban, "Fumator " becomes fumator after removing the space (the TRIM function) and converting to lowercase (the LOWER function), "87 mg/dl" becomes 87 after deleting the unit, and "5,2" becomes the number 5.2. For two numerical values, the hypothesis can be

checked independently, because the file contains a redundant derived variable, BMI: for P057, the height is $\sqrt{\frac{77.7}{26.3}} = \sqrt{2.954} = 1.719m$, that is, 172 cm, exactly the value 1.72 converted from meters to centimeters; for P102, the weight is 27.0 x 1.64² = 27.0 x 2.6896 = 72.6 kg, so 7.26 is the correct value with the decimal separator misplaced. The value 1540 is most likely 154 with a doubled digit, but it could just as well have been a mistyped 150 or 140; nothing in the file allows it to be checked, so here confirmation in the source record is mandatory. The two missing values (the age of patient P042 and the 6-month blood glucose of patient P134) cannot be inferred in any way: only the source document can fill them in (70 years and 135 mg/dL, respectively). When the source is not available, the value is left missing and noted in the log; an invented value is more dangerous than a missing one, even when the hypothesis seems obvious.

8. After the corrections, check the result. The "brut_lucru" sheet should have 160 data rows, and the statistics should match those of the clean file: =AVERAGE(M2:M161) gives 131.13; =COUNT(S2:S161) gives 160; =MIN(I2:I161) gives 145 cm; =MAX(M2:M161) gives 169 mmHg; =COUNTBLANK(D2:D161) gives 0; =SORT(UNIQUE(C2:C161)) displays only F and M. If all checks pass, your cleaned dataset is identical to mg2026_curat.csv.

Notice how much the errors would have changed the results had they not been found. On the uncleaned raw file (161 rows, with the value 1540 and the duplicate), =AVERAGE(M2:M162) gives 139.70 mmHg instead of 131.13, and =STDEV(M2:M162) gives 111.89 mmHg instead of 13.88: a single mistyped digit moved the group mean to the threshold of hypertension and multiplied the standard deviation almost eightfold. The median, =MEDIAN(M2:M162), stays at 131 mmHg, exactly as in the clean file, which illustrates the robustness of the median discussed in Chapter 2. When the mean and median of a symmetric variable differ this much, and the standard deviation is of the same size as the mean, the first suspect is a data entry error.

L1.4 Exploring the JASP Interface

JASP is the program in which you will do most of the analyses this semester. The interface has three areas: the data (on the left, like a spreadsheet), the options panel of the current analysis (in the middle) and the results (on the right), which update as soon as you change an option.

1. Open JASP. From the main menu (the button with three lines at the top left) choose Open > Computer > Browse and select mg2026_curat.csv. The data appear on the left.
2. Check the type of each variable. Next to the name of each column there is an icon: a ruler for numerical variables (Scale), three stepped bars for ordinal variables (Ordinal) and three circles for nominal variables (Nominal). JASP guesses the type on import, but not always correctly: educatie, activitate and satisfactie are imported as numbers, although they are ordinal; diabet_ogtt and scadere5 are imported as numbers, although they are dichotomous nominal variables. Click the icon and change the type. This step is not cosmetic: JASP offers different analyses depending on the declared type of the variable.
3. Choose Descriptives > Descriptive Statistics. Move the variables varsta, imc0, tas0 and glicemie0 into the Variables box. In the Statistics section, in addition to the default options (Valid, Missing, Mean, Std. deviation, Minimum, Maximum), also check Median, Quartiles, Skewness, Kurtosis and S.E. mean.
4. Compare the results with those from Google Sheets and from Table 2.12 in Chapter 2. You should obtain: varsta 56.29 ± 8.77 years, median 56; imc_0 $29.04 \pm \frac{3.91\ \text{kg}}{m^2}$, median 28.25; tas_0 131.13 ± 13.88 mmHg, median 131, standard error 1.097;

glicemie$_0$ 105.44 ± 15.76 mg/dL, median 104.5. JASP displays results with a decimal point, which is also the form you use in reports.

5. Move the variable sex into the Split box. The results are displayed separately for women and men; for SBP you should see 131.37 ± 14.47 mmHg in women (n = 86) and 130.85 ± 13.25 mmHg in men (n = 74).
6. For a nominal variable, for example fumat, check the Frequency tables option (in the Tables section or directly below the list of variables, depending on the version). The frequency table should show 74 never smokers, 33 former smokers and 53 current smokers.
7. Save the analysis as a JASP file (File > Save As, LP1_Surname_FirstName.jasp). The JASP file keeps the data, the options and the results together, which makes the analysis reproducible. Results can also be exported as PDF or HTML (File > Export Results).

L1.5 Exploring the MedCalc Interface

MedCalc has a spreadsheet-type interface, with menus for analyses and graphs. It is the preferred program for ROC curves and for the analysis of diagnostic tests, which you will use in Sessions 11-12.

1. Open MedCalc and import the file mg2026_curat.csv via File > Open, choosing the text/CSV file type. In the import window, check that the separator is the comma and that the decimal symbol is the point.
2. Choose Statistics > Summary statistics, select the variable tas0 and run the analysis. MedCalc displays the number of observations, the mean, the 95% confidence interval for the mean (128.96-133.30 mmHg, which you will learn about in Chapter 4), the standard deviation, the standard error, the median, the percentiles and, optionally, the normality tests.
3. Choose Graphs > Box-and-whisker plot for tas0, optionally with sex as the grouping variable. Compare the graph with Figure 2.3D in Chapter 2.
4. If a MedCalc license is not available on your computer, open the free online calculators on the manufacturer's website. Try the calculator for the confidence interval of a proportion: enter 29 patients with diabetes out of 160 and note the result; you will interpret it in Chapter 4 (the Wilson method gives 12.9-24.8%).

L1.6 Epi Info in Brief

Epi Info 7 (latest version 7.2.7) was developed by the CDC for field epidemiology and works entirely without an internet connection. The CDC stopped development of and support for the program in September 2025; the program remains usable but no longer receives updates. It has three main components: Form Designer (building data collection forms), Enter (data entry) and Analysis (analysis, in two variants: Classic Analysis, which is command-based, and Visual Dashboard, which is menu-based).

1. Open Epi Info and choose Classic Analysis. With the READ command (the Read button in the command list), import the file mg2026_curat.csv, choosing the text/CSV source type.
2. Run the FREQ command for the variable fumat and the MEANS command for the variable tas0. Check that you obtain the same frequencies (74, 33, 53) and the same mean (131.13 mmHg) as in the other programs.

3. Note that Epi Info also displays the program generated by the commands, which can be saved and run again. You will return to Epi Info in the sessions on risk analysis, where 2x2 tables with relative risk and odds ratio are obtained directly.

L1.7 Uploading the Results to Moodle

1. Download the Google Sheets spreadsheet as an Excel file: File > Download > Microsoft Excel (.xlsx). The file must contain the sheets: date (the demonstration dataset), dictionar, brut_original, brut_lucru and jurnal.
2. Prepare a short document (one page, PDF) with: the table of the 12 problems found and corrected, the check statistics (n, mean and standard deviation of SBP before and after cleaning) and two or three sentences about what you learned from the cleaning exercise.
3. Upload to Moodle, under the LP1 activity, the files LP1_Surname_FirstName.xlsx, LP1_Surname_FirstName.jasp and LP1_Surname_FirstName.pdf, before the deadline shown on the platform [52].

Additional Exercises

1. On the BD_Studenti file on Moodle (the student database used in previous years in the discipline): determine the sample size (N), classify each variable by type, write a complete data dictionary (name, description, type, unit or codes, plausible range) and use the functions MIN, MAX, COUNT, COUNTBLANK and UNIQUE to check whether the file contains missing values, values outside the plausible ranges or inconsistent coding. Add a calculated column with each student's BMI. No numerical results are given; the check is done in the next session.
2. Problem with data in the statement. A colleague entered the data of eight patients from a hypertension study in the table below. Find the errors, propose a way of checking each one and compute the mean SBP before and after correction, assuming that the source record confirms the value 152 mmHg for P4 and that the missing value for patient P7 cannot be recovered.

Table L1.2. Data entered for eight patients (exercise 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. The errors are: (1) P3, sex "m" instead of "M" (inconsistent coding; detected with UNIQUE; standardized to M after confirmation); (2) P4, SBP 1520 mmHg, biologically impossible (detected with MAX; corrected to 152 according to the source record); (3) P5, fumat "Nefumator" capitalized (UNIQUE; standardized to nefumator); (4) P7, SBP missing (COUNTBLANK; remains missing and is noted in the log); (5) row P2 duplicated (COUNTIF on id; the second row is deleted). Before correction, the mean SBP computed by =AVERAGE on the 7 numerical values (including the duplicate and the value 1520) is (138 + 145 + 132 + 1520 + 128 + 160 + 145)/7 = 2,368/7 = 338.3 mmHg, an absurd value. After correction, 7 distinct patients remain, 6 of them with a known SBP: (138 + 145 + 132 + 152 + 128 + 160)/6 = 855/6 = 142.5

mmHg. The AVERAGE function ignores the empty cell, but the number of patients analyzed (n = 6 of 7) must be reported explicitly.

3. On the file mg2026_curat.csv, create a new column, "obez" (obese), with the formula =IF(K2>=30, 1, 0), and count the obese patients with =SUM(...). You should obtain 54. Then create a column "tas_crescuta" (elevated SBP) with =IF(M2>=140, 1, 0); you should obtain 41 patients. Discuss what information is lost through these dichotomizations (Chapter 2, Section 2.1.4).

What to submit and indicative grading scheme

The three files described in L1.7 are submitted on Moodle. The indicative grading scheme, out of 10 points: the demonstration dataset with a correct structure (one row per patient, variable names without spaces) and BMI computed with a formula, 1.5 points; the complete data dictionary, 1.5 points; working validation rules, 1 point; correct import of the clean file, with the check statistics, 1 point; identification of the 11 errors and the duplicate, 2 points (one point for each half of the list); the change log, with the source of each correction, 1 point; the JASP file with the variable types set correctly and the descriptive statistics, 1 point; the PDF document with the interpretation, 1 point. One point is lost if the final statistics do not match those of the clean file (for example, a mean SBP different from 131.13 mmHg).

Offline variant

If the network is down, the session can be done entirely on the local computer. The demonstration dataset, the dictionary, validation and cleaning are done in LibreOffice Calc or Microsoft Excel, with the same functions (in the Romanian-language version of Excel, function names are translated, for example MEDIE instead of AVERAGE, and the argument separator is ";", as with the Romania setting); data validation is found in the Data > Data Validation menu. JASP and Epi Info work without an internet connection once installed. The mg2026 files are copied from the instructor's USB stick or from the laboratory's local server. Uploading to Moodle is done later, within the deadline shown.

Practical Session 2. Describing and Organizing Data: The Medical Questionnaire, Frequency Tables and Graphs

Objective

In this session you follow the complete path of a variable, from the question put to the patient to the table and graph in a report. You first build a short medical questionnaire in Google Forms and see how the type of each question determines the type of the variable you will analyze. Then, on the mg2026 dataset, you build frequency tables, compute descriptive indicators and draw the appropriate graphs, first in Google Sheets, then in JASP. All expected results are given in the text, so that you can check your work yourselves.

By the end, the student should be able to:

- design a short questionnaire in Google Forms, choosing for each question the type that produces the desired variable and the appropriate validation rules;
- link the form to a Google Sheets spreadsheet and recode the answers for analysis;
- build tables of absolute, relative and cumulative frequencies with COUNTIF, COUNTIFS, FREQUENCY and with pivot tables;
- compute in Google Sheets and in JASP the mean, median, mode, standard deviation, quartiles, coefficient of variation, skewness and excess kurtosis;
- build the histogram, box plot, bar chart and pie chart and interpret the shape of the distribution.

Tools

Software: Google Forms and Google Sheets (the university Google account) [43-46], JASP 0.98.1 [47]. Files: mg2026_curat.csv (or the spreadsheet cleaned in Practical Session 1), BD_Studenti for the additional exercises. Approximate time: 100 minutes (about 30 minutes for the questionnaire and 70 for the descriptive analysis).

The columns of the mg2026 dataset are those from Practical Session 1: C sex, D varsta, E mediu, F educatie, G fumat, H activitate, K imc0, M tas0, R trigliceride, S glicemie0, AB satisfactie; the data occupy rows 2-161.

Procedure

L2.1 Designing a Short Medical Questionnaire in Google Forms

The scenario: a family practice wants to find out how often its patients with hypertension measure their blood pressure at home and how satisfied they are with the control of their disease. The questionnaire is anonymous, is completed in the waiting room on the practice's tablet and has no more than 10 questions.

1. Open Google Forms from your university account and create a new form, named "LP2 Home blood pressure self-monitoring". In the description, state the purpose of the questionnaire, the fact that it is anonymous and voluntary, and how long it takes to complete.
2. The first question is the consent question: "I agree to complete this anonymous questionnaire, and my answers may be used for statistical purposes", of the multiple-choice type, with the options "Yes" and "No", required. Use section branching (Go to

section based on answer) so that the answer "No" sends the respondent directly to the end of the form [148]. Without consent, nothing is collected.

3. Add the questions below, choosing the type indicated. Table L2.1 shows the link between the question type in Google Forms and the type of variable that will result. This is the most important link in this session: the type of the variable is decided when you write the question, not when you do the analysis.

Table L2.1. Question types in Google Forms and the resulting variables.

Question	Type in Google Forms	Validation	Resulting variable
Your age (completed years)	Short answer	Number, between 18 and 100	discrete numerical (age in years)
Sex	Multiple choice: female, male	required	dichotomous nominal
Area of residence	Multiple choice: urban, rural	required	dichotomous nominal
How many times did you measure your blood pressure at home in the past week?	Short answer	Whole number, between 0 and 50	discrete numerical
Last systolic blood pressure value measured at home (mmHg), if you know it	Short answer	Number, between 70 and 260; not required	continuous numerical, with missing values allowed
Which device do you use?	Multiple choice: upper arm, wrist, I do not have a device	required	nominal
What prevents you from measuring more often?	Checkboxes (multiple answers)	-	several dichotomous variables (one for each option)
How satisfied are you with the control of your blood pressure?	Linear scale 1-5 (1 very dissatisfied, 5 very satisfied)	required	ordinal
Educational level	Drop-down: lower secondary school, high school or vocational school, university, postgraduate	required	ordinal

4. Set the validation rules. For the numerical questions, click the three dots in the corner of the question, choose Response validation and set "Number", "Between", with the limits from the table, plus a clear error message ("Please enter a number between 70 and 260") [45]. Validation prevents exactly the errors you cleaned in Practical Session 1: the text "87 mg/dl" instead of 87, a blood pressure of 1540, a weight of 7.26 kg, an age of 5 years.

5. Do not ask for identifying data: no name, no e-mail address (in Settings, the collection of e-mail addresses must be turned off), no full date of birth. Age in years is enough. This is the direct application of the principle of data minimization (Chapter 1, Section 1.8.3).

6. Test the form: complete it 5-6 times, as if you were different patients, with varied answers, including one who refuses consent and one who does not know their blood pressure. Also ask the colleague next to you to complete it 2-3 times from their phone.

L2.2 Linking the Form to Google Sheets and Recoding the Answers

1. In the form, open the Responses tab and choose Link to Sheets, with the option to create a new spreadsheet [46]. Google creates a spreadsheet in which each response is a row and each question a column; the first column contains the timestamp of submission. The

structure is exactly the one recommended in Chapter 2: one row per respondent, one column per variable [147].

2. Note how the different question types appear. Answers to multiple-choice questions appear as text ("female"), the linear scale as a number (1-5), and checkboxes as a text with all the checked options, separated by commas, in the same cell. The latter is not a good format for analysis: each option must be turned into a separate column, with 1 (checked) and 0 (unchecked). For the option "no time", for example, use the formula =IF(ISNUMBER(SEARCH("no time", H2)), 1, 0).
3. Recode the ordinal variables entered as text. For education, in a new column: =IFS(J2="lower secondary school", 1, J2="high school or vocational school", 2, J2="university", 3, J2="postgraduate", 4). Ordinal analysis needs numerical codes in the correct order; text would be sorted alphabetically, and the categories would end up in the correct order only by chance.
4. Remove the rows of respondents who refused consent (if the branching worked, the other columns are empty) and rename the columns with short names, without spaces: varsta, sex, mediu, nr_masurari, tas_casa, aparat, satisfactie, educatie. Add a "dictionar" sheet.

L2.3 Frequency Tables in Google Sheets on the mg2026 Dataset

For the descriptive analysis, work on the mg2026 dataset, which has 160 patients and allows meaningful results. Open the spreadsheet cleaned in Practical Session 1 or import mg2026_curat.csv again (with the procedure from L1.2) and work in a new sheet, "frecvente" (frequencies).

1. Nominal variable: smoking status. In A2:A4 write the categories nefumator, fost fumator, fumator. In B2 write =COUNTIF(date!$G$2:$G$161, A2) and copy it down. In C2, the relative frequency: =B2/SUM($B$2:$B$4)*100. You should obtain 74 (46.3%), 33 (20.6%) and 53 (33.1%). Replace the name of the data sheet ("date") with the name of your own sheet.
2. Ordinal variable: satisfaction (column AB). In A7:A11 write the values 1-5. Absolute frequency: =COUNTIF(date!$AB$2:$AB$161, A7). Relative frequency: =B7/160*100. Cumulative absolute frequency: in D7 =B7, in D8 =D7+B8, copied down. Cumulative relative frequency: =D7/160*100. The result should be Table 2.4 in Chapter 2: frequencies 4, 19, 54, 55, 28; relative frequencies 2.5; 11.9; 33.8; 34.4; 17.5%; cumulative frequencies 2.5; 14.4; 48.1; 82.5; 100.0%. The cumulative frequency passes 50% at the value 4, so the median is 4.
3. The same for education (column F): 20, 57, 64, 19 patients (12.5; 35.6; 40.0; 11.9%), with cumulative frequencies 12.5; 48.1; 88.1; 100.0% and median 3. For physical activity (column H): 21, 42, 60, 34, 3 patients, with median 3.
4. Continuous variable grouped into classes: SBP (column M). First compute the number of classes using Sturges' rule, =1+3.322*LOG10(160), which gives 8.32, and the width =(MAX(date!M2:M161)-MIN(date!M2:M161))/8, which gives 8.4; round the width to 10 mmHg. In A14:A20 write the lower limits 100, 110,..., 160. The frequency of each class with COUNTIFS: =COUNTIFS(date!$M$2:$M$161, ">="&A14, date!$M$2:$M$161, "<"&A14+10). You should obtain 11, 21, 35, 52, 25, 13, 3, that is, Table 2.5 in Chapter 2.
5. The same frequencies are obtained with the FREQUENCY function, which takes the data and the upper limits of the classes. Write the upper limits 109, 119, 129, 139, 149, 159 in F14:F19, then in a single cell =FREQUENCY(date!M2:M161, F14:F19). The function

returns 7 values (the last for values above 159) and writes them automatically in the cells below. Check that the sum of the frequencies is 160.

6. The BMI classes used in clinical practice have unequal widths: below 25 (normal weight), 25-29.9 (overweight), 30-34.9 (class I obesity), 35-39.9 (class II), ≥ 40 (class III). Build this table with COUNTIFS on column K. Expected result: 18, 88, 43, 8 and 3 patients; obese in total 43 + 8 + 3 = 54 (33.8%).

L2.4 Pivot Tables and Contingency Tables

A pivot table automatically summarizes the data by one or more categorical variables, without formulas [44].

1. Select the data (A1:AB161) and choose Insert > Pivot table, creating a new sheet.
2. In the pivot table editor, add the variable fumat to Rows, the variable sex to Columns, and the variable id to Values, summarized by COUNTA. You should obtain the sex x fumat contingency table: women, 16 former smokers, 30 current smokers, 40 never smokers; men, 17, 23 and 34; totals 33, 53, 74 and 160.
3. Change the display of the values to percentages of the column total (Show as > % of column). Compare the proportion of current smokers among women (30/86 = 34.9%) and among men (23/74 = 31.1%).
4. Build a second pivot table, brat x satisfactie. Expected result: arm A, satisfaction 1-5: 1, 13, 28, 9, 2; arm B: 3, 3, 14, 24, 9; arm C: 0, 3, 12, 22, 17. Even before any statistical test, the table shows that patients in arm C are more satisfied: 39 of 54 (72.2%) chose 4 or 5, compared with 11 of 53 (20.8%) in arm A. The appropriate test for this comparison, Mann-Whitney or Kruskal-Wallis, will be applied in Practical Session 7.
5. The pivot table can also compute means: put brat in Rows and tas0 in Values, summarized by AVERAGE. You should obtain 131.53 (A), 132.79 (B) and 129.11 (C) mmHg.

L2.5 Descriptive Statistics in Google Sheets

In a new sheet, "descriptive", build a table with the variables in columns and the indicators in rows. The formulas for SBP (column M) are:

Table L2.2. Google Sheets formulas for the descriptive statistics of SBP and the expected results.

Indicator	Formula (United States setting)	Expected result
n	=COUNT(date!M2:M161)	160
Mean	=AVERAGE(date!M2:M161)	131,13
Median	=MEDIAN(date!M2:M161)	131
Mode	=MODE(date!M2:M161)	138
Standard deviation	=STDEV(date!M2:M161)	13,88
Variance	=VAR(date!M2:M161)	192,66
Standard error	=STDEV(date!M2:M161)/SQRT(COUNT(date!M2:M161))	1,097
Coefficient of variation (%)	=STDEV(date!M2:M161)/AVERAGE(date!M2:M161)*100	10,6
Minimum, maximum	=MIN(date!M2:M161), =MAX(date!M2:M161)	102, 169
Q1, Q3	=QUARTILE(date!M2:M161, 1), =QUARTILE(date!M2:M161, 3)	122, 140
10th and 90th percentiles	=PERCENTILE(date!M2:M161, 0.1), =PERCENTILE(date!M2:M161, 0.9)	112; 148,2

Indicator	Formula (United States setting)	Expected result
Skewness	=SKEW(date!M2:M161)	0,01
Excess kurtosis	=KURT(date!M2:M161)	-0,10

Note the argument 0.1 in the PERCENTILE function: in the United States setting used here, the decimal is written with a point and arguments are separated by commas (with the Romania setting, the decimal comma and the semicolon are used instead). Copy the formulas for the other variables as well, changing the column letter. The expected results:

Table L2.3. Expected results for other variables in mg2026.

Variable (column)	Mean	SD	Median	Q1	Q3	Skewness	Excess kurtosis
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

Also compute the geometric mean of triglycerides, =GEOMEAN(date!R2:R161), which should give 144.0 mg/dL, and compare it with the arithmetic mean (161.18) and the median (151.5). Compute the limits for extreme values of BMI: =QUARTILE(date!K2:K161, 3)+1.5*(QUARTILE(date!K2:K161, 3)-QUARTILE(date!K2:K161, 1)) gives 39.2 kg/m², and =COUNTIF(date!K2:K161, ">"&39.2) gives 3 patients.

L2.6 Graphs in Google Sheets

1. Bar chart for fumat. Select the frequency table from L2.3, step 1 (the categories and the absolute frequencies), then Insert > Chart. In the chart editor choose the "Column chart" type. In the Customize tab add the title ("Smoking status, mg2026, n = 160"), the axis titles and the data labels. Check that the vertical axis starts at 0.
2. Pie chart for education. Select the frequency table for education (with the category labels as text, not the codes 1-4) and choose the "Pie chart" type. Display the percentages on the slices: 12.5; 35.6; 40.0; 11.9%. Do not use the 3D version.
3. Histogram of blood glucose. Select the glicemie0 column (S1:S161) and choose the "Histogram" type. Under Customize > Histogram set the bucket size to 10. The expected shape: a broad peak between 90 and 120 mg/dL (34 patients in each of the classes 90-99, 100-109 and 110-119), with a short tail to the right, up to 154 mg/dL. Also try a bucket size of 5, then of 20, and observe how the appearance changes.
4. The box plot is not available as a standard chart type in Google Sheets. You will build it in JASP, in the next step. You can, however, display the five-number summary of SBP in a small table: 102, 122, 131, 140, 169.

L2.7 Descriptive Analysis and Graphs in JASP

1. Open mg2026_curat.csv in JASP and check the variable types, as in Practical Session 1: educatie, activitate and satisfactie ordinal; sex, mediu, fumat, brat, diabet_ogtt nominal.
2. Descriptives > Descriptive Statistics, with the variables varsta, imc0, tas0, glicemie0, trigliceride. Check Mean, Std. deviation, Median, Quartiles, Minimum, Maximum, Skewness, Kurtosis, Shapiro-Wilk test (you will interpret it in Chapter 3). Compare with Table L2.3; the values should match.

3. For the variables fumat, educatie and satisfactie, check Frequency tables. JASP displays the frequencies, the percentages, the valid percentages (which exclude missing values) and the cumulative percentages, exactly as in Table 2.4.
4. In the Basic plots (or Plots) section, check Distribution plots for imc0 and trigliceride. Compare the histograms with Figure 2.2 in Chapter 2: both distributions have a tail to the right.
5. For the box plot, put tas0 in Variables and sex in Split, then check Boxplots (optionally also the Boxplot element, Violin element and Jitter element options, which also show the individual points). The result should resemble Figure 2.3D: median 131 mmHg in both sexes, with the boxes almost overlapping. In the table, women au 131.37 ± 14.47 mmHg$(n = 86)$, and men 130.85 ± 13.25 mmHg$(n = 74)$.
6. Repeat the box plot for imc0, without splitting by sex. Identify on the graph the three extreme values above 39.2 kg/m² (40.0; 43.4 and 48.3).

L2.8 Interpreting the Distribution

For the report you submit, write one interpretation paragraph each for SBP, BMI and triglycerides. The model for SBP: "Systolic blood pressure at enrollment had an approximately symmetric distribution (skewness 0.01, excess kurtosis -0.10), with a mean of 131.13 ± 13.88 mmHg, practically equal to the median (131 mmHg; Q1-Q3 122-140), and values between 102 and 169 mmHg. For this variable, the mean and standard deviation are the appropriate indicators." For BMI and triglycerides, use the skewness, the relationship between mean and median, and the box plot to justify reporting by median (Q1-Q3).

Additional Exercises

1. On the BD_Studenti file on Moodle: produce the appropriate frequency tables and graphs for gender, eye color, hours of sleep and type of housing; compute the measures of central tendency and dispersion for age, BMI (calculated with the formula from weight and height), hours of sleep and hours of study; compare, with an appropriate graph, eye color in girls and boys and hours of sleep in girls and boys. Justify the choice of each indicator and each graph according to the type of variable. No numerical results are given; they are discussed in the next session.
2. Table with missing values: HbA1c in patients with diabetes. A physician grouped the HbA1c values of the 40 patients with type 2 diabetes on their list. Fill in the values A, B, C, D, E and F and answer: what proportion of patients reached the HbA1c target < 7%? What proportion have HbA1c ≥ 8%?

Table L2.4. HbA1c in 40 patients with type 2 diabetes.

HbA1c (%)	**Absolute frequency**	**Relative frequency (%)**	**Cumulative frequency (%)**
< 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. $A = \frac{9}{40} \times 100 = 22.5\%$ (check: 12.5 + 22.5 = 35.0). B = 0.30 x 40 = 12 patients. C = 35.0 + 30.0 = 65.0%. $D = \frac{8}{40} \times 100 = 20.0\%$ (check: 65.0 + 20.0 = 85.0). F = 100.0 - 85.0 = 15.0%, and E = 0.15 x 40 = 6 patients. Check of the sum: 5 + 9 + 12 + 8 + 6 = 40. The HbA1c

target < 7% was reached by 5 + 9 = 14 patients, that is, 35.0% (the cumulative frequency of the 6.5-6.9 class). HbA1c ≥ 8% is found in 8 + 6 = 14 patients, that is, 20.0 + 15.0 = 35.0%. The median class is 7.0-7.9: the 20th and 21st ordered values lie in this class, because the cumulative frequency goes from 35.0% (14 patients) to 65.0% (26 patients). Clinical interpretation: only a third of the patients are at target, and another third have poor glycemic control that calls for intensification of treatment.

3. Table with missing values: satisfaction with the services of an outpatient clinic. A satisfaction survey was answered by 120 patients, on a scale from 1 (very dissatisfied) to 5 (very satisfied). Fill in the missing values, determine the median and the mode, and compute the proportion of patients who are satisfied or very satisfied.

Table L2.5. Satisfaction of 120 patients of an outpatient clinic.

Satisfaction	**Absolute frequency**	**Relative frequency (%)**	**Cumulative absolute frequency**	**Cumulative frequency (%)**
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 (check: 6 + 15 = 21). $B = \frac{21}{120} \times 100 = 17.5\%$ (or 5.0 + 12.5). $C = \frac{33}{120} \times 100 = 27.5\%$ (check: 17.5 + 27.5 = 45.0). D = 0.35 x 120 = 42 patients. E = 54 + 42 = 96 (check: 96/120 = 80.0%). F = 0.20 x 120 = 24 patients (check: 96 + 24 = 120). Median: n = 120, so the median is the mean of the values in positions 60 and 61; the cumulative frequency is 54 at the value 3 and 96 at the value 4, so both positions have the value 4, and the median is 4. The mode is also 4 (42 patients). Satisfied or very satisfied: 35.0 + 20.0 = 55.0% (66 patients). The correct way to report this ordinal variable is "median 4 (Q1-Q3 3-4)"; the quartiles are found in the same way as the median, at positions 30-31 (value 3) and 90-91 (value 4).

4. Build a data collection form in Google Forms for 5 variables of the mg2026 dataset (sex, age, SBP, blood glucose, satisfaction), with the appropriate validations, and check that you cannot enter any of the errors in mg2026_brut.csv (for example, SBP 1540, blood glucose "87 mg/dl", sex "Feminin").

What to submit and indicative grading scheme

Upload to Moodle: the link to the Google Forms form (or a screenshot of the questions), the Google Sheets spreadsheet downloaded as .xlsx (the sheets for responses, dictionar, frecvente, pivot, descriptive), the JASP file and a 1-2 page PDF report with the frequency tables, the descriptive statistics, at least four graphs (bar, pie, histogram, box plot) and the interpretation paragraphs from L2.8. The indicative grading scheme, out of 10 points: the questionnaire, with correctly chosen question types, validations and consent, 2 points; linking to Sheets and recoding, 1 point; correct frequency tables (fumat, satisfaction, SBP by class), 2 points; the pivot tables, 1 point; correct descriptive statistics, with the indicator appropriate to the type of variable, 2 points; the graphs, with titles, axes and units, 1 point; the interpretation of the distribution, 1 point.

Offline variant

Without an internet connection, the questionnaire is designed on paper (with the same question types and the same validation rules written as instructions), and the answers are entered in a LibreOffice Calc or Microsoft Excel spreadsheet, with data validation set from the Data menu. The frequency tables are built with the same functions (COUNTIF, COUNTIFS, FREQUENCY; in the Romanian-language version of Excel, CONTARDACĂ, FRECVENȚĂ), the pivot tables

from the Insert > PivotTable menu, and the descriptive statistics and graphs (including the box plot) in JASP, which works offline. The Google form is built later, as homework.

Practical Session 3. Distributions and Normality Testing

Objective

This session turns Chapter 3 into working steps. We start from five physiological values measured at the enrollment of patients in the mg2026 study (age, systolic blood pressure, BMI, fasting plasma glucose and triglycerides) and ask, for each one, whether its distribution can be considered normal. The answer is not given by a single number: we build it from graphs, from shape indicators and from normality tests, and then use it to decide which descriptive indicators to report and which type of test we will be able to apply in the following sessions. We work in parallel in JASP, MedCalc and Google Sheets, because in the practical exam you will have to choose one of them.

By the end, the student should be able to:

- obtain in JASP and MedCalc the skewness, the excess kurtosis, their standard errors, the Shapiro-Wilk test, the histogram with the normal curve and the Q-Q plot for a continuous variable;
- compute in Google Sheets the skewness (SKEW), the excess kurtosis (KURT), their standard errors, z-scores (STANDARDIZE) and theoretical proportions under the normal curve (NORM.DIST, NORM.INV);
- build a Q-Q plot by hand in Google Sheets;
- apply a logarithmic transformation, check its effect and back-transform the mean into a geometric mean;
- identify extreme values using the z-score and the 1.5 x IQR rule;
- write a reasoned decision of the type "variable X can/cannot be treated as normal", in the format required for the discipline's assignments.

Tools

Software: JASP 0.98.1 or later [47]; MedCalc 24.0 [50] (or its online calculators); Google Sheets with English (United States) regional settings (decimal point, "," as the argument separator) [43]; optionally Epi Info 7.2.7 [51]. Files: mg2026_curat.csv (160 rows), from Moodle, the Emediqual platform [52]; for the additional exercises, the discipline's classic files BD_Studenti and Probleme BPOC (COPD problems). Approximate time: 100 minutes.

After import into Google Sheets, the columns used today are: D = varsta, K = imc0, M = tas0, R = trigliceride, S = glicemie0, V = hba1c0. The data occupy rows 2-161. The expected values in this session come from the dataset's results file (computed with Python) and match those displayed by JASP and MedCalc; differences in the last decimal may arise from rounding.

Procedure

L3.1 Preparing the File

1. Open mg2026_curat.csv in JASP (File > Open > Computer). In the data view, check that varsta, imc0, tas0, glicemie0, trigliceride and hba1c0 are marked as continuous variables (the ruler icon). If a variable appears as nominal (the three-circles icon), this usually means that a cell contains text; go back to Chapter 2 and to mg2026_brut.csv to see where the problem may come from.

2. In JASP, create a new variable with the natural logarithm of triglycerides: click the "+" sign to the right of the last column, name it ln_trig, choose "R" as the computation mode and write the formula log(trigliceride) (in R, the log function is the natural logarithm). The first value should be 5.1417, because the first patient has triglycerides of 171 mg/dL.
3. Import the same file into Google Sheets (File > Import > Upload; comma separator). Check that the BMI decimals were read as numbers (right-aligned), not as text. In column AC write the header ln_trig, and in AC2 the formula =LN(R2), copied down to row 161.

L3.2 The Shape of the Distribution in JASP

4. Descriptives > Descriptive Statistics menu. Move into the Variables box: varsta, tas0, imc0, glicemie0, trigliceride and ln_trig.
5. In the Statistics section check: Mean, Median, Std. deviation, Minimum, Maximum, Quartiles; under Distribution check Skewness, Kurtosis and Shapiro-Wilk test. JASP will also automatically display the standard errors of skewness and kurtosis.
6. In the Plots section (or Basic plots, depending on the version) check Distribution plots (with the option to display the density), Q-Q plots and Boxplots.
7. Compare the JASP table with Table L3.1. All values should match up to rounding. The standard error of skewness is the same for all variables (0.192), because it depends only on n; the same holds for the standard error of kurtosis (0.381).

Table L3.1. Self-check values on mg2026 (n = 160).

Variable	Mean	SD	Median	Skewness	Excess kurtosis	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. For each variable, compute the ratios skewness / 0.192 and kurtosis / 0.381 and write them in your notebook. You should obtain, for example, 7.42 and 9.64 for BMI, 4.97 and 2.94 for triglycerides, and 1.92 and -0.60 for blood glucose. Mark the ratios that exceed 2 in absolute value.
9. Look at the Q-Q plots. For tas0 and varsta, the points should follow the diagonal. For imc0, the last 3-4 points on the right rise well above the line (the patients with BMI 40.0, 43.4 and 48.3). For triglycerides, the points form a concave-up curve, typical of positive skewness. For ln_trig, the curve straightens out. Sketch in your notebook the Q-Q shape for BMI and for triglycerides.
10. Control question: how do you interpret blood glucose, for which Shapiro-Wilk gives p = 0.037, but the skewness relative to its standard error is below 2 and the mean differs from the median by less than 1 mg/dL? The expected answer, which you will find argued in Section 3.8.5 of the course, is that the departure from normality is small, explainable by patients with undiagnosed diabetes, and does not prevent the use of the mean and of parametric tests at n = 160; it is, however, reported transparently.
11. Optionally, check normality by group: in the same window, move the variable sex into the Split box. For blood glucose you will notice that the departure comes mainly from the group of women (skewness 0.57, Shapiro-Wilk p = 0.022), whereas in men the distribution is compatible with normality (p = 0.736). For tas0 by arm (Split: brat), all

three tests are nonsignificant (p = 0.782; 0.373; 0.818). Remember the idea: when comparing groups, what matters is normality within each group.

L3.3 The Same Analyses in MedCalc

12. In MedCalc, choose File > Open and select mg2026_curat.csv (on import, confirm that the first line contains the variable names and that the separator is the comma).
13. Statistics > Summary statistics menu. Under Variable choose tas0. In the options window, check the test for normal distribution (Shapiro-Wilk, or alternatively Kolmogorov-Smirnov) and, if available, the option for shape statistics. The output shows the mean, SD, standard error, 95% CI for the mean, the median with its CI, skewness and kurtosis (each with the p value of its own test) and the conclusion of the normality test ("accept Normality" or "reject Normality").
14. Compare the values with Table L3.1. Pay attention to two details. First: MedCalc's wording "accept Normality" is a shorthand; in your report write "the normality hypothesis is not rejected". Second: for the Kolmogorov-Smirnov test, check in the program's documentation whether the Lilliefors correction is applied. On BMI, the test without correction gives p ≈ 0.08, and with the Lilliefors correction p ≈ 0.002; Shapiro-Wilk gives p < 0.001. The difference shows why the classic test, without correction, is too lenient.
15. Graphs: Graphs > Histogram (check the normal curve overlay) and Graphs > Normal plot (MedCalc's Q-Q plot). For triglycerides, the histogram with the superimposed normal curve clearly shows that the actual peak lies to the left of the peak of the curve, and the right tail extends beyond the curve.
16. Repeat for imc0, glicemie0 and trigliceride. For triglycerides, in the Summary statistics window also check the logarithmic transformation option (if your version offers it), or create a new variable in the MedCalc data sheet, defined by the formula LN(trigliceride). The geometric mean displayed should be 144.0 mg/dL.

L3.4 Shape Indicators and the Q-Q Plot in Google Sheets

17. In a new sheet, named "Forma" (shape), build a table with the variables in rows and the indicators in columns. For SBP, the formulas are:

- n: =COUNT(Date!M2:M161), result 160;
- mean: =AVERAGE(Date!M2:M161), result 131.13;
- median: =MEDIAN(Date!M2:M161), result 131;
- SD: =STDEV(Date!M2:M161), result 13.88;
- skewness: =SKEW(Date!M2:M161), result 0.01;
- excess kurtosis: =KURT(Date!M2:M161), result -0.10;
- Q1 and Q3: =QUARTILE(Date!M2:M161, 1) and =QUARTILE(Date!M2:M161, 3), results 122 and 140.

Copy the row for the other variables, changing the column. The SKEW and KURT functions in Google Sheets give the same sample-corrected values as JASP and MedCalc.

18. The standard errors of skewness and kurtosis, with the exact formulas, for n in cell B2:

- SE of skewness: =SQRT(6*B2*(B2-1)/((B2-2)*(B2+1)*(B2+3))), result 0.192;
- SE of kurtosis: =2*C2*SQRT((B2^2-1)/((B2-3)*(B2+5))), where C2 contains the SE of skewness, result 0.381.

Compute the ratios and apply conditional formatting (Format > Conditional formatting) to cells with an absolute value greater than 2, with a red background. Google Sheets has no built-in Shapiro-Wilk test; for the p value, use JASP or MedCalc.

19. Build the Q-Q plot for SBP. In a new sheet, "QQ_TAS":

- in A2:A161 copy the SBP values and sort them in ascending order (Data > Sort range);
- in B2 write the rank, =ROW()-1, and copy it down (1 to 160);
- in C2 compute the cumulative proportion, =(B2-0.5)/160;
- in D2 compute the theoretical normal quantile, =NORM.S.INV(C2); the first value is -2.73, the last 2.73;
- select D2:D161 and A2:A161 (D on the X axis), Insert > Chart > Scatter chart; add a linear trendline.

The points should lie very close to the line; the correlation between the two columns, =CORREL(D2:D161, A2:A161), is 0.996. Repeat for triglycerides: the correlation drops to 0.972, and the points curve away at the right end. The W statistic of the Shapiro-Wilk test is based on exactly this idea: how well the ordered values line up with the normal quantiles.

20. Histogram: select column R (trigliceride), Insert > Chart, Chart type: Histogram chart. Under Customize > Histogram, try class widths (Bucket size) of 20 and then of 50 mg/dL and observe how the appearance changes. This is why we do not judge normality from the histogram alone.

L3.5 z-Scores and Proportions under the Normal Curve

21. In the data sheet, column AD, header z_tas. In AD2: =STANDARDIZE(M2, AVERAGE($M$2:$M$161), STDEV($M$2:$M$161)), copied down to row 161. The first patient has an SBP of 139 mmHg and z = 0.57. Check that the mean of the column is 0 and the SD is 1 (=AVERAGE(AD2:AD161), =STDEV(AD2:AD161)).
22. Count the patients with a z-score greater than 2 in absolute value: =COUNTIF(AD2:AD161, ">2") + COUNTIF(AD2:AD161, "<-2"). Result: 10 patients (6.25%), close to the 4.6% expected theoretically for a normal distribution. No patient has |z| > 3; the maximum z-score is 2.73, and the minimum -2.10.
23. The z-score and percentile of a patient with an SBP of 160 mmHg: =STANDARDIZE(160, 131.13, 13.88) gives 2.08; =NORM.S.DIST(2.08, TRUE) gives 0.9812. The patient is at the 98th percentile. Check the same result in Table E.1 of Appendix E (row 2.0, column 0.08).
24. The theoretical and the observed proportion for SBP ≥ 140 mmHg:

- theoretical: =1-NORM.DIST(140, AVERAGE(M2:M161), STDEV(M2:M161), TRUE), result 0.261 (26.1%);
- observed: =COUNTIF(M2:M161, ">=140")/COUNT(M2:M161), result 0.256 (25.6%; 41 patients).

25. The interval that should contain 95% of patients: =NORM.INV(0.025, AVERAGE(M2:M161), STDEV(M2:M161)) and =NORM.INV(0.975, ...) give 103.9 and 158.3 mmHg. Count how many patients actually lie between the two limits with =COUNTIFS(M2:M161, ">=103.9", M2:M161, "<=158.3"): the result is 150, that is, 93.8%.
26. Repeat steps 24-25 for BMI, with the cutoff of 30 kg/m². The normal model predicts 40.3% obese patients, but in the sample there are 33.8%. Write in your notebook why the normal model is wrong here (the positive skewness pulls the mean up and "widens" the curve).

L3.6 Logarithmic Transformation of Triglycerides

27. In the "Forma" sheet, add the row for ln_trig (column AC). The skewness should fall from 0.95 to -0.28, and the excess kurtosis from 1.12 to -0.28. In JASP, Shapiro-Wilk for ln_trig gives p = 0.200, compared with p < 0.001 for triglycerides.
28. Also try the square root transformation: column AE, =SQRT(R2). The skewness should be 0.32. Compare with the logarithm and justify your choice (see Table 3.7 in the course). Also try the reciprocal, =1/R2: the skewness increases to 1.48, meaning that the transformation is too strong.
29. The geometric mean, in two ways: =EXP(AVERAGE(AC2:AC161)) and =GEOMEAN(R2:R161). Both should give 144.0 mg/dL. Compare with the arithmetic mean (161.18) and the median (151.5).
30. The 95% confidence interval for the geometric mean is computed on the logarithmic scale and back-transformed:

- lower limit: =EXP(AVERAGE(AC2:AC161) - CONFIDENCE.T(0.05, STDEV(AC2:AC161), 160));
- upper limit: =EXP(AVERAGE(AC2:AC161) + CONFIDENCE.T(0.05, STDEV(AC2:AC161), 160)).

Sheets displays 133.44 and 155.42 mg/dL, the values from Lecture 3. As an experiment, round the limits on the logarithmic scale to two decimals before exponentiation and observe how far each limit moves; the conclusion, which you write in your notebook: on the logarithmic scale, round only at the end.

31. The reference interval for triglycerides on the logarithmic scale: =EXP(AVERAGE(AC2:AC161) - 1.96*STDEV(AC2:AC161)) and =EXP(AVERAGE(AC2:AC161) + 1.96*STDEV(AC2:AC161)) give 55.3 and 374.9 mg/dL. Compare with the naive interval, computed on the untransformed values, 11.6-310.7 mg/dL, and with the observed percentiles, =PERCENTILE(R2:R161, 0.025) and =PERCENTILE(R2:R161, 0.975), which give 53.8 and 364.0 mg/dL.

L3.7 Extreme Values

32. For BMI, compute Tukey's fences: Q1 = 26.2, Q3 = 31.4, IQR = 5.2; lower fence =QUARTILE(K2:K161, 1)-1.5*(QUARTILE(K2:K161, 3)-QUARTILE(K2:K161, 1)) = 18.4; upper fence = 39.2. Count the values above the upper fence with COUNTIF: the result is 3 (40.0; 43.4; 48.3).
33. Repeat for triglycerides: the fences are -37.4 and 347.6 mg/dL, and 5 values exceed the upper fence (364, 364, 373, 405, 442). For ln_trig, no value falls outside the fences. The conclusion, which you write in your notebook: the "extreme values" of triglycerides are the normal tail of a log-normal distribution, not errors.
34. The z-score of the patient with a BMI of 48.3 is 4.92. Check it with STANDARDIZE. Decide, in writing, what you do with this patient (expected answer: the value is checked in the source record and kept, because it is plausible, and the median and IQR or nonparametric tests are used for BMI; the patient is not deleted).

L3.8 Writing the Conclusion in the a-g Format

35. For BMI, write up the solution in the format used for the discipline's assignments. The expected model:

(a) Research hypothesis: the BMI of patients at cardiometabolic risk in the population studied has a normal distribution.

(b) Statistical hypotheses: H0, BMI comes from a population with a normal distribution; H1, the distribution of BMI in the population is not normal.

(c) Method: the Shapiro-Wilk test, recommended for small and medium samples, supplemented by skewness and kurtosis relative to their standard errors and by the Q-Q plot.

(d) Procedure: JASP, Descriptives > Descriptive Statistics, with the options from steps 5-6.

(e) Result: skewness 1.42 (SE 0.192), excess kurtosis 3.68 (SE 0.381), $W = 0.904$, $p < 0.001$; the Q-Q plot shows large departures in the right tail.

(f) Statistical conclusion: $p < 0.05$, H0 is rejected.

(g) Research conclusion: the distribution of BMI departs significantly from normality, with positive skewness due to a small group of patients with severe obesity. BMI will be described by the median and IQR (28.25; 26.2-31.4 kg/m²), and comparisons will use nonparametric tests or will also report their results.

36. Do the same for SBP (expected conclusion: normality is not rejected, $p = 0.303$; the mean is reported $\pm$SD, 131.13 ± 13.88 mmHg).

Additional Exercises

Exercises 1-3 are solved on the discipline's classic files, from Moodle; the numerical results are not given here, and the instructor checks them in the laboratory.

1. The BD_Studenti file. For the variables Varsta, IMC (calculated as weight / height², with height in meters), Ore_somn and Inaltime, obtain in JASP the skewness, the excess kurtosis, their standard errors and the Shapiro-Wilk test, plus the Q-Q plots. Repeat the analysis separately by sex (Split: Gen). For each variable, write a decision sentence like those in step 35. Watch out for variables with few distinct values (hours of sleep reported in whole hours): what effect does this have on the Shapiro-Wilk test?
2. The Probleme BPOC file (COPD problems). Analyze the distribution of baseline vital capacity (CVV1), baseline heart rate (FC1), baseline maximum (systolic) blood pressure (TMAX1) and baseline peak expiratory flow (DEM1). For each, state whether you would use the unpaired t-test or the Mann-Whitney test for the comparison by sex, and why. For FC1 and FC2, check the normality of the difference FC1 - FC2 (create a new variable in JASP), because it is the difference, not each value separately, that matters for the paired t-test.
3. On either of the two files, choose the most skewed variable, apply the logarithmic transformation and report the skewness before and after, the geometric mean and its 95% confidence interval.
4. Problem with data in the statement (with solution). In a reference population of full-term newborns, birth weight has a mean of 3400 g and a standard deviation of 450 g, with an approximately normal distribution (fictitious data). (a) What proportion of newborns weigh less than 2500 g? (b) What weight separates off the heaviest 10% of newborns? (c) A newborn weighs 4500 g: compute the z-score and the percentile.

Solution. (a) $z = (2500 - 3400) / 450 = -2.00$; $\Phi(-2.00) = 1 - 0.9772 = 0.0228$. Approximately 2.3% of newborns would weigh less than 2500 g. In Sheets: =NORM.DIST(2500, 3400, 450, TRUE). (b) $\Phi(z) = 0.90$ for $z = 1.28$; $x = 3400 + 1.28 \times 450 = 3976$ g (=NORM.INV(0.9, 3400, 450) gives 3977 g). (c) $z = (4500 - 3400) / 450 = 2.44$; $\Phi(2.44) = 0.9927$; the newborn is at the 99.3rd percentile. Note: in reality, the distribution of birth weight has a heavier left tail than the normal distribution (preterm infants, growth restriction), so the actual proportion below 2500 g is usually higher than the normal model predicts; the exercise illustrates once again that the model must be checked against the data.

5. Problem with data in the statement (with solution). In 50 patients, a variable has a skewness of 0.62 and an excess kurtosis of 0.40. The standard errors for n = 50 are 0.337 (skewness) and 0.662 (kurtosis). Is the distribution compatible with normality?

Solution. The ratios are 0.62 / 0.337 = 1.84 and 0.40 / 0.662 = 0.60, both below 2. The indicators do not signal a significant departure, but a skewness of 0.62 is not negligible in size: with 50 patients, the power to detect it is limited. The decision is made by also looking at the histogram, the Q-Q plot and the difference between the mean and the median; if the variable is a serum concentration or a duration, it is prudent to also report the median (IQR) and to check the result with a nonparametric test.

What to submit and indicative grading scheme

Upload to Moodle, named LP3_Surname_FirstName, a single file (PDF or exported Google document) containing:

- the shape table for the five variables and ln_trig, with the mean, median, SD, skewness, excess kurtosis, the ratios to the standard errors and the Shapiro-Wilk p (2 points);
- the Q-Q plots for SBP and triglycerides, one from JASP or MedCalc and one built by hand in Google Sheets (2 points);
- the calculations from L3.5: the z-score and percentile for 160 mmHg, the theoretical and observed proportions for SBP ≥ 140 and for BMI ≥ 30, with a two-sentence comment (2 points);
- the results of the logarithmic transformation: skewness before and after, the geometric mean and its 95% CI (1.5 points);
- the analysis of extreme values for BMI and triglycerides, with the reasoned decision (1 point);
- the solution in the a-g format for BMI and SBP (1.5 points).

The mistakes that most often lower the grade: "the test confirmed normality" instead of "the normality hypothesis is not rejected"; using the FALSE argument in NORM.DIST; deleting the patient with a BMI of 48.3; reporting the mean ± SD for BMI without any comment on skewness.

Offline variant

If there is no internet access, Google Sheets cannot be used, but JASP and MedCalc run locally. All steps in L3.2 and L3.3 can be done entirely offline. The steps in L3.4-L3.7 can be redone in any locally installed spreadsheet program (Microsoft Excel or LibreOffice Calc), with the same function names (SKEW, KURT, STANDARDIZE, NORM.DIST, NORM.INV, NORM.S.INV, GEOMEAN, QUARTILE, PERCENTILE); in the Romanian versions of Excel, the argument separator is ";", as with the Romania setting. Epi Info 7.2.7 can compute the mean, SD and quartiles in the Classic Analysis module (the MEANS command), but it does not offer normality tests. As a last resort, z-scores and proportions under the curve are computed by hand, with Table E.1 in Appendix E, on the values from Table L3.1.

Practical Session 4. Estimation and Confidence Intervals: The Effect of Sample Size

Objective

This session puts Chapter 4 into practice. We compute confidence intervals for means and for proportions on the mg2026 dataset, first with the software, then by hand in Google Sheets, to see that there is no "black box": every number displayed by JASP or MedCalc can be reconstructed from the mean, the standard deviation and n. On the same data, we build the classic 68.3%, 95.4% and 99.7% intervals for values and for the mean, we simulate in Google Sheets how the interval changes when we take samples of 10, 40 and 160 patients, and we end with sample size calculation in MedCalc, G*Power and Epi Info. This is one of the few sessions in which you will "see" sampling variability with your own eyes.

By the end, the student should be able to:

- obtain the 95% CI and the 99% CI for a mean in JASP, MedCalc and Google Sheets and check them by hand calculation;
- build and interpret, on the same data, the intervals for individual values and for the mean;
- compute the 95% CI for a proportion by the Wald, Wilson and exact methods;
- organize a simple simulation in Google Sheets and describe how the width and position of the interval depend on sample size;
- compute the sample size for a mean, for a proportion and for the comparison of two means, with at least two programs;
- report an estimate correctly: point value, 95% CI, unit of measurement, method.

Tools

Software: JASP 0.98.1 or later [47]; MedCalc 24.0 [50]; G*Power [77]; Epi Info 7.2.7, the StatCalc module [51] (the CDC stopped development of and technical support for Epi Info in September 2025; the program remains usable, and each step also has a MedCalc or G*Power alternative); Google Sheets with English (United States) regional settings [43]. Files: mg2026_curat.csv from Moodle [52]; for the additional exercises, BD_Studenti and Probleme BPOC (COPD problems). Approximate time: 100 minutes.

The Google Sheets columns used today: M = tas0, P = colesterol, Q = hdl, S = glicemie0, $X = \text{diabet}_{\text{ogtt}}\left(\frac{0}{1}\right)$, D = varsta. The data occupy rows 2-161.

Procedure

L4.1 The Confidence Interval for the Mean in JASP

1. Open mg2026_curat.csv in JASP. T-Tests > One Sample T-Test menu. Move tas0 into the Variables box. Leave Test value at 0: we are not interested in the test but in the interval, and with a test value of 0 the difference displayed is the mean itself.
2. Check "Location parameter" (or "Mean difference", depending on the version) and "Confidence interval", with the level at 95%. Also check "Descriptives". Expected result: mean 131.13 mmHg, SD 13.88, SE 1.097, 95% CI 128.96-133.30 mmHg. The t value and p displayed (the test against 0) have no clinical meaning and are ignored.

3. Change the confidence level to 99%: the interval becomes 128.27-133.99 mmHg. Then to 90%: 129.32-132.95 mmHg. Write down the three intervals and their widths (3.6; 4.3; 5.7 mmHg).
4. Repeat for glicemie0 (expected 95% CI 102.98-107.90 mg/dL) and for colesterol (203.63-215.15 mg/dL).
5. Optional: in Descriptives > Descriptive Statistics, recent versions of JASP offer a confidence interval for the mean directly, and the bootstrap option allows intervals for other statistics. If you have them, check that they give the same 95% CI for SBP.

L4.2 The Same Interval in MedCalc

6. In MedCalc, Statistics > Summary statistics, variable tas0. The output contains the rows "Arithmetic mean", "95% CI for the mean", "Standard deviation", "Standard error of the mean" and, further down, "Median" with "95% CI for the median". Compare: CI for the mean 128.96-133.30 mmHg; the interval for the median should be around 129-134 mmHg (programs may use slightly different versions of the rank method).
7. For triglycerides, the same menu displays a CI for the mean of 149.26-173.09 mg/dL. If you activate the logarithmic transformation option (where available), MedCalc reports the geometric mean of 144.0 mg/dL with a 95% CI of 133.44-155.42 mg/dL. Write down why the two intervals do not contradict each other (they estimate different parameters; see Section 4.9.1).

L4.3 Hand Calculation in Google Sheets

8. In a new sheet, "IC_medie" (CI for the mean), build the table step by step for SBP:

- B2, n: =COUNT(Date!M2:M161), result 160;
- B3, mean: =AVERAGE(Date!M2:M161), result 131.13;
- B4, SD: =STDEV(Date!M2:M161), result 13.88;
- B5, SE: =B4/SQRT(B2), result 1.097;
- B6, critical t: =T.INV.2T(0.05, B2-1), result 1.975;
- B7, margin: =B6*B5, result 2.167;
- B8, lower limit: =B3-B7, result 128.96;
- B9, upper limit: =B3+B7, result 133.30.

9. Check the margin with the dedicated function: =CONFIDENCE.T(0.05, B4, B2) should also give 2.167. Also compute =CONFIDENCE.NORM(0.05, B4, B2), which uses z = 1.96 and gives 2.151. The difference is small at n = 160; in step 16 you will see it grow.
10. For the 99% CI, change 0.05 to 0.01 in B6: t = 2.607, margin 2.861, interval 128.27-133.99 mmHg.

L4.4 Intervals for Values and Intervals for the Mean

11. In the same sheet, build Table L4.1 with formulas (k in column A: 1, 2, 3):

- lower limit for values: =$B$3-A12*$B$4; upper: =$B$3+A12*$B$4;
- lower limit for the mean: =$B$3-A12*$B$5; upper: =$B$3+A12*$B$5.

Table L4.1. Self-check values for SBP (mg2026).

Confidence	Individual values ($m \pm k \times s$)	Mean ($m \pm k \times \mathrm{SE}$)
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

Confidence	Individual values ($m \pm k \times s$)	Mean ($m \pm k \times \text{SE}$)
99.7% ($k = 3$)	89.5-172.8 mmHg	127.84-134.42 mmHg

12. Count how many patients actually lie in each interval for values, with =COUNTIFS(Date!M2:M161, ">="&C12, Date!M2:M161, "<="&D12). Expected results: 112 patients (70.0%), 150 (93.8%) and 160 (100%). Then count how many patients lie in the 95.4% interval for the mean: only 25 patients. Write in your notebook, in your own words, why the second number is so small and why this is not a problem.
13. Repeat the table for age (m = 56.29; s = 8.77): the 95.4% interval for values is 38.75-73.83 years, and for the mean 54.90-57.68 years. This is exactly the type of exam problem: "estimate with 95.4% confidence the values of age" versus "estimate with 95.4% confidence the mean age in the population".

L4.5 The Confidence Interval for a Proportion

14. The proportion of patients with OGTT-confirmed diabetes. In the "IC_prop" sheet:

- B2, k: =SUM(Date!X2:X161), result 29;
- B3, n: =COUNT(Date!X2:X161), result 160;
- B4, p: =B2/B3, result 0.181;
- B5, z: =NORM.S.INV(0.975), result 1.96;
- B6, Wald SE: =SQRT(B4*(1-B4)/B3), result 0.0305;
- B7 and B8, the Wald limits: =B4-B5*B6 and =B4+B5*B6, results 0.122 and 0.241.

15. The Wilson interval, in three cells:

- B10, the center: =(B4+B5^2/(2*B3))/(1+B5^2/B3), result 0.189;
- B11, the margin: =B5/(1+B5^2/B3)*SQRT(B4*(1-B4)/B3+B5^2/(4*B3^2)), result 0.059;
- B12 and B13, the limits: =B10-B11 and =B10+B11, results 0.129 and 0.248.

Wilson 95% CI: 12.9%-24.8%, compared with Wald 12.2%-24.1%.

16. The exact (Clopper-Pearson) interval: in JASP, Frequencies > Binomial Test, variable diabet_ogtt (set as nominal), check "Confidence interval"; for level 1 the result is 12.5%-25.0%. In Epi Info, Classic Analysis, the command FREQ diabet_ogtt displays the same exact limits. In MedCalc, the calculator for one proportion (Tests > One proportion, or the online calculators) also gives the exact interval.
17. Sensitivity test: in B2 and B3 type 2 and 20 manually (two complications in 20 patients). The Wald interval becomes -3.1% to 23.1%, with an impossible negative limit; the Wilson interval is 2.8%-30.1%. Then type 0 and 50: Wald gives 0-0%, Wilson 0-7.1%, and the rule of three gives 3/50 = 6%. Restore the original formulas at the end.

L4.6 Simulation: How the Interval Depends on Sample Size

In this part, we treat the 160 patients in mg2026 as a "population" with a known mean, 131.13 mmHg, and draw random samples from it, exactly as in Figure 4.2 of the course.

18. In a new sheet, "Simulare" (simulation), write the header "esantion" (sample) in A1. In A2 write the formula =INDEX(Date!$M$2:$M$161, RANDBETWEEN(1, 160)) and copy it down to A161. Each cell picks one of the 160 patients at random (with replacement, so the same patient may appear several times). Each time the sheet is modified, the values change; to force a new draw, press Delete on an empty cell or use File > Settings > Calculation > Recalculation: "On change and every minute".
19. Compute the intervals for the first 10, the first 40 and all 160 values in column A:

- n = 10, in D2: =AVERAGE(A2:A11); E2: =CONFIDENCE.T(0.05, STDEV(A2:A11), 10); the limits =D2-E2 and =D2+E2; the width =2*E2;
- n = 40: the same formulas on A2:A41, with 40;
- n = 160: on A2:A161, with 160.

In the next column write =IF(AND(131.13>=F2, 131.13<=G2), "YES", "NO"), which tells you whether the interval contains the "population" mean.

20. Copy the results (Edit > Paste special > Values only) into a table of repetitions, then trigger a new draw and repeat. Do at least 20 repetitions (40 in teams of two students). For each n, compute at the end the mean of the widths and the number of intervals that contain 131.13.
21. What you should observe. In 10,000 repetitions done by computer, the mean width of the 95% CI was 19.3 mmHg for n = 10 (with large variation, from about 12 to 27 mmHg between samples), 8.8 mmHg for n = 40 (7.2-10.4) and 4.3 mmHg for n = 160 (3.9-4.7). The proportion of intervals that contained the population mean was 95.3%, 94.9% and 94.9%, respectively, that is, very close to 95% for all three sizes. With 20 repetitions, you will probably have 18-20 "YES" results for each n; if a "NO" appears, it is not an error but exactly the promised 5%. Also compare with the fixed samples in the course (Table 4.3): n = 10, CI 116.38-143.62 mmHg (width 27.2); n = 40, 123.08-133.67 (10.6); n = 160, 128.96-133.30 (4.3).
22. The conclusion, which you write in your notebook in 3-4 sentences: the width of the interval decreases approximately with $\sqrt{n}$ (from 10 to 40 patients it is halved, and likewise from 40 to 160); the 95% confidence level does not depend on n but refers to the proportion of intervals that contain the parameter; in small samples, not only the width but also the position of the interval varies considerably from one repetition to the next.
23. The version with a normal population, optional. Replace the formula in A2 with =NORM.INV(RAND(), 131, 14), which generates values from a normal population with mean 131 and SD 14 mmHg, and repeat steps 19-20, checking the coverage of 131. This is exactly the simulation in Figure 4.2.

L4.7 Sample Size Calculation

24. For a mean, by hand in Sheets: =ROUNDUP((1.96*14/3)^2, 0) gives 84 patients (margin ±3 mmHg, σ = 14 mmHg). Try margins of 2 and 5 mmHg: 189 and 31. Some programs use the t distribution instead of z and give slightly more (for ±3 mmHg, 87); the difference is of no practical importance.
25. For a proportion: =ROUNDUP(1.96^2*0.2*0.8/0.05^2, 0) gives 246 (anticipated prevalence 20%, margin ±5 percentage points). With p = 0.5, 385.
26. Epi Info, StatCalc > Population survey. Enter: Population size 1000, Expected frequency 20%, Acceptable margin of error 5%, Design effect 1, Clusters 1. At the 95% confidence level, the program displays 198 subjects: this is the sample of 246 corrected for a finite population of 1000 (246 / (1 + 245 / 1000) = 197.6). Change Population size to 999999: the result becomes 246. Do the same calculation with Population size 500 (expected result: 166).
27. G*Power, comparison of two means: Test family "t tests", Statistical test "Means: Difference between two independent means (two groups)", Type of power analysis "A priori: Compute required sample size". Enter 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. Result: 64 per group, 128 in total. Repeat with d = 0.3 (176 per group) and with d = 0.8 (26 per group). The

"Determine" button lets you compute d from the means and standard deviation: for a difference of 5 mmHg and an SD of 14 mmHg, d = 0.357, and the required sample size is 125 per group.

28. MedCalc, Sampling menu: choose the calculation for comparing two means, enter α = 0.05, β = 0.20 (power 80%), difference 5 and SD 14, and compare with G*Power. Then use the calculation for a single proportion (prevalence 20%, precision ±5%).
29. Adjustment for losses: if you anticipate 20% losses, =ROUNDUP(64/(1-0.2), 0) gives 80 per group. Check why the formula =64*1.2 (result 77) is wrong: compute how many patients remain out of 77 after losses of 20%.
30. The power of an existing sample: in 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. The result is around 61%: the comparison of SBP between the sexes in mg2026 had modest power for a difference of 5 mmHg. Remember: this calculation is done for a clinically relevant difference set in advance, not for the observed difference.

L4.8 Reporting an Estimate

31. For SBP, blood glucose and the prevalence of diabetes, write one article-style reporting sentence each. Model: "Mean SBP at enrollment was 131.13 mmHg (SD 13.88; 95% CI for the mean 128.96-133.30 mmHg; n = 160). The prevalence of OGTT-confirmed diabetes was 18.1% (29/160; Wilson 95% CI 12.9-24.8%)." Check that each sentence contains: the estimate, the unit of measurement, the interval with its level, the method (for proportions) and n.

Additional Exercises

Exercises 1-3 are solved on the discipline's classic files; the numerical results are not given here.

1. The BD_Studenti file. (a) Estimate, with 68.3% confidence, the values of the students' height. (b) Estimate, with 95.4% confidence, the mean ideal height in the student population. (c) Estimate, with 99.7% confidence, the students' BMI values and, separately, the mean BMI in the population. Comment on the difference between the last two intervals. (d) Compute the 95% CI with the t distribution for the mean hours of sleep, separately for girls and boys, and say whether the intervals overlap.
2. The Probleme BPOC file (COPD problems). Compute the 95% CI for the mean baseline heart rate (FC1) and for the mean of the difference FC1 - FC2. Compute the Wilson 95% CI for the proportion of patients with dyspnea at enrollment (DISPN = 1), for the whole group and separately for patients under/over 50 years.
3. On BD_Studenti, draw random samples of 10 and of 30 students in Google Sheets (the method from step 18) and compare the width of the 95% CI for mean height with that obtained on the whole file.
4. Problem with data in the statement (with solution). A pilot study in 15 hypertensive patients finds a mean decrease in SBP of 9 mmHg, with an SD of 12 mmHg. (a) Compute the 95% CI for the mean decrease. (b) How many patients would be needed to estimate the mean decrease with a margin of ±2 mmHg?

Solution. (a) SE $= \frac{12}{\sqrt{15}} = \frac{12}{3.873} = 3.098$ mmHg; df = 14, t = 2.145; margin = 2.145 x 3.098 = 6.65 mmHg; 95% CI: 9 ± 6.65, that is, 2.35-15.65 mmHg. The interval excludes 0, but it is very wide: the true decrease could be minor or substantial. $(b)\ n = \left(\frac{1.96\times12}{2}\right)^2 = (11.76)^2 = 138.3$, hence 139 patients (with the t distribution, a program will give 1-2 more patients). Note that the SD from

the pilot is itself imprecise with 15 patients; a prudent calculation would use a somewhat larger value.

5. Problem with data in the statement (with solution). In a survey in a family practice, 45 of 150 patients vaccinated against influenza report a local reaction. Compute the Wald and Wilson 95% CIs.

Solution. $p = \frac{45}{150} = 0.30$; SE $= \sqrt{\frac{0.30 \times 0.70}{150}} = 0.0374$; Wald: 0.30 ± 0.073, that is, 22.7%-37.3%. Wilson: $\frac{z^2}{n} = 0.0256$; center = (0.30 + 0.0128) / 1.0256 = 0.305; margin = $\left(\frac{1.96}{1.0256}\right) \times \sqrt{0.0014 + 0.0000427} = 1.911 \times 0.0380 = 0.0726$; CI 23.2%-37.8%. For p = 30% and n = 150, the two methods give very similar results; the difference becomes important only for extreme proportions or small n.

What to submit and indicative grading scheme

Upload to Moodle, named LP4_Surname_FirstName, a file containing:

- the 90%, 95% and 99% confidence intervals for mean SBP, with the screenshot from JASP or MedCalc and the calculation table from Sheets (2 points);
- Table L4.1 completed for SBP and for age, with the number of patients in each interval and the comment from step 12 (2 points);
- the 95% CI for the proportion with diabetes by the Wald, Wilson and exact methods, plus the sensitivity test from step 17 (1.5 points);
- the simulation table (at least 20 repetitions for n = 10, 40, 160), with the mean of the widths, the number of intervals containing 131.13 and the conclusion from step 22 (2.5 points);
- the sample size calculations from L4.7, with the screenshots from G*Power or MedCalc and Epi Info (1.5 points);
- the reporting sentences from step 31 (0.5 points).

Common mistakes: using the standard deviation instead of the standard error for the interval of the mean (or vice versa); interpreting the 95% CI of the mean as the interval containing 95% of patients; rounding the sample size down; adjusting for losses by multiplying by 1.2.

Offline variant

JASP, MedCalc, G*Power and Epi Info run locally, so steps L4.1, L4.2, L4.5 (the JASP and Epi Info versions) and L4.7 can be done without the internet. For L4.3, L4.4 and the simulation, a locally installed spreadsheet program (Excel or LibreOffice Calc) is used, with the same functions: AVERAGE, STDEV, T.INV.2T, CONFIDENCE.T, CONFIDENCE.NORM, NORM.S.INV, NORM.INV, RAND, RANDBETWEEN, INDEX, COUNTIFS. In Excel, recalculation of the random draws is forced with the F9 key. If no computer is available, the intervals are calculated by hand, with the critical values from Tables E.2 and E.3 of Appendix E, using the statistics in Table L4.1.

Practical Session 5. Statistical Hypothesis Testing: Formulating Hypotheses, Errors and Decisions

Objective

This session turns the reasoning of Chapter 5 into practical steps: from a clinical question to the hypotheses H0 and H1, from the data to the test statistic and the p value, and from the p value to a correctly worded decision. We work with the simplest test, the one-sample t-test, because there every step can be checked by hand. The second part of the session is a simulation: you will generate in Google Sheets hundreds of "studies" in which the null hypothesis is true and see with your own eyes that approximately 5% of them come out "significant". The simulation will help you understand, better than any definition, what α means and why a nonsignificant result does not prove the absence of an effect.

By the end, the student should be able to:

- formulate the statistical hypotheses H0 and H1 (in words and mathematically) for a given clinical question and justify the choice of a two-sided or one-sided test;
- apply the one-sample t-test in JASP, MedCalc and Google Sheets and read in the output the t statistic, the degrees of freedom, the p value, the confidence interval and the effect size;
- compute the t statistic by hand and obtain the p value from the table and with the functions T.DIST.2T and T.INV.2T;
- simulate the type I error and the power of the test in Google Sheets and interpret the proportion of significant results;
- compute, with a simple formula and with a program, the number of patients needed for a given power;
- write up the solution to a problem in the a-g format used for the discipline's assignments.

Tools

Software: JASP 0.98.1 [47]; MedCalc 24.0 [50]; Google Sheets, with English (United States) regional settings (the argument separator is ",", and the decimal separator is the point) [43]; optionally G*Power for power calculation [77]. Files: mg2026_curat.csv (160 patients) from the discipline's Moodle platform [52]; for the additional exercises, the classic files Date_HbA1c, BD_Studenti and BPOC. Approximate time: 100 minutes (20 minutes on hypotheses, 35 minutes on the t-test in the three programs, 25 minutes on the simulation, 20 minutes on power and the assignment).

The columns of mg2026_curat.csv, imported into Google Sheets with the header in row 1 and the data in rows 2-161, are: 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. Check the order after import; the formulas below assume this layout.

Procedure

L5.1 Formulating the Hypotheses

Before opening any program, write down the hypotheses on paper. This is the rule that prevents most mistakes in the practical exam: the test is chosen according to the question, not according to what "comes out" in the program.

1. Read the question and identify the population, the variable and the comparison. Example: "Does the mean SBP of patients at cardiometabolic risk differ from 125 mmHg?" Population: adults at cardiometabolic risk in family practices. Variable: SBP at enrollment (numerical, mmHg). Comparison: the population mean versus a fixed value, 125 mmHg.
2. Write the research hypothesis in clinical language: "the mean SBP of patients at cardiometabolic risk differs from 125 mmHg".
3. Write H0 in words and mathematically: "the mean SBP in the population is 125 mmHg", $\mu = 125$. Write H1: "the mean SBP in the population differs from 125 mmHg", $\mu \neq 125$.
4. Decide between two-sided and one-sided. By default, two-sided; a one-sided test requires a written justification before the analysis (for example, a non-inferiority study).
5. Set $\alpha = 0.05$.

Classroom exercise (10 minutes, in pairs). Write the hypotheses H0 and H1, in words and mathematically, for the questions below; state the parameter (μ, μD, π, ρ) and the type of test (two-sided or one-sided).

- (a) Does the mean fasting plasma glucose in mg2026 differ from 100 mg/dL?
- (b) Does the mean HbA1c differ from 5.7%?
- (c) Does fasting plasma glucose change between enrollment and 6 months in the same patients?
- (d) Does SBP differ between women and men?
- (e) Does the decrease in blood glucose differ among the three arms of the study?
- (f) Does the proportion of patients with OGTT-confirmed diabetes differ between obese and non-obese patients?
- (g) Are BMI and SBP correlated?
- (h) A new glucose meter does not underestimate blood glucose by more than 5 mg/dL compared with the laboratory.

Check. For (a): H0: $\mu = 100$; H1: $\mu \neq 100$. For (c): H0: $\mu D = 0$, where D = glicemie0 - glicemie6; H1: $\mu D \neq 0$. For (e): H0: $\mu A = \mu B = \mu C$; H1: at least one mean differs. For (g): H0: $\rho = 0$; H1: $\rho \neq 0$. For (h): non-inferiority hypotheses, one-sided, with the 5 mg/dL margin set in advance. Hypotheses are always written in terms of population parameters, never in terms of the sample mean.

L5.2 The One-Sample t-Test in JASP

1. Open JASP and load the file: main menu (the three lines at the top left) > Open > Computer > Browse > mg2026_curat.csv. Check in the data view that tas0 and glicemie0 appear as continuous variables (the ruler icon); if not, click the icon and choose Scale.
2. From the module bar choose T-Tests > One Sample T-Test (in the Classical section).
3. Move tas0 into the Variables box. In the Test value box write 125.
4. Under Tests check Student and Wilcoxon signed-rank. Under Alt. Hypothesis leave "≠ Test value" (two-sided).
5. Under Additional Statistics check: Location parameter (the mean difference from the tested value), Confidence interval (95%), Effect size (Cohen's d), Descriptives and, if you want a graph, Descriptives plots.
6. Under Assumption Checks check Normality (Shapiro-Wilk) and, optionally, Q-Q plot.

What you should obtain for tas0 versus 125: mean 131.13 mmHg, SD 13.88, n = 160; Student's test t = 5.59, df = 159, p < 0.001; mean difference (location parameter) 6.13 mmHg, with a 95% CI of 3.96-8.30 mmHg; Cohen's d = 0.44; Shapiro-Wilk p = 0.303; Wilcoxon test p < 0.001. Note that JASP reports the confidence interval for the difference from 125, not for the mean; by adding 125 you obtain the interval for the mean, 128.96-133.30 mmHg, known from Chapter 4.

7. Change Test value to 130 and read the output again. You should obtain t = 1.03, p = 0.304, and a 95% CI of the difference from -1.04 to 3.30 mmHg (which contains zero). Write the correct decision: "H0 is not rejected", not "H0 is accepted".
8. Repeat the analysis for glicemie0, with Test value 100. You should obtain: mean 105.44 mg/dL, SD 15.76; t = 4.36, df = 159, p < 0.001; mean difference 5.44 mg/dL (95% CI 2.98-7.90); d = 0.35; Shapiro-Wilk p = 0.037 (slight departure from normality); Wilcoxon p < 0.001. Discuss with your colleague: is the Wilcoxon test necessary here? With n = 160 and a small skewness (0.37), the t-test is robust, and the two tests lead to the same conclusion.
9. Repeat for hba1c0 with Test value 5.7. Expected result: t = -3.48, df = 159, p < 0.001; the mean is below 5.7%.

L5.3 The Same Test in MedCalc

1. Open MedCalc and import the file: File > Import > CSV (or File > Open, with the CSV file type). Check the decimal separator in the import options, because the file uses the decimal point.
2. From the Statistics menu choose T-tests > One sample t-test.
3. Under Variable select tas0, and under Test mean (the value to compare with) write 125. Confirm with OK.
4. Read in the results window: the mean, the 95% CI for the mean, the difference from the tested value, the t statistic, the degrees of freedom and the p value. The results should match those from JASP (t = 5.59; p < 0.0001).
5. MedCalc also allows direct calculation from summary data, without a file: in the Tests menu look for the test for one mean (Test for one mean), enter the mean 131.13, SD 13.88, n = 160 and the value 125. The result should be the same. This is a useful method for checking the figures in an article.

L5.4 The Same Test in Google Sheets, Step by Step

Google Sheets has no dedicated function for the one-sample t-test, but you can build it from simple functions; that is precisely why it is the best exercise for understanding. Work in a new sheet (for example, "LP5") and write the labels in column A and the formulas in column B.

1. B1, mean: =AVERAGE(Date!M2:M161) (if the data are in the "Date" sheet). Result: 131.13.
2. B2, standard deviation: =STDEV(Date!M2:M161). Result: 13.88.

3. B3, n: =COUNT(Date!M2:M161). Result: 160.
4. B4, the tested value μ0: 125.
5. B5, standard error: =B2/SQRT(B3). Result: 1.097.
6. B6, t statistic: =(B1-B4)/B5. Result: 5.59.
7. B7, degrees of freedom: =B3-1. Result: 159.
8. B8, two-sided p value: =T.DIST.2T(ABS(B6), B7). Result: approximately 0.0000001, that is, $p < 0.001$. The function requires a positive value, which is why we use ABS.
9. B9, the critical value for two-sided $\alpha = 0.05$: =T.INV.2T(0.05, B7). Result: 1.975.
10. B10 and B11, the 95% CI limits of the mean: =B1-B9*B5 and =B1+B9*B5. Result: 128.96 and 133.30.
11. B12, the automatic decision: =IF(B8<0.05, "H0 is rejected", "H0 is not rejected").
12. B13, the effect size d: =(B1-B4)/B2. Result: 0.44.

An elegant check with the T.TEST function. The one-sample t-test is equivalent to a paired t-test between the data and a constant column equal to μ0. Write 125 in cells Date!AD2:AD161 (write 125 in AD2 and drag down) and compute =T.TEST(Date!M2:M161, Date!AD2:AD161, 2, 1). The result should be identical to that in B8. The third argument (2) requests the two-sided test, and the fourth (1) requests the paired test.

Now change the value in B4 to 130, then to 140, and watch how t, p, the decision and the position of μ0 relative to the confidence interval change simultaneously. Fill in the table below; it reproduces, in your own sheet, the equivalence of the three routes to a decision from Chapter 5.

μ0	t	p	μ0 in the 95% CI?	Decision
125	5,59	< 0,001	no	H0 is rejected
130				
132				
140				

L5.5 Computing the p Value by Hand from the t Statistic

In the written exam you do not have a statistical calculator, but you must be able to bracket the p value between thresholds using the table of critical values (Table 5.6 in the chapter).

1. Compute t and the degrees of freedom.
2. Find the row with the degrees of freedom in the table (or the nearest smaller row, to be on the safe side).
3. Compare |t| with the critical values for 0.05, 0.01 and 0.001. If |t| is smaller than the critical value for 0.05, then $p > 0.05$; if it lies between the values for 0.05 and 0.01, then $0.01 < p < 0.05$; and so on.
4. Then check in Google Sheets with =T.DIST.2T(t, df).

Bracketing exercises (solve first with the table, then check in Sheets): $t = 2.10$ with $df = 20$ (the exact result should be 0.049); $t = 0.83$ with $df = 9$ (0.43); $t = 3.10$ with $df = 9$ (0.013); $t = 1.80$ with $df = 25$ (two-sided 0.084; one-sided =T.DIST.RT(1.8, 25) = 0.042); $t = 4.36$ with $df = 159$ (< 0.001). For the next-to-last example, discuss: why is it dangerous to decide, after seeing the data, that the test "should" be one-sided?

Exercise 5 of the session: compute t by hand for a small sample. Take the first 10 patients in the file (rows 2-11) and test whether the mean SBP differs from 125 mmHg. Compute the mean, SD, SE, t and df, and bracket p using the table. Compare with the result on 160 patients and

explain, in two sentences, why the conclusion may be different. (The result depends on the patients chosen; in the chapter, a random sample of 10 patients gave t = 0.83, p = 0.43.)

L5.6 Simulating the Type I Error in Google Sheets

The aim is to see what α = 0.05 means. We will generate 200 "studies" in which two groups of 20 patients each come from the same population (SBP with mean 130 mmHg and SD 14 mmHg). By construction, H0 is true in every study. Any significant result is therefore a type I error.

1. Create a new sheet, "Simulare". In cell A1 write "studiul" (study), in B1:U1 the labels for group 1 (g1_1... g1_20), in V1:AO1 the labels for group 2 (g2_1... g2_20), and in AP1 "p".
2. In A2 write 1, in A3 =A2+1, and drag down to A201 (200 studies).
3. In B2 write =NORM.INV(RAND(), 130, 14) and copy the formula into the whole block B2:AO201. Each cell is now a random SBP value from a normal population with mean 130 and SD 14.
4. In AP2 write =T.TEST(B2:U2, V2:AO2, 2, 2) and drag down to AP201. Each row receives the p value of an independent-samples t-test between the two groups of the same study.
5. In AR1 write "proportion p < 0.05", and in AR2 =COUNTIF(AP2:AP201, "<0.05")/200. Format as a percentage.
6. Modify any empty cell (for example, type something in AT1 and delete it) to force the RAND function to recalculate. Write down the proportion obtained 5 times.

What you should obtain: proportions around 5%, usually between 2% and 8% (with 200 studies, the random variation is about ±3 percentage points). In the simulation in the chapter (1000 studies), the proportion was 4.9%. Also build a histogram of the p values in column AP (Insert > Chart > Histogram): it should be approximately flat, with p values spread uniformly between 0 and 1. The conclusion to write in your notebook: when H0 is true, the test declares approximately α of the studies "significant"; α is the false alarm rate we accept.

7. Simulating power. Change the formula in group 2 (V2:AO201) to =NORM.INV(RAND(), 137, 14), that is, a real difference of 7 mmHg between the groups (d = 0.5). Now H0 is false, and the proportion of significant results estimates the power. What you should obtain: around 34% (the theoretical value for d = 0.5 and 20 patients per group is 33.8%). In other words, two out of three such studies would miss a real effect of 7 mmHg. This is the type II error.
8. Also try a large difference, with the group 2 mean equal to 144 mmHg (d = 1): the power should rise to approximately 87%.
9. Multiple comparisons. Set the group 2 mean back to 130 (H0 true). Think of each group of 10 rows as a single "article" with 10 tests. How many of the 20 "articles" have at least one p < 0.05? Theoretically, 1 - 0.95^10 = 40% of them. Check by counting, or with =COUNTIF on blocks of 10 rows.

L5.7 Power and Sample Size Calculation

1. The approximate formula for the number of patients per group when comparing two means, in Google Sheets: =2*(NORM.S.INV(0.975)+NORM.S.INV(0.8))^2*σ^2/Δ^2. For σ = 14 mmHg and Δ = 7 mmHg (d = 0.5): the result is 62.8, that is, approximately 63 per group; the exact calculation, with the t distribution, gives 64 per group (the value used in Chapter 5). For Δ = 5 mmHg: approximately 123 per group. For Δ = 4.2 mmHg (d = 0.3): approximately 175 per group (exactly 176).
2. The approximate formula for power, given n per group: =NORM.S.DIST(d*SQRT(n/2)-NORM.S.INV(0.975), TRUE). For d = 0.5 and n = 20: approximately 0.35, very close to the simulation. For d = 0.5 and n = 64: approximately 0.81.

3. In JASP, power is calculated in the Power module (if it does not appear in the module bar, activate it with the "+" button on the right of the bar): choose the independent-samples t-test, enter the effect size, α and the desired power, and read n.
4. In MedCalc, the Sampling menu offers sample size calculation for comparing two means (enter the difference, SD, α and β). Alternatively, G*Power: t tests > Means: Difference between two independent means (two groups) > A priori.
5. Fill in the table and comment on how n changes when power increases or α decreases.

Δ (mmHg)	σ	d	α	Power	n per group (formula)
7	14	0,5	0,05	80%	approximately 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%	

Self-check values (approximate formula, rounded up): 85 per group for 90% power; 94 per group for α = 0.01; 123-124 per group for Δ = 5 mmHg (123.07 before rounding); 31 per group for Δ = 10 mmHg. The exact calculation with the t distribution gives values 1-2 patients larger (for example, 86 and 96 for the first two rows).

L5.8 The Assignment in the a-g Format: A Complete Model

Problem. Is the mean fasting plasma glucose at enrollment in mg2026 different from the threshold of 100 mg/dL?

(a) Research hypothesis: the mean fasting plasma glucose in the population of adults at cardiometabolic risk in family practices differs from the 100 mg/dL threshold for impaired fasting glucose.

(b) Statistical hypotheses. H0: the mean fasting plasma glucose in the population is 100 mg/dL; μ = 100. H1: the mean differs from 100 mg/dL; μ ≠ 100. Two-sided test, α = 0.05.

(c) Method: one numerical variable, one mean compared with a fixed value, σ unknown; the one-sample t-test. Conditions: independent observations; normality checked with Shapiro-Wilk (p = 0.037, a slight departure, acceptable at n = 160 thanks to the central limit theorem); additional check with the Wilcoxon test.

(d) Procedure: JASP, T-Tests > One Sample T-Test, variable glicemie0, Test value 100, with Student and Wilcoxon, Location parameter, Effect size, Descriptives, Normality; check in Google Sheets with the formulas from L5.4.

(e) Result: $m = 105.44$ mg/dL, s = 15.76, n = 160; t(159) = 4.36; p < 0.001; mean difference 5.44 mg/dL (95% CI 2.98-7.90); d = 0.35; Wilcoxon p < 0.001. Graph: the histogram of blood glucose with a vertical line at 100 mg/dL, or the JASP descriptives plot with the mean and 95% CI.

(f) Statistical conclusion: p < 0.001 < 0.05, H0 is rejected; the difference is highly significant.

(g) Research conclusion: the mean fasting plasma glucose of patients at cardiometabolic risk is higher than the 100 mg/dL threshold, by approximately 3-8 mg/dL; the typical patient in the study lies in the range of impaired fasting glucose. The effect is small to medium (d = 0.35), and the result describes the population studied, not the general population.

Additional Exercises

Solve each exercise in the a-g format, in JASP and in Google Sheets. For the files on Moodle, the results are not given; check each other's work with your colleagues and compare the values obtained in the two programs.

1. The Date_HbA1c file. Is the mean HbA1c in the first year of the data different from the therapeutic target of 7%? Formulate the hypotheses, check normality and apply the appropriate test. Repeat for the last year. Comment on the difference between statistical significance and reaching the target at the individual level.
2. The BD_Studenti file. Does the mean height of the female students differ from a reference value chosen by the instructor (for example, the national mean for young women, announced in the session)? And that of the male students? Which of the two tests has more power, and why?
3. The BPOC file (COPD data). Is the patients' mean baseline heart rate different from 80 beats/min? Check normality; if appropriate, also report the Wilcoxon test.
4. With data in the statement (with complete solution). In 16 treated patients with hypothyroidism, TSH has a mean of 3.1 mIU/L and an SD of 1.2 mIU/L. The laboratory target is 2.5 mIU/L. (a) H0: μ = 2.5; H1: μ ≠ 2.5. (b) SE $= \frac{1.2}{\sqrt{16}} = 0.30$. (c) $t = \frac{3.1-2.5}{0.30} = 2.00$, df = 15. (d) critical t (15, 0.05) = 2.131; 2.00 < 2.131, so p > 0.05 (exactly, T.DIST.2T(2, 15) = 0.064). (e) 95% CI for the mean: $3.1 \pm 2.131 \times 0.30 = 2.46 - \frac{3.74 m_{\text{UI}}}{L}$, an interval that contains 2.5. (f) H0 is not rejected. (g) The data do not allow the conclusion that mean TSH differs from the target; the interval, however, is wide, so the result is inconclusive, and with a larger sample the same mean difference of 0.6 mIU/L could become significant.
5. With data in the statement (with complete solution). A planned study aims to detect a difference in fasting plasma glucose of 8 mg/dL between two groups, with $\sigma = 16$ mg/dL, two-sided α = 0.05 and 80% power. $d = \frac{8}{16} = 0.5$; $n = \frac{2\times(1.96+0.84)^2\times16^2}{8^2} = \frac{2\times7.84\times256}{64} = 62.7$, that is, 63 patients per group (64 with the exact calculation). If 15% loss to follow-up is expected, 64 / 0.85 = 75.3, that is, 76 patients per group are recruited.

What to submit and indicative grading scheme

Submit on Moodle a single file (PDF or document) containing: the table with the hypotheses for the 8 questions in L5.1 (1 point); the result of the t-test for tas0 versus 125 and versus 130 from JASP, with a screenshot of the table and the correctly worded decision (2 points); the Google Sheets sheet from L5.4, with the formulas visible and the table completed for the four μ0 values (2 points); the result of the simulation from L5.6 (the 5 proportions, the histogram of the p values, the estimated power for the 7 mmHg difference) with a 3-4 sentence interpretation (2 points); the completed power table from L5.7 (1 point); an assignment in the a-g format on one of the additional problems 1-3 (2 points). Total: 10 points. Points are deducted for: hypotheses written with the sample mean, the expression "H0 is accepted", p reported as 0.000, a missing confidence interval and a missing research conclusion in words.

Offline variant

If the network or the platform is down, the session can be done entirely with a pocket calculator and the table of critical t values from Chapter 5, plus a locally installed JASP (the program does not need the internet after installation). The mg2026_curat.csv file is copied from the instructor's USB stick. Instead of Google Sheets, any locally installed spreadsheet program (LibreOffice Calc or Excel) is used, with the same functions: T.DIST.2T, T.INV.2T, NORM.INV, RAND, COUNTIF (in the Romanian-language LibreOffice, function names may appear translated, and the argument separator is ";", as with the Romania setting). The simulation in L5.6 can also be done "manually", in a group: each student rolls two dice 20 times, records the sum and compares their mean with that of the student at the next desk using a t-test calculated by hand; across the

group, the proportion of "significant" comparisons will be small, and the discussion about the type I error remains the same.

Practical Session 6. Parametric Tests: The t-Test and ANOVA

Objective

In this session you apply the parametric tests from Chapter 6 to the mg2026 dataset: the paired-samples t-test, the independent-samples t-test (Student and Welch), one-way ANOVA with post hoc tests, and repeated measures ANOVA. We work in parallel in JASP, MedCalc and Google Sheets. The aim is not to press buttons, but to recognize the design, to check the conditions of application before reading the p value, and to be able to reproduce in Google Sheets, with simple formulas, a result obtained in specialized software. Each step gives the figures you should obtain, so that you can check your work yourselves.

By the end, the student should be able to:

- choose among the paired t-test, the independent-samples t-test and ANOVA on the basis of the design and formulate the hypotheses in words and mathematically;
- run the paired t-test and the independent-samples t-test in JASP and MedCalc, checking normality and equality of variances, and interpret the difference between the Student and Welch versions;
- run one-way ANOVA in JASP, with Levene's test, the Q-Q plot of the residuals, the Tukey and Games-Howell post hoc tests and the plot of the means;
- build the ANOVA table in Google Sheets with the functions AVERAGEIF, DEVSQ, FILTER and F.DIST.RT and use T.TEST with types 1, 2 and 3;
- run and interpret repeated measures ANOVA, including Mauchly's test and the Greenhouse-Geisser correction;
- write up the results in the a-g format, with the effect size and the confidence interval.

Tools

Software: JASP 0.98.1 [47]; MedCalc 24.0 [50]; Google Sheets with English (United States) regional settings ("," as the argument separator, decimal point) [43]. Files from Moodle [52]: mg2026_curat.csv; for the additional exercises, BD_lp3 (blood glucose at 0, 3 and 6 months, insurance status), BPOC (COPD data: heart rate, FEV1, bronchial patency index before and after treatment, drug received) and BD_Studenti (height, ideal height, sex, study group). Approximate time: 100 minutes (10 minutes preparing the data, 25 minutes on the paired and independent tests, 35 minutes on ANOVA, 15 minutes on repeated measures, 15 minutes on Google Sheets and the assignment).

The columns of mg2026_curat.csv, in the order in the file: 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.

Procedure

L6.1 Preparing the Data: The Change Variables

The main question of the mg2026 study concerns the decrease in blood glucose at 6 months by arm. For ANOVA we need a new variable, calculated for each patient.

1. In JASP, in the data view, click the "+" sign after the last column (Add computed column). Name: dglic; type: Scale; method: R formula or the visual builder. Formula: glicemie0 - glicemie6. Confirm with Compute column. The mean of the new column should be 10.19 mg/dL.
2. In the same way, create dgr = greutate0 - greutate6 (weight loss, kg). Mean: 2.31 kg.
3. In Google Sheets, in the data sheet, write "dglic" in AC1, =S2-U2 in AC2 and drag down to AC161; write "dgr" in AD1, =J2-Y2 in AD2 and drag down. Check: =AVERAGE(AC2:AC161) gives 10.19; =AVERAGE(AD2:AD161) gives 2.31.
4. Note the sign convention: positive values mean a decrease. A common mistake in the exam is to interpret the sign of the difference the wrong way round.

L6.2 The Paired-Samples t-Test in JASP

The question: does fasting plasma glucose change between enrollment and 6 months in the same patients? H0: μD = 0; H1: μD ≠ 0, where D = glicemie0 - glicemie6.

1. T-Tests > Paired Samples T-Test. Move the pair glicemie0 and glicemie6 into the Variable pairs box (the first variable on the left, the second on the right).
2. Under Tests check Student and Wilcoxon signed-rank; alternative hypothesis "Measure 1 ≠ Measure 2".
3. Under Additional Statistics check Location parameter, Confidence interval, Effect size, Descriptives and Descriptives plots (or the "raincloud" plot, if it is available in your version).
4. Under Assumption Checks check Normality (Shapiro-Wilk). JASP tests the normality of the differences, which is exactly the right condition.

What you should obtain: mean blood glucose 105.4 mg/dL at enrollment and 95.2 mg/dL at 6 months; t = 11.14, df = 159, p < 0.001; mean difference 10.19 mg/dL, 95% CI 8.38-11.99; Cohen's d (for paired data, dz) = 0.88; Shapiro-Wilk on the differences p = 0.725; Wilcoxon test p < 0.001.

5. Add the pairs glicemie0-glicemie3 and glicemie3-glicemie6 to the same analysis. Expected values: 0-3 months, difference 6.08 mg/dL (95% CI 4.61-7.54), t = 8.22, dz = 0.65, Shapiro-Wilk p = 0.133; 3-6 months, difference 4.11 mg/dL (95% CI 2.10-6.13), t = 4.03, p < 0.001, dz = 0.32, Shapiro-Wilk p = 0.868. Caution: three paired tests on the same data are multiple comparisons; the correct analysis for three time points is repeated measures ANOVA (L6.5).
6. The pair tas0-tas6: difference 4.78 mmHg (95% CI 3.36-6.19), t = 6.68, df = 159, p < 0.001, dz = 0.53, Shapiro-Wilk p = 0.129.
7. Control exercise: deliberately run, incorrectly, the independent-samples t-test on the same data (blood glucose at enrollment and at 6 months treated as two groups). In Google Sheets, =T.TEST(S2:S161, U2:U161, 2, 2). Compare the p value with that of the paired test and explain the difference (Chapter 6, Section 6.3.3).

L6.3 The Independent-Samples t-Test in JASP

The question: does SBP differ between women and men? H0: μF = μM; H1: μF ≠ μM.

1. T-Tests > Independent Samples T-Test. Move tas0 into Variables and sex into Grouping Variable.
2. Under Tests check Student, Welch and Mann-Whitney.
3. Under Additional Statistics check Location parameter, Confidence interval, Effect size (Cohen's d), Descriptives and Descriptives plots.
4. Under Assumption Checks check Normality and Equality of variances (Levene's test).

What you should obtain: women 131.37 ± 14.47 mmHg$(n = 86)$, men 130.85 ± 13.25 mmHg$(n = 74)$; Shapiro-Wilk p = 0.426 (F) and 0.648 (M); Levene p ≈ 0.52 (the variances can be considered equal); Student's test t = 0.24, df = 158, p = 0.81; Welch's test t = 0.24, df = 157.4, p = 0.81; difference F - M 0.52 mmHg (95% CI approximately -3.8 to 4.9); d =

0.04; Mann-Whitney p = 0.80. JASP computes the difference in the alphabetical order of the groups (F minus M), which is why the sign is positive; in the chapter we reported M minus F, -0.5 mmHg. The sign does not change the conclusion, but it must be stated in the text.

5. Repeat for glicemie0 by sex. Expected values: F 104.81 ± 16.30, M 106.16 ± 15.19; Levene p ≈ 0.43; Student t = -0.54, df = 158, p = 0.591; Welch p = 0.589; difference F - M -1.35 mg/dL (95% CI -6.27 to 3.57); d = -0.09. Pay attention to the normality check: in the group of women, Shapiro-Wilk p = 0.022. With 86 female patients and a moderate departure, the t-test remains acceptable; the Mann-Whitney test (p = 0.44) confirms the conclusion. Report both results in your assignment.
6. Repeat for inaltime by sex: difference F - M -13.37 cm (95% CI -15.28 to -11.45), t = -13.79, df = 158, p < 0.001, d ≈ -2.19; Levene p > 0.7. An enormous effect, also evident on the box plot.
7. Repeat for varsta by sex: F 55.42 ± 8.95, M 57.31 ± 8.50; Welch t = -1.37, p = 0.17; d = -0.22. Discuss why checking age by sex matters before interpreting other differences between the sexes (age is a possible confounder).

L6.4 One-Way ANOVA in JASP

The main question of the study: does the decrease in blood glucose at 6 months differ among arms A, B and C? H0: μA = μB = μC; H1: at least one mean differs.

1. ANOVA > ANOVA. Move dglic into Dependent Variable and brat into Fixed Factors.
2. Under Display check Descriptive statistics and Estimates of effect size (η² and ω²).
3. Under Assumption Checks check Homogeneity tests (Levene) and Q-Q plot of residuals. Under Homogeneity corrections leave None for the main analysis; also check Welch for comparison.
4. Under Post Hoc Tests move brat into the box on the right and check Tukey; add Games-Howell and Bonferroni for comparison; check Confidence intervals and Effect size.
5. Under Descriptives Plots put brat on the horizontal axis and check Display error bars, Confidence interval 95%.
6. Under Nonparametrics move brat for the Kruskal-Wallis test.
7. For normality by group, use Descriptives > Descriptive Statistics, with dglic in Variables and brat in Split, and check Shapiro-Wilk test.

What you should obtain: A 3.0 ± 8.8 mg/dL($n = 53$), B 11.1 ± 11.2($n = 53$), C 16.3 ± 10.6($n = 54$). The ANOVA table: brat SS = 4806.9, df = 2, MS = 2403.4; Residuals SS = 16459.5, df = 157, MS = 104.8; F = 22.93, p < 0.001; η² = 0.226; ω² = 0.215. Levene's test: p = 0.089 or p = 0.163, depending on the version of the test used by the program (centered on the mean or on the median); in both cases above 0.05. Welch's ANOVA: F = 26.05, p < 0.001. Shapiro-Wilk by arm: A p = 0.575, B p = 0.052, C p = 0.903. Tukey: B - A 8.1 mg/dL (95% CI 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 and 0.038. Kruskal-Wallis: H = 36.63, df = 2, p < 0.001.

8. Repeat for dgr (weight loss). Expected values: 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. Repeat for tas0 with fumat as the factor. Expected values: former smokers 127.55 ± 14.39($n = 33$), current smokers 137.04 ± 14.63($n = 53$), never smokers 128.50 ± 11.76($n = 74$); F(2, 157) = 7.85, p < 0.001 (0.0006); η² = 0.091; Levene p ≈ 0.11; Tukey: current smokers - former smokers 9.49 mmHg (p = 0.004), current smokers - never smokers 8.54 mmHg (p = 0.001), never smokers - former smokers 0.95 mmHg (p = 0.94).
10. Interpret on paper, in 4-5 sentences, the result from point 7: which arms differ, how large the effect is, and what the confidence interval of the C - B difference, which starts at 0.6 mg/dL, means.

L6.5 Repeated Measures ANOVA in JASP

The question: does blood glucose change between 0, 3 and 6 months? H0: μ0 = μ3 = μ6.

1. ANOVA > Repeated Measures ANOVA. Under Repeated Measures Factors rename the factor "time" and the levels "0 months", "3 months", "6 months".
2. Move glicemie0, glicemie3 and glicemie6 into Repeated Measures Cells, in the correct order.
3. Under Assumption Checks check Sphericity tests, and under Sphericity corrections check None and Greenhouse-Geisser.
4. Under Post Hoc Tests move time and check Bonferroni. Under Descriptives Plots put time on the horizontal axis.

What you should obtain: means 105.4 / 99.4 / 95.2 mg/dL; $F(2, 318) = 64.97$, $p < 0.001$; Mauchly's test $W = 0.873$, $p < 0.001$ (sphericity is not satisfied); Greenhouse-Geisser $\varepsilon = 0.887$, corrected degrees of freedom 1.77 and 282.2, $p < 0.001$; Bonferroni comparisons: all three pairs $p < 0.001$. The Friedman test (ANOVA > Repeated Measures ANOVA > Nonparametrics, or Chapter 7): $\chi^2 = 74.67$, $p < 0.001$.

5. Mixed model (optional, for students who finish early): add brat to Between Subject Factors. Expected values: effect of time $F(2, 314) = 73.56$, $p < 0.001$; effect of arm $F(2, 157) = 8.50$, $p < 0.001$; time × arm interaction $F(4, 314) = 11.51$, $p < 0.001$. Means at 0 / 3 / 6 months: A 107.8 / 105.7 / 104.8; B 106.3 / 101.3 / 95.2; C 102.3 / 91.2 / 86.0 mg/dL. The significant interaction means that the course over time differs between arms: this is, in fact, the question about the effectiveness of the interventions. Draw the plot of the means with one line for each arm (Descriptives Plots, brat in Separate lines).

L6.6 The Same Analyses in MedCalc

1. Import mg2026_curat.csv (File > Import or File > Open, CSV type) and check that the numerical variables are recognized as numbers.
2. The paired test: Statistics > T-tests > Paired samples t-test; Variable 1: glicemie0, Variable 2: glicemie6. The result should match L6.2 (mean difference, 95% CI, $t = 11.14$, $p < 0.0001$).
3. The independent test: Statistics > T-tests > Independent samples t-test. For Sample 1 choose the variable tas0 with the selection criterion sex="F", and for Sample 2 the same variable with sex="M". MedCalc also reports the F test for equality of variances, the classic t-test and the version corrected for unequal variances.
4. ANOVA: Statistics > ANOVA > One-way analysis of variance. Under Data you can write the expression glicemie0-glicemie6 directly (or a previously computed variable); under Factor choose brat. Check Levene's test and a post hoc test (Tukey-Kramer or Bonferroni). Compare the table with the one from JASP.
5. From summary data, without a file: the Tests > Comparison of means (t-test) menu lets you enter the mean, SD and n for two groups. Enter the BMI values from a published article, $28.15 \pm \frac{6.49\ \mathrm{kg}}{m^2} (n = 52)$ and $24.47 \pm \frac{4.60\ \mathrm{kg}}{m^2} (n = 84)$ [71], and compare the result with the p value reported by the authors ($p = 0.001$). You should obtain a t of around 3.6-3.9 (depending on the version of the test) and $p < 0.001$.

L6.7 t-Tests and the ANOVA Table in Google Sheets

The function T.TEST(range1, range2, tails, type) gives the p value directly. The "tails" argument is 1 or 2 (use 2, two-sided). The "type" argument is 1 for paired samples, 2 for independent samples with equal variance (Student) and 3 for independent samples with unequal variances (Welch).

1. The paired test: =T.TEST(S2:S161, U2:U161, 2, 1). Result: approximately 1.1 × 10^-21, that is, $p < 0.001$.

2. The independent test by sex, with FILTER to separate the groups: =T.TEST(FILTER(M2:M161, C2:C161="F"), FILTER(M2:M161, C2:C161="M"), 2, 2) gives 0.814 (Student), and with type 3 it gives 0.813 (Welch).
3. The t statistic for the paired test, built by hand: in separate cells compute the mean of the differences =AVERAGE(AC2:AC161) (10.19), the standard deviation =STDEV(AC2:AC161) (11.57), the standard error =STDEV(AC2:AC161)/SQRT(160) (0.915), t = mean / SE (11.14), p =T.DIST.2T(t, 159) and the 95% CI limits = mean ± T.INV.2T(0.05, 159) × SE (8.38 and 11.99).
4. The F test for the ratio of variances: =F.TEST(FILTER(M2:M161, C2:C161="F"), FILTER(M2:M161, C2:C161="M")) gives approximately 0.44. The F test is sensitive to departures from normality; Levene's test, from JASP, is preferable, but here both lead to the same conclusion.
5. The ANOVA table for dglic by arm, built with formulas. In a new sheet write the letters A, B, C in A2:A4. In B2 =COUNTIF(Date!B$2:B$161, A2) (n per arm: 53, 53, 54). In C2 =AVERAGEIF(Date!B$2:B$161, A2, Date!AC$2:AC$161) (means 3.02; 11.09; 16.33). In D2 =DEVSQ(FILTER(Date!AC$2:AC$161, Date!B$2:B$161=A2)) (sum of squared deviations in each arm: 3987.0; 6532.5; 5940.0). Drag the formulas down to row 4.
6. The grand mean: =AVERAGE(Date!AC2:AC161) (10.19). Total SS: =DEVSQ(Date!AC2:AC161) (21266.4). Within-groups SS: =SUM(D2:D4) (16459.5). Between-groups SS: =SUMPRODUCT(B2:B4, (C2:C4-grand mean)^2) or, more simply, total SS - within-groups SS (4806.9).
7. MS between = SS between / 2 = 2403.4; MS within = SS within / 157 = 104.8; F = MS between / MS within = 22.93; p =F.DIST.RT(F, 2, 157), approximately 1.8 × 10^-9; critical value =F.INV.RT(0.05, 2, 157) = 3.05; eta² = SS between / SS total = 0.226. Arrange the results in a table with the columns Source, SS, df, MS, F, p, exactly as in Table 6.5 in the chapter.
8. The add-on version: from Extensions > Add-ons you can install a statistical analysis add-on (for example, XLMiner Analysis ToolPak), which offers "Anova: Single Factor". The add-on requires the data for each group in a separate column; you can obtain them with =FILTER(Date!AC2:AC161, Date!B2:B161="A") and the analogous formulas for B and C. The result should match the table built by hand.

L6.8 The Assignment in the a-g Format: The Model for ANOVA

(a) Research hypothesis: the decrease in fasting plasma glucose at 6 months differs among patients who received standard counseling, counseling with an app, and counseling with an app and teleconsultation.

(b) H0: the mean decrease in blood glucose is the same in the three populations, μA = μB = μC. H1: at least one arm has a mean decrease different from the others. α = 0.05.

(c) Method: a numerical variable (dglic) compared across three independent groups; one-way ANOVA, followed by Tukey's test; conditions checked: independence (individual randomization), normality by arm (Shapiro-Wilk p between 0.052 and 0.903), homogeneity of variances (Levene p > 0.05, SD ratio 1.27). The alternative in case of departure: Welch's ANOVA and Kruskal-Wallis.

(d) Software: JASP, ANOVA > ANOVA, with Post Hoc Tests Tukey and Descriptives Plots; check in Google Sheets with the table built in L6.7.

(e) Result: 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, η² = 0.226; Tukey B - A 8.1 (95% CI 3.4-12.8), C - A 13.3 (8.6-18.0), C - B 5.2 (0.6-9.9), all p < 0.05. Graph: box plot by arm or the plot of the means with 95% CIs.

(f) Statistical conclusion: p < 0.001, H0 is rejected; the differences between arms are highly significant, and all three pairs differ significantly.

(g) Research conclusion: adding the self-monitoring app to counseling increased the decrease in fasting plasma glucose at 6 months by approximately 8 mg/dL, and monthly teleconsultation brought an additional benefit of approximately 5 mg/dL compared with the app, with a less precise estimate (95% CI 0.6-9.9 mg/dL). The treatment arm explains almost a quarter of the variability in the decrease in blood glucose. The study is randomized, so the differences can be attributed to the interventions; the data, however, are simulated for teaching purposes.

Additional Exercises

Exercises 1-6 are done on the discipline's classic files on Moodle; the results are not given, and each is solved in the a-g format, with a check of the conditions of application and with the appropriate graph. Exercises 7 and 8 have the data in the statement and a complete solution.

1. The BD_lp3 file. Are there significant differences between blood glucose at time zero and blood glucose at 3 months? And between 3 and 6 months? Then answer the question for all three time points with repeated measures ANOVA and compare the conclusions.
2. The BD_lp3 file. Does blood glucose at 6 months differ between insured and uninsured patients? Does age differ between the two groups? If age differs, how could this influence the interpretation of the difference in blood glucose?
3. The BD_lp3 file. Does the course of blood glucose at 0, 3 and 6 months differ between insured and uninsured patients? Use the mixed model from L6.5 (insurance status as the between-subjects factor) and interpret the interaction term.
4. The BPOC file (COPD data). Did heart rate decrease significantly after treatment (FC1 vs. FC2)? Did percent-predicted FEV1 increase (VEMSP1 vs. VEMSP2)? Check the normality of the differences and, if appropriate, report the Wilcoxon test.
5. The BPOC file. Does the baseline bronchial patency index (IPB1) differ among patients who received drugs A, B and C? Apply ANOVA with Levene's test and the appropriate post hoc test; if the variances differ, use Welch's ANOVA and Games-Howell.
6. The BD_Studenti file. (a) Does height differ between the sexes? (b) Does actual height differ from the declared ideal height (paired test)? (c) Does height differ among study groups? For (c), discuss why a significant result might merely reflect the sex composition of the groups.
7. Data in the statement, with solution. Heart rate (beats/min) in 6 patients with COPD before and after a cardioselective beta-blocker: before 96, 88, 102, 90, 94, 100; after 84, 82, 90, 86, 85, 91. The differences are 12, 6, 12, 4, 9, 9; $m_D = \frac{52}{6} = 8.67$; the deviations from the mean: 3.33; -2.67; 3.33; -4.67; 0.33; 0.33; sum of squares 11.11 + 7.11 + 11.11 + 21.78 + 0.11 + 0.11 = 51.33; $s_D = \sqrt{\frac{51.33}{5}} = 3.20$; SED $= \frac{3.20}{\sqrt{6}} = 1.31$; $t = \frac{8.67}{1.31} = 6.63$, df = 5; critical t (5, 0.05) = 2.571, so $p < 0.05$ (exactly $p = 0.0012$). 95% CI: 8.67 ± 2.571 × 1.31 = 5.30-12.03 beats/min. Conclusion: heart rate decreased significantly, on average by about 9 beats/min; with 6 patients and no control group, the result is only indicative.
8. Data in the statement, with solution. Hemoglobin (g/dL) in three groups of 4 pregnant women each, in the first, second and third trimesters (independent groups): 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. The means: 12.80; 11.90; 11.60; grand mean 12.10. SS between = 4 × [(0.70)² + (-0.20)² + (-0.50)²] = 4 × 0.78 = 3.12, df = 2, MS = 1.56. SS within: 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, df = 9, MS = 0.0711. $F = \frac{1.56}{0.0711} = 21.9$; critical F (2, 9, 0.05) = 4.26, so p < 0.05 (exactly p < 0.001). $\eta^2 = \frac{3.12}{3.76} = 0.83$. Tukey's test: q(3, 9, 0.05) = 3.95; HSD = $3.95 \times \sqrt{\frac{0.0711}{4}} = 3.95 \times 0.133 = 0.53$ g/dL; T1 - T2 = 0.90 and T1 - T3 = 1.20 are significant, T2 - T3 = 0.30 is not. Conclusion: mean hemoglobin is higher in the first trimester than in the second and third trimesters, consistent with the physiological hemodilution of pregnancy; the difference between the second and third trimesters could not be demonstrated in this very small sample.

What to submit and indicative grading scheme

Submit on Moodle a single file containing: the results of the paired test for glicemie0-glicemie6 (the JASP table, the normality test on the differences, the graph) with the interpretation (2 points); the results of the independent test for tas0 and glicemie0 by sex, with Levene's test and the justification for the choice among Student, Welch and Mann-Whitney (2 points); the ANOVA for dglic by arm, with the check of the conditions, Tukey's test and the plot of the means (2 points); the ANOVA table built in Google Sheets, with the formulas visible (1.5 points); the repeated measures ANOVA with Mauchly's test and the Greenhouse-Geisser correction (1 point); an assignment in the a-g format on one of the additional exercises 1-6 (1.5 points). Total: 10 points. Points are deducted for: the independent test applied to paired data, normality checked on each series instead of on the differences in the paired test, multiple t-tests instead of ANOVA, no post hoc test after a significant F, missing effect size and confidence interval, and the wording "H0 is accepted".

Offline variant

Without a network, JASP and MedCalc work locally after installation; the mg2026_curat.csv file is copied from the instructor's USB stick. The Google Sheets part is replaced with LibreOffice Calc or Excel: the functions T.TEST, T.DIST.2T, T.INV.2T, DEVSQ, AVERAGEIF, COUNTIF, F.DIST.RT and F.INV.RT exist there too (in Excel, FILTER is available in recent versions; otherwise, sort the data by arm and select the range of each group manually). In Excel, ANOVA can also be obtained from Data > Data Analysis > Anova: Single Factor, after activating the Analysis ToolPak add-in. If no computer is available, additional exercises 7 and 8 are solved entirely by hand, with the tables of critical t and F values and the q table for Tukey's test.

Practical Session 7. Nonparametric Tests: Mann-Whitney, Wilcoxon, Kruskal-Wallis, Friedman and Chi-Square

Objective

In this session you apply the tests from Chapter 7 on the computer, to the mg2026 dataset, in three different programs, and check that you obtain the same results. The emphasis is not on pressing buttons but on three things that will matter in the practical exam: choosing the right test for the type of variable and design, reading the output tables correctly (which statistic, which p value, which effect size) and formulating the conclusion. Each step gives you the figures you should obtain, so that you can check yourselves; if your result differs, look for the mistake before moving on.

By the end, the student should be able to:

- apply in JASP the Mann-Whitney, Wilcoxon, Kruskal-Wallis (with Dunn's post hoc test) and Friedman tests and extract the statistic, the p value and the effect size;
- build in JASP a contingency table with the expected frequencies, the chi-square test, the Yates correction, Fisher's test and Cramér's V;
- enter a 2 x 2 table and a 2 x n table in Epi Info StatCalc and interpret the output (uncorrected chi-square, Yates, Fisher, OR, RR, the test for trend);
- reproduce in Google Sheets the observed frequencies, the expected frequencies and the p value of the chi-square test;
- check the conditions of application (expected frequencies, paired or independent data) and choose the correct version of the test;
- formulate the statistical conclusion and the clinical conclusion for each test.

Tools

- JASP 0.98.1 [47], with the default modules T-Tests, ANOVA and Frequencies.
- MedCalc 24.0 [50] (on the laboratory computers).
- Epi Info 7.2.7 [51], the StatCalc component and, optionally, Classic Analysis.
- Google Sheets, with English (United States) regional settings (the argument separator is ",", and the decimal separator is the point).
- The file mg2026_curat.csv (160 patients, 28 columns) from the discipline's Moodle platform.
- Approximate time: 100 minutes (JASP 50 minutes, Epi Info 15 minutes, Google Sheets 20 minutes, conclusions and upload 15 minutes).

The columns of the file, in order, are: 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. The data are in rows 2-161.

Procedure

L7.1 Preparing the Data in JASP

1. Open JASP and load the file: main menu (the three lines at the top left) > Open > Computer > Browse, then select mg2026_curat.csv.

2. Check the type of each variable, shown by the icon to the left of the column name (a ruler for continuous, stepped bars for ordinal, three circles for nominal). Click the icon to change it. Set brat, sex, mediu, fumat, diabet_ogtt and scadere5 as **nominal**; educatie, activitate and satisfactie as **ordinal**; and the rest as **continuous**.
3. Create the variable "dglic" (the decrease in blood glucose at 6 months): click the "+" sign next to the last column, enter the name dglic, choose the continuous type and the R formula as the computation method, then type glicemie0 - glicemie6 and press Compute column. Check: the mean of the dglic column must be 10.19.
4. In the same way, create the variable "obez", nominal, with the R formula ifelse(imc0 >= 30, 1, 0). You should obtain 54 values of 1 and 106 values of 0.
5. Create the variable "glic100_0" with the formula ifelse(glicemie0 >= 100, 1, 0) and "glic100_6" with the formula ifelse(glicemie6 >= 100, 1, 0). You should obtain 97 and 62 values of 1, respectively.

Why does setting the variable type matter? JASP will not let you put a continuous variable in the "Rows" box of a contingency table and, conversely, it offers only the tests suited to the declared type. A wrong type is the most common reason why "the test does not show up" in the lab.

L7.2 The Mann-Whitney U Test in JASP: Satisfaction in Arm C Versus Arm A

The Mann-Whitney test compares two groups, so you must temporarily exclude arm B.

1. Click the name of the brat column; in the label window that opens, uncheck "Filter" for the value B (or use the filter button, the funnel symbol). The rows of the patients in arm B turn gray: they are excluded from the analyses. Check the bottom bar: 107 active rows.
2. T-Tests > Classical > Independent Samples T-Test.
3. Dependent Variables: satisfactie (if JASP does not accept the ordinal variable, temporarily set it as "Scale"). Grouping Variable: brat.
4. Under Tests, uncheck Student and check **Mann-Whitney**. Under Additional Statistics, check **Effect size** (for Mann-Whitney, JASP reports the rank-biserial correlation) and **Descriptives**.
5. For the description, use Descriptives > Descriptive Statistics, with satisfactie under Variables and brat under Split, checking Median and Quartiles (or Interquartile range).

What you should obtain: median 3 (Q1 2, Q3 3) in arm A and 4 (Q1 3, Q3 5) in arm C; W (the Mann-Whitney statistic reported by JASP) = 596.5 or 2,265.5, depending on which group is taken as the reference (the order A, C gives 596.5); $p < 0.001$; rank-biserial correlation 0.58 in absolute value (the sign depends on the order of the groups). Write on the worksheet: "$U = \frac{2265.5}{596.5}$ are the same statistic, seen from the two groups; U1 + U2 = 54 x 53 = 2,862".

6. Repeat for arms B and A (now filter out arm C). You should obtain $p < 0.001$ and a rank-biserial correlation of 0.42. Then for C and B: $p = 0.079$, correlation 0.19.
7. At the end, reactivate all arms (check "Filter" again for B and C). Forgetting an active filter is a classic source of wrong results in the following exercises.

Written interpretation (on the worksheet): "Satisfaction was higher in arm C than in arm A (median 4 vs. 3; Mann-Whitney $p < 0.001$; rank-biserial correlation 0.58, large effect)."

L7.3 The Wilcoxon Test for Paired Data in JASP

Exercise A: blood glucose at 0 and 6 months, all patients.

1. T-Tests > Classical > Paired Samples T-Test.
2. Move glicemie0 and glicemie6 into the Variable Pairs box (they form a pair).

3. Under Tests, check both Student and **Wilcoxon signed-rank**. Under Additional Statistics, check Effect size and Descriptives, and under Assumption Checks, Normality (Shapiro-Wilk on the differences).

What you should obtain: paired t-test t = 11.14, df 159, p < 0.001; Shapiro-Wilk on the differences p = 0.725 (the differences are normal, so the t-test is the main analysis); Wilcoxon p < 0.001. When both tests agree, report the test chosen in advance; here, the paired t-test, with a mean difference of 10.19 mg/dL (95% CI 8.38-11.99).

Exercise B: weight at 0 and 6 months, arm A only.

1. Filter out arms B and C (53 active rows remain).
2. Paired Samples T-Test with the pair greutate0 - greutate6, checking Student, Wilcoxon signed-rank and Normality.

What you should obtain: Wilcoxon W = 870 (the sum of the positive ranks for the difference greutate0 - greutate6; if you reverse the order of the variables you obtain 456), p = 0.052; paired t-test t = 2.07, df 52, p = 0.043; Shapiro-Wilk on the differences p = 0.696. The median of the differences is 1.1 kg, the mean 0.71 kg.

Discuss on the worksheet, in 4-5 lines: which test is the correct one here and why (the distribution of the differences is normal, so the paired t-test is justified); why Wilcoxon gives a slightly larger p value (it uses only the ranks, and its relative efficiency is about 95%); and why we are not allowed to choose the test after seeing which p value falls below 0.05.

3. Optionally, the sign test: in the column of differences, count how many are positive (32), negative (19) and zero (2). In Frequencies > Binomial Test, on a variable that codes the sign of the difference, with Test value 0.5, you should obtain p = 0.092.

L7.4 The Kruskal-Wallis Test with Dunn's Post Hoc Test in JASP

1. Reactivate all rows. ANOVA > Classical > ANOVA.
2. Dependent Variable: dglic; Fixed Factors: brat.
3. Open the **Nonparametrics** section and move brat into the Kruskal-Wallis box. Check **Dunn's post hoc test** (in some versions, Dunn appears in the Post Hoc Tests section, under Type: Dunn).
4. For comparison, also request the classical ANOVA, with Assumption Checks > Homogeneity tests and Post Hoc Tests > Tukey.

What you should obtain: Kruskal-Wallis H = 36.63, df 2, p < 0.001; Dunn with Bonferroni correction: A vs. B p = 0.0003, A vs. C p < 0.001, B vs. C p = 0.12 (unadjusted 0.041). ANOVA: F(2, 157) = 22.93, p < 0.001; Levene p = 0.163; Tukey C vs. B p = 0.024.

5. Repeat for satisfactie (Dependent Variable: satisfactie). You should obtain H = 32.16, df 2, p < 0.001; Dunn-Bonferroni: A vs. B p = 0.0003, A vs. C p < 0.001, B vs. C p = 0.29.
6. Compute the effect size $\varepsilon^2 = \frac{H}{N-1}$ by hand: 36.63/159 = 0.23 for dglic and 32.16/159 = 0.20 for satisfaction.

Check question (answer on the worksheet): why is the B-C comparison significant with Tukey (p = 0.024) but not with Dunn-Bonferroni (p = 0.12)? Which is the main analysis for dglic, and why?

L7.5 The Friedman Test in JASP

1. ANOVA > Classical > Repeated Measures ANOVA.
2. In the Repeated Measures Factors box, rename the factor "RM Factor 1" to "time" and the levels to "0 months", "3 months", "6 months".
3. Move glicemie0, glicemie3 and glicemie6, in this order, into the Repeated Measures Cells box.

4. Open the **Nonparametrics** section, move the time factor into the Friedman box and check Conover's post hoc tests.
5. For the description: Descriptives > Descriptive Statistics on the three variables, with Median and Quartiles.

What you should obtain: medians 104.5, 99.0 and 92.5 mg/dL; Friedman χ^2 = 74.67, df 2, p < 0.001; Kendall's W = 0.23 (if JASP does not display it, compute it: 74.67/(160 x 2)); repeated measures ANOVA F = 64.97, p < 0.001. The Conover post hoc comparisons should show significant differences between all three time points; as an independent check, the Wilcoxon tests with Bonferroni correction in Table 7.7 of Chapter 7 give adjusted p < 0.001 (0-3 months), 0.0007 (3-6 months) and < 0.001 (0-6 months).

6. Filter to arm A only and repeat. You should obtain Friedman χ^2 = 6.04, p = 0.049, with medians of 111, 104 and 105 mg/dL. Discuss: the overall result is borderline; does any comparison remain significant after correction? (Answer to check: no, the smallest Bonferroni-adjusted value is 0.075.)

L7.6 Contingency Tables and the Chi-Square Test in JASP

Exercise A: arm and weight loss ≥ 5% (3 x 2 table).

1. Reactivate all rows. Frequencies > Classical > Contingency Tables.
2. Rows: brat; Columns: scadere5.
3. Under Statistics, check χ^2 and, under Nominal, **Phi and Cramér's V**. Under Cells, check Counts **Expected** and Percentages **Row**.

What you should obtain: observed frequencies 6/47, 16/37, 16/38; expected frequencies 12.59/40.41 (A), 12.59/40.41 (B), 12.82/41.18 (C); χ^2 = 6.77, df 2, p = 0.034; Cramér's V = 0.21; row percentages 11.3%, 30.2% and 29.6%. Check the condition for applying the test: the smallest expected frequency is 12.6, so the test is valid.

4. The chart: Descriptives > Descriptive Statistics, the variable scadere5 with Split by brat, Plots > Distribution plots; or, more simply, build the bar chart of the percentages in Google Sheets (step L7.8). Remember the point made in the lab: for chi-square, the chart is not produced from the Contingency Tables menu but from Descriptives.

Exercise B: obesity and diabetes (2 x 2 table).

1. Contingency Tables with Rows: obez, Columns: diabet_ogtt.
2. Check χ^2, χ^2 **continuity correction** and, in the section for 2 x 2 tables, **Log odds ratio** and **Fisher's exact test**. Under Cells: Expected, Row percentages.

What you should obtain: observed 95/11 (non-obese) and 36/18 (obese); expected 86.79/19.21 and 44.21/9.79; χ^2 = 12.70, df 1, p < 0.001; with continuity correction 11.20, p = 0.0008; Fisher p = 0.0008; φ = 0.28. Log odds ratio = 1.46, hence $\mathrm{OR} = e^{1.46} = 4.32 (; \mathrm{CI}\, 95\%\, 1.86 - 10.03)$, regardless of whether you compute it as obese versus non-obese.

3. Write on the worksheet why, for this table, the three variants (Pearson, Yates, Fisher) lead to the same conclusion (all expected frequencies are above 5, the smallest being 9.79), and in what situation they would have given different conclusions.

Exercise C: a situation in which the chi-square test is not valid.

Filter to patients under 50 years only (add a computed variable ifelse(varsta < 50, 1, 0) and filter by it) and build the obez x diabet_ogtt table for this subgroup. Note how many cells have an expected frequency below 5 and which test you choose. (The subgroup has 34 patients and only 2 cases of diabetes; the appropriate test is Fisher's.)

L7.7 The Same Tests in MedCalc

Open mg2026_curat.csv in MedCalc (File > Open, file type text/CSV; check that the decimal comma is recognized correctly). Menu names may differ slightly between versions; look for the tests by name.

1. Mann-Whitney: the Statistics menu, under the rank-based tests, Mann-Whitney test (independent samples). Variable: satisfactie; Grouping/Filter: brat = "A" for the first sample and brat = "C" for the second. What you should obtain: the same p value (< 0.001) as in JASP.
2. Wilcoxon: Wilcoxon test (paired samples), with greutate0 and greutate6, Filter: brat = "A". What you should obtain: p = 0.052.
3. Kruskal-Wallis: under analysis of variance, Kruskal-Wallis test, Variable: dglic (create it beforehand with the Variable editor or type the expression glicemie0-glicemie6 directly), Factor: brat. What you should obtain: H = 36.63, p < 0.001, plus the post hoc comparisons offered by the program.
4. Friedman: Friedman test, with the three blood glucose columns. What you should obtain: χ^2 = 74.67 (or a very close value, depending on the correction for ties), p < 0.001.
5. Chi-square: under contingency tables, Chi-squared test, with brat and scadere5. What you should obtain: χ^2 = 6.77, df 2, p = 0.034. MedCalc also has the test for trend (Chi-squared test for trend), useful here because the arms have increasing intensity: what you should obtain is a χ^2 value for trend of about 4.9, with p = 0.027.
6. McNemar: McNemar test, with glic100_0 and glic100_6 (created as 0/1 variables). What you should obtain: discordant pairs 36 (≥ 100 only at enrollment) and 1 (≥ 100 only at 6 months), p < 0.001. The value of the statistic depends on the variant displayed: 33.11 without correction or 31.24 with correction.

L7.8 2 x 2 and 2 x n Tables in Epi Info StatCalc

Epi Info StatCalc works with tables that are already summarized: it does not need the data file, only the frequencies.

2 x 2 table: obesity x diabetes.

1. Open Epi Info 7 > StatCalc > Tables (2 x 2, 2 x n).
2. Enter: exposed (obese) with the disease (diabetes) 18, exposed without the disease 36, unexposed with the disease 11, unexposed without the disease 95. Pay attention to the orientation of the table: in StatCalc, the rows are the exposure, the columns are the disease, and the top left cell is "exposed and diseased".
3. Read the output. What you should obtain: Chi-square uncorrected 12.70 (p = 0.0004); Chi-square Yates corrected 11.20 (p = 0.0008); Fisher exact, two-sided, about 0.0008; Odds ratio 4.32 (95% CI 1.86-10.03); Risk ratio 3.21 (95% CI 1.64-6.31). StatCalc also displays a Mantel-Haenszel variant of the chi-square, slightly smaller than the uncorrected one (12.62), and mid-p intervals; do not confuse them with the values from JASP.
4. Interpret, on the worksheet: diabetes is 3.2 times more frequent in obese patients (the prevalence ratio, cross-sectional study), and the association is statistically significant.

2 x n table: arm and weight loss ≥ 5%.

5. In the same module, add columns to obtain three levels of exposure (A, B, C) and enter the cases (weight loss ≥ 5%: 6, 16, 16) and the non-cases (47, 37, 38). What you should obtain: chi-square about 6.77, with 2 degrees of freedom, p = 0.034.
6. StatCalc > Chi square for trend: enter the exposure scores 0, 1, 2 for arms A, B, C, with the cases and non-cases above. What you should obtain: chi-square for trend about 4.9, p = 0.027 (Epi Info uses the Mantel variant of the test, which may differ in the second decimal place). Explain on the worksheet why the test for trend has 1 degree of freedom, while the overall test has 2.

Optionally, in Classic Analysis: READ (the CSV file), then TABLES brat scadere5 for a 3 x 2 table with the chi-square displayed automatically, and MEANS dglic brat, which displays the ANOVA, Bartlett's test and the Kruskal-Wallis test in the same output.

L7.9 Chi-Square Step by Step in Google Sheets

Import mg2026_curat.csv into a new sheet (File > Import > Upload, Separator type: comma). Work in a second sheet, named "chi".

1. The observed table (arm in the rows, scadere5 in the columns). In cells A2:A4 type A, B, C; in B1 type "yes", in C1 "no". In B2 type the formula:

   =COUNTIFS(mg2026!$B$2:$B$161, $A2, mg2026!$AA$2:$AA$161, 1)

   and in C2:

   =COUNTIFS(mg2026!$B$2:$B$161, $A2, mg2026!$AA$2:$AA$161, 0)

   Copy the formulas down to row 4. You should obtain 6/47, 16/37, 16/38.
2. The totals: in D2 =SUM(B2:C2) (copy down), in B5 =SUM(B2:B4) (copy to the right), in D5 =SUM(D2:D4). You should obtain 53, 53, 54 for the rows, 38 and 122 for the columns, 160 in total.
3. The expected table, in F2:G4. In F2 type:

   =$D2*B$5/$D$5

   and copy the formula into the whole block F2:G4. Notice the $ symbols: the row of the column totals and the column of the row totals stay fixed. You should obtain 12.59/40.41, 12.59/40.41, 12.82/41.18.
4. The p value directly: =CHISQ.TEST(B2:C4, F2:G4). You should obtain 0.034. Caution: the function takes the observed table and the expected table, not the raw data, and it applies no continuity correction.
5. The χ^2 statistic: =SUMPRODUCT((B2:C4-F2:G4)^2/F2:G4). You should obtain 6.77. Check p: =CHISQ.DIST.RT(6.77, 2) gives 0.034, and the critical value =CHISQ.INV.RT(0.05, 2) gives 5.99.
6. Cramér's V: =SQRT(6.77/(160*1)), that is, 0.21.
7. The chart: select the column with the row percentages (compute in H2 =B2/D2, formatted as a percentage) and insert a bar chart (Insert > Chart, Column chart). It should look like Figure 7.4 in Chapter 7.
8. Ranks for a manual Mann-Whitney test (optional): for the example in Chapter 7 (the scores 2, 3, 3, 5, 6 and 4, 6, 7, 7, 8, 9 in A2:A12), in B2 type =RANK.AVG(A2, $A$2:$A$12, 1) and copy down. The sum of the ranks of the first five values, =SUM(B2:B6), must be 17.5, and U1 = 17.5 - 5 x 6/2 = 2.5.

Repeat steps 1-6 for the obesity x diabetes table (using a helper column with =IF(K2>=30, 1, 0) in the data sheet). You should obtain $\chi^2 = 12.70$ and $p = 0.0004$ (without the Yates correction).

Additional Exercises

1. In the BPOC file (COPD data) on Moodle, choose an ordinal or a skewed numerical variable and compare it between the sexes with the Mann-Whitney test; report the medians, U, p and the rank-biserial correlation. Justify why you did not use the t-test.
2. In the same file, compare a variable measured at the beginning and at the end of treatment (for example, a respiratory function value from time point 1 and time point 2) with the

Wilcoxon test and with the paired t-test. Check the normality of the differences and decide which test is the appropriate one.

3. In the BD_Studenti file, investigate the association between sex and the opinion about biostatistics (contingency table). Check the expected frequencies; if one of the categories has too few students, propose a meaningful merging of categories and redo the test.
4. In the Date_HbA1c file, compare the proportion of patients who reach the therapeutic target (HbA1c < 7.5%) between the first and the last year. Caution: if they are the same patients, the appropriate test is McNemar's, not the ordinary chi-square test.
5. Problem with data in the statement: in a cohort study of retirees, 30 of 80 people exposed to dust and 30 of 240 unexposed people showed a decrease in lung capacity. Build the table, compute the expected frequencies and χ^2 by hand, then check in Epi Info StatCalc.

   Solution: the total with a decrease is 60 of 320 (18.75%). Expected frequencies: exposed, $\frac{80\times60}{320} = 15$ with a decrease and 65 without; unexposed, $\frac{240\times60}{320} = 45$ with a decrease and 195 without. Observed: 30/50 and 30/210. |O - E| = 15 in each cell. $\chi^2 = \frac{225}{15} + \frac{225}{65} + \frac{225}{45} + \frac{225}{195} = 15 + 3.46 + 5 + 1.15 = 24.62$, df 1, $p < 0.001$. The risk in the exposed is 30/80 = 37.5%, in the unexposed 30/240 = 12.5%, RR = 3.0. Exposure to dust is a significant risk factor for decreased lung capacity in this sample.

What to submit and indicative grading scheme

Upload to Moodle a single PDF file (or the Google document exported as PDF) containing:

- the screenshots of the JASP output tables for steps L7.2-L7.6 (2 points);
- the completed verification table, with the values you obtained next to those in this text, and an explanation of any difference (2 points);
- the Epi Info output for the 2 x 2 table and for the test for trend, with the interpretation of the OR and RR (1.5 points);
- the Google Sheets sheet (link or screenshot) with the observed and expected tables and the formulas visible (1.5 points);
- the answers to the check questions in L7.3, L7.4 and L7.6 (2 points);
- the statistical and clinical conclusion, in 2-3 sentences, for each of the five tests (1 point).

Total: 10 points. Points are deducted for: a test applied to wrongly filtered data (a filter left active by mistake), the mean and standard deviation reported for satisfaction, a chi-square test reported without checking the expected frequencies, and the conclusion "H0 is accepted".

Offline variant

If the network or the Moodle platform is not available, the file mg2026_curat.csv is copied from the instructor's USB stick or from the shared folder of the laboratory. JASP, MedCalc and Epi Info work locally, without a connection. Step L7.9 (Google Sheets) is replaced by the same calculation in a locally installed spreadsheet (LibreOffice Calc or Excel), where the functions have the same names (CHISQ.TEST, CHISQ.DIST.RT, RANK.AVG, COUNTIFS) and the argument separator is again "," in the English configuration. If no program is available, work by hand on the obesity x diabetes table (expected frequencies, χ^2, df, comparison with 3.84) and on the Mann-Whitney example in Figure 7.1, using the tables of critical values in the appendix of this book. The file with the results is submitted at the next practical session.

Practical Session 8. Choosing the Statistical Test and Standardized Reporting of Results

Objective

The previous practical sessions taught you one test at a time. In this session you are no longer told which test to apply. You receive 15 research questions formulated on the mg2026 dataset, and for each one you must choose the test yourselves, justify it, apply it in the software and report the result in the discipline's standardized format (a-g). In the last part of the session you assess a colleague's report using a common grid. It is, in practice, a dress rehearsal for the practical exam, and the mistakes made here are the cheapest mistakes you will ever make.

By the end, the student should be able to:

- identify, for a given research question, the aim of the analysis, the type of outcome variable and the design (independent or paired, number of groups);
- choose the appropriate test and its alternative, using the algorithm in Chapter 8, and check the conditions of application in the software;
- apply the test in JASP (or MedCalc, Epi Info) and extract the statistic, the degrees of freedom, the p value and an effect size;
- write up the complete solution in the a-g format, with an appropriate chart;
- critically assess a colleague's report using a grid and formulate constructive comments;
- recognize the questions for which statistical testing is not the right answer (for example, baseline comparisons in a randomized trial).

Tools

- JASP 0.98.1 [47] (the main program); MedCalc 24.0 [50] and Epi Info 7.2.7 [51] for checks; Google Sheets for the simple charts and for the report document.
- The file mg2026_curat.csv from Moodle, with the variables computed in Practical Session 7 (dglic, obez, glic100_0, glic100_6). For this session, also create: dgr (greutate0 - greutate6), dtas (tas0 - tas6) and tas140_0, tas140_6 (ifelse(tas0 >= 140, 1, 0), and likewise for tas6).
- The report template (section L8.2), available on Moodle as a Google document.
- Figure 8.1 (the algorithm for choosing the test) and Table 8.1 (the summary of tests) from Chapter 8, open alongside.
- Approximate time: 100 minutes (review 10 minutes, solving the problems 60 minutes, peer assessment 20 minutes, final discussion and upload 10 minutes).

Procedure

L8.1 Reviewing the Algorithm (10 Minutes)

Before turning to the computer, work on paper. For each question in the list in L8.3, fill in only the first four columns of the table below, without opening the software. Only after you have decided on the test do you move on to JASP. The exercise forces you to choose the test from the design, not from what "comes out".

Table L8.1. The decision sheet (to be completed before the analysis).

No.	Aim (difference / association / agreement / prediction)	Outcome variable and its type	Design (groups, independent or paired)	Test chosen	Alternative (if the conditions are not met)
1					
...					
15					

The instructor checks each student's decision sheet before the calculations begin (for large groups, by spot checks). A wrong choice is corrected now, not after 40 minutes of work.

L8.2 The Standardized Report Template

Each problem is written up according to the template below. The total text for one problem must not exceed half a page, plus the chart. Conciseness is part of the requirement: a good clinical report says everything that is needed, and nothing more.

Table L8.2. The a-g report template.

Component	What to write	Example wording
(a) Research hypothesis	one sentence, in clinical language	"Women have higher HDL cholesterol than men."
(b) Statistical hypotheses	H0 and H1 in words and mathematically; two-sided	"H0: μF = μM, there is no difference between the mean HDL values; H1: μF ≠ μM."
(c) Method and justification	the type of the variables; the design; checking the conditions, with the p values of the checking tests; the test chosen and the alternative	"HDL numerical, sex nominal with 2 categories, independent groups; SW p = 0.41 and 0.48; Levene p = 0.034, hence Welch's t-test."
(d) Solution and software	the program, the menu, the options checked	"JASP, T-Tests > Independent Samples T-Test: Welch, Descriptives, Effect size, Assumption checks."
(e) Result, with a chart	descriptive statistics by group; the test statistic, df, p; the effect size with 95% CI; the chart	"F 53.92 ± 11.70; M 44.41 ± 9.66; t = 5.63, df 157.7, p < 0.001; difference 9.51 (95% CI 6.18-12.85); d = 0.88; box plot."
(f) Statistical conclusion	the decision regarding H0	"p < 0.001, we reject H0: the difference is statistically significant."
(g) Research conclusion	what it means clinically, with the effect size and the limitations	"Women have HDL cholesterol 9.5 mg/dL higher on average; large effect; cross-sectional analysis in a sample of patients at cardiometabolic risk."

Writing rules, checked during assessment:

- p values are written with three decimal places (p = 0.034), and those below 0.001 as "p < 0.001"; never "p = 0.000";
- normally distributed numerical variables are described by mean ± SD, skewed and ordinal ones by median (IQR), and nominal ones by n (%);
- the decision is worded as "we reject H0" or "we do not reject H0", never "we accept H0";
- the chart must match the type of data: a box plot or a chart of means with CIs for numerical variables, a bar chart (percentages) for nominal and ordinal variables, and a scatter plot for the association between two numerical variables;

- decimals are written with a decimal point.

An example of a complete report, on a question that is not part of the list below, shows how long a good answer actually needs to be.

The question: does fasting plasma glucose at baseline differ between patients with and without diabetes confirmed by the OGTT?

(a) Patients in whom the OGTT confirmed diabetes have higher fasting plasma glucose than those without diabetes. (b) H0: the mean fasting plasma glucose is the same in the two groups (μdiabetes = μno diabetes); H1: the means differ (μdiabetes ≠ μno diabetes), two-sided test. (c) Blood glucose is numerical, diabetes is nominal with two categories, and the groups are independent and unequal in size (29 and 131). Shapiro-Wilk by group: p = 0.198 (diabetes) and p = 0.061 (no diabetes); Levene p = 0.457. Test chosen: unpaired t-test; alternative: Mann-Whitney. (d) JASP, T-Tests > Independent Samples T-Test; Dependent Variables: glicemie0; Grouping Variable: diabet_ogtt; checked: Student, Welch, Mann-Whitney, Descriptives, Effect size, Assumption checks; the chart from Descriptives (box plot by group). (e) Without diabetes 101.85 ± 13.97 mg/dL$(n\,131)$, with diabetes 121.62 ± 13.19 mg/dL$(n\,29)$; t = -6.96, df 158, p < 0.001 (Welch t = -7.22, df 43.1, p < 0.001); difference 19.77 mg/dL (95% CI 14.25-25.28); Cohen's d = 1.43; Mann-Whitney p < 0.001; box plot attached. (f) p < 0.001 < 0.05, so we reject H0: the difference is statistically significant. (g) Patients with confirmed diabetes had fasting plasma glucose about 20 mg/dL higher on average, a very large effect. The result is to be expected, because fasting plasma glucose is one of the diagnostic criteria; if we wanted to know how well fasting plasma glucose identifies diabetes confirmed by the OGTT, the question would be one of prediction, and the appropriate method would be the ROC curve (AUC 0.844 in mg2026, Chapter 12).

L8.3 The 15 Problems

Solve the problems in order. Problems 1-10 are compulsory for all students; problems 11-15 are divided among the groups (each student solves at least two, as indicated by the instructor). The significance level is α = 0.05, and the tests are two-sided. The expected answers, for the instructor, are in the answer file for Chapter 8.

Problem L1. Does waist circumference differ between women and men? Pay attention to the result of the normality test in each group and to the size of the groups: decide, and argue, whether you use the t-test or the Mann-Whitney test.

Problem L2. Do triglycerides differ between patients with and without diabetes confirmed by the OGTT? Take into account the distribution of triglycerides (Chapter 3) and the very unequal size of the groups.

Problem L3. Does baseline HbA1c differ among the three arms? Solve it technically, then discuss in (g) whether this question makes sense in a randomized trial.

Problem L4. Does baseline total cholesterol differ among the three arms? (The same discussion as for L3; compare the two results.)

Problem L5. Does weight loss at 6 months (dgr) differ among the three arms? If the overall test is significant, specify which arms differ from each other.

Problem L6. Does BMI differ among never smokers, former smokers and current smokers? Check the distribution of BMI in each group and the homogeneity of variances before choosing the test.

Problem L7. Does satisfaction at 6 months differ between women and men?

Problem L8. Does age differ between patients from urban and rural areas?

Problem L9. Does the proportion of patients with a weight loss of at least 5% differ between women and men? Compute and report the expected frequencies.

Problem L10. Is smoking status associated with the area of residence?

Problem L11. Is smoking status distributed in the same way in the three arms? Compute the test, then explain in (g) how you interpret a baseline difference in a randomized trial.

Problem L12. Did the proportion of patients with fasting plasma glucose ≥ 100 mg/dL change between baseline and 6 months, separately in arm A and in arm B? Choose the appropriate test for paired data and comment on the difference between the two arms.

Problem L13. Is baseline fasting plasma glucose associated with baseline HbA1c? Build the scatter plot and choose the appropriate correlation coefficient.

Problem L14. Did fasting plasma glucose change at 0, 3 and 6 months in arm A? Check the conditions for repeated measures ANOVA (normality, sphericity) and decide which test you set as the main one. Then compare the result with that of the alternative and discuss.

Problem L15. Is age associated with the FINDRISC score? Choose the correlation coefficient and discuss in (g) whether the association may be partly "built in" by the way the FINDRISC score is calculated.

L8.4 Peer Assessment (20 Minutes)

After you have finished problems 1-10, exchange reports with the colleague next to you (by sharing the Google document, with comment rights). Each of you assesses the other's report for three problems chosen by the instructor, using the grid in Table L8.3. Comments are written directly in the document, in the margin of the text. The aim is not to "catch" mistakes, but to find them now, before the examiner does.

Table L8.3. The peer assessment grid (for one problem; 0 = absent or wrong, 1 = partial, 2 = complete and correct).

Criterion	What you check	Score
1. Hypotheses	H0 and H1 are written in words and mathematically, with the correct parameter (μ, Me, p)	0-2
2. Choice of test	the test matches the type of variable and the design; the alternative is mentioned	0-2
3. Checking the conditions	normality by group or on the differences, Levene, expected frequencies, as appropriate, with values	0-2
4. Result	the appropriate description (mean ± SD, or median and IQR, or n and %); the statistic, df, p; the effect size	0-2
5. Chart	the type of chart fits the data; the axes have a title and a unit	0-2
6. Conclusions	the correct statistical decision ("we do not reject", not "we accept"); the clinical conclusion takes into account the effect size and the design	0-2

In addition to the score, the assessor writes one positive comment ("what you did very well") and one concrete suggestion for improvement. After the assessment, each student has 5 minutes to correct their report. The final report keeps the colleague's comments and marks what was corrected.

Questions worth asking during the assessment, from the experience of the lab:

- Did your colleague check normality on the whole sample or in each group?

- For a problem with paired data (L12, L14), did they use a test for paired data?
- Did they report the mean for satisfaction (L7)? If so, the description does not match the test.
- Did they write "there is no difference" after a nonsignificant result, instead of "we did not find a significant difference"?
- In L3, L4 and L11, did they notice that testing baseline differences in a randomized trial is questionable?
- In L14, did they choose the test before seeing the p value, or did they choose the one that "comes out"?

L8.5 Final Discussion and Upload

In the last 10 minutes, the instructor discusses with the whole group three problems in which the choice was most often wrong and one in which the tests gave discordant results (usually L1, L12 or L14). Each student uploads the final report (PDF) to Moodle, with the decision sheet, the 10-12 solved problems and the colleague's comments.

Additional Exercises

1. Using the BPOC file (COPD data) on Moodle, formulate three research questions yourselves: one that requires a test for two independent groups, one for paired data and one for two qualitative variables. Solve them in the a-g format, without looking at the solutions from other sessions.
2. Using the Date_HbA1c file, solve the problem "Does mean HbA1c vary significantly across the years studied?" in two variants: assuming independent measurements (different patients each year) and assuming repeated measurements on the same patients. Explain why the test changes.
3. Problem with data in the statement, with the complete solution. A study compared the duration of postoperative pain (days) after two hernioplasty techniques: technique 1 (n = 6): 2, 3, 3, 4, 5, 9; technique 2 (n = 6): 4, 5, 6, 6, 8, 10. Choose and apply the test.

   Solution. (a) Technique 1 produces a shorter duration of pain. (b) H0: the distributions of duration are the same (Me1 = Me2); H1: the distributions differ. (c) The variable is numerical, but it is a duration in days, in groups of 6 each, with an outlying value (9 and 10 days): Mann-Whitney U; independent groups. (d) Calculation by hand. The combined ordered series: 2, 3, 3, 4, 4, 5, 5, 6, 6, 8, 9, 10. The ranks: 2 → 1; 3 and 3 → 2.5; 4 and 4 → 4.5; 5 and 5 → 6.5; 6 and 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; check $28 + 50 = 78 = \frac{12\times13}{2}$. U1 = 28 - 21 = 7; U2 = 50 - 21 = 29; U = 7. (e) The critical value for n1 = n2 = 6 (Table 7.15) is 5; medians 3.5 and 6 days; $r_{\text{rb}} = 1 - \frac{2\times7}{36} = 0.61$. (f) U = 7 > 5, we do not reject H0 (exact p about 0.09). (g) Pain lasted less after technique 1 (median 3.5 vs. 6 days), with a large estimated effect, but the difference is not statistically significant in 12 patients; a larger study is needed.

What to submit and indicative grading scheme

Upload to Moodle a single PDF file with:

- the decision sheet completed before the analysis (Table L8.1), for all 15 problems (1 point);
- problems L1-L10 solved in the a-g format, each with its chart (5 points, 0.5 points per problem);
- at least two problems from L11-L15, in the same format (1.5 points);

- the assessment grid completed for the colleague's report, with the positive comment and the suggestion for improvement (1.5 points);
- the list of corrections made after the peer assessment (1 point).

Total: 10 points. Errors that cancel the score for a problem: a test for independent groups applied to paired data; a chi-square test reported with expected frequencies below 5 without any discussion; a test chosen after seeing the result (for example, reporting the alternative only because it gives $p < 0.05$); the conclusion "we accept H0".

Offline variant

Without access to Moodle, the data file is copied from the shared folder of the laboratory or from a USB stick, and the report is written in a local text editor and submitted at the next session. The peer assessment is done on printed sheets: the students exchange their reports and fill in the grid in Table L8.3 by hand. If the computers are not available either, the problems are solved on paper starting from the descriptive statistics provided by the instructor (means, standard deviations, medians, frequency tables), and the requirement is limited to components (a), (b), (c), (f) and (g) of the format, plus the hand calculation of a chi-square or Mann-Whitney test chosen by the instructor.

Practical Session 9. Correlation and Regression Analysis: Pearson, Spearman and Simple Linear Regression

Objective

This session turns Chapter 9 into working steps: starting from two variables measured on the same patients, the student draws the correlation diagram, chooses the appropriate coefficient, tests it, estimates the regression line, checks the residuals and formulates the conclusion in the format of the discipline's assignments. We work on the mg2026 dataset, with four different programs, so that you see that the results coincide regardless of the tool and so that you can manage with any of them in the practical exam.

By the end, the student should be able to:

- build a correlation diagram with the regression line, the equation and R^2 displayed;
- compute the Pearson, Spearman and Kendall coefficients with the p value and the confidence interval, in JASP and MedCalc;
- obtain in Google Sheets r, R^2, the slope, the intercept, the standard error of the estimate and a prediction, using the functions CORREL, PEARSON, RSQ, SLOPE, INTERCEPT, FORECAST, STEYX and LINEST;
- run a simple linear regression in JASP, MedCalc and Epi Info and read the table of coefficients;
- check the conditions of the regression on the residual plots;
- write up the solution to a correlation and regression problem in the a-g format.

Tools

Software: JASP 0.98.1 [47]; MedCalc 24.0 [50]; Epi Info 7.2.7 [51]; Google Sheets, with English (United States) regional settings (the decimal separator is the point, and the argument separator in formulas is ",") [43]. Files: mg2026_curat.csv (160 rows, from Moodle, the Emediqual platform [52]); for the additional exercises, the discipline's classic files BD_Studenti and Probleme BPOC (COPD problems). Approximate time: 100 minutes.

In mg2026_curat.csv, the columns used today are (after import into Google Sheets): D = varsta, H = activitate, I = inaltime, J = greutate0, K = imc0, M = tas0, Q = hdl, R = trigliceride, S = glicemie0, V = hba1c0. The data occupy rows 2-161.

Procedure

L9.1 The First Step: The Correlation Diagram

1. Open the file in JASP (File > Open > Computer > mg2026_curat.csv). Check in the data view that imc0, tas0, glicemie0 and hba1c0 are marked as continuous variables (the ruler icon), and activitate and educatie as ordinal. If JASP has read a variable incorrectly, click the icon in the column header and change the type.
2. In JASP, the Descriptives > Descriptive Statistics menu: put tas0 and imc0 in the Variables box; under Plots, check "Scatter plots" (in some versions the option is called "Correlation plots"). You will see the cloud of points with BMI on one axis and SBP on the other. Ask yourselves the four questions from Chapter 9: shape, direction, strength,

outliers. The expected answer: an elongated, increasing, fairly scattered cloud, with a few patients with a BMI above 40 who draw the eye to the right.

3. In Google Sheets: import the file (File > Import > Upload, with the comma as separator; check that the decimals were read correctly). Select columns K and M (BMI and SBP), then Insert > Chart. In the Chart editor, under Chart type, choose Scatter chart. Make sure that BMI is on the X axis (the first column selected). In the Customize > Series tab, check Trendline, under Type choose Linear, under Label choose "Use equation" and check "Show R²". The chart should display an equation of the form y = 1.43x + 89.5 and $R^2 = 0.164$.
4. Repeat the chart for glicemie0 (S) and hba1c0 (V). The cloud is visibly tighter ($R^2 = 0.545$). Compare the two charts: this is what the difference between a weak correlation (r = 0.405) and a moderate-to-strong one (r = 0.738) looks like.

L9.2 Correlation Coefficients in JASP

5. The Regression > Correlation menu (in recent versions, the module is called Regression and contains "Correlation"). Move into the Variables box: imc0, tas0, varsta, glicemie0, hba1c0, hdl, trigliceride.
6. In the Sample Correlation Coefficient section, check Pearson's r, Spearman's rho and Kendall's tau-b. In Additional Options, check "Report significance", "Flag significant correlations" and "Confidence intervals" (95%). Under Plots, check "Scatter plots" and, for the overall picture, "Heatmap".
7. JASP displays a matrix. For the pairs of interest, you should find the values in Table L9.1.

Table L9.1. Self-check values for mg2026 (n = 160).

Pair	Pearson r (95% CI)	p	Spearman rho	Kendall tau-b	Colton
BMI - SBP	0,405 (0,266-0,527)	< 0,001	0,372	0,258	weak
age - SBP	0,256 (0,106-0,396)	0,001	0,174	0,122	weak
blood glucose - HbA1c	0,738 (0,659-0,802)	< 0,001	0,728	0,557	moderate
HDL - triglycerides	0.076 (-0.080 to 0.229)	0,34	0,075	0,050	no correlation
height - weight	0,652 (0,552-0,733)	< 0,001	0,693	0,505	moderate

8. For the ordinal variable activitate, run Spearman with imc0 separately: rho = -0.012, p = 0.88. For educatie and satisfactie: rho = 0.125, p = 0.115. Note that Pearson is not reported for ordinal variables; if you check it, JASP computes it anyway, but its interpretation is not justified.
9. What to write in your notebook: for BMI-SBP, why we report both coefficients (BMI is not normally distributed: Shapiro-Wilk p < 0.001, skewness 1.42), and the fact that the two values (0.405 and 0.372) are close, so the extreme values do not distort the result much.

L9.3 Simple Linear Regression in JASP

10. The Regression > Linear Regression menu. Put tas0 under Dependent Variable and imc0 under Covariates. Method: Enter.

11. Under Statistics, check: Estimates, Confidence intervals (95%), Model fit, Descriptives; under Residuals, check Durbin-Watson and Casewise diagnostics (with the default threshold of 3 standard deviations).
12. Under Plots, check: "Residuals vs. predicted", "Residuals histogram" (with "Standardized residuals") and "Q-Q plot standardized residuals".
13. The Model Summary table should show R = 0.405, $R^2 = 0.164$, R^2 adjusted $= 0.158$ and RMSE = 12.73. In JASP, the RMSE (root mean square error) is the standard error of the estimate (SEE) from the lectures.
14. The ANOVA table: the sum of squares for the regression 5,015.2, for the residuals 25,617.0, total 30,632.2; F(1, 158) = 30.93, p < 0.001.
15. The Coefficients table: the intercept 89.47 (SE 7.56; t = 11.84; CI 74.54-104.40), the slope for imc0 1.435 (SE 0.258; t = 5.56; p < 0.001; CI 0.925-1.944). Check that the t for the slope is the same as the one obtained by hand for r in Chapter 9, and that $F = t^2 (5.56^2 = 30.93)$.
16. The plots: the residuals versus the predicted values should form a horizontal band with no curvature and no funnel shape (compare with Figure 9.5 in the chapter); the histogram and the Q-Q plot should show an approximately normal distribution. If you want a formal test, save the residuals (the option of adding the residuals as a new variable in the data, where the version allows it) and apply Shapiro-Wilk from Descriptives: p = 0.878.
17. Repeat the regression with varsta as the predictor: SBP = 108.28 + 0.406 x age, CI for the slope 0.166-0.646, $R^2 = 0.066$. And with glicemie0 as the predictor of hba1c0: HbA1c = 3.379 + 0.0209 x blood glucose, $R^2 = 0.545$, RMSE = 0.301.

L9.4 Google Sheets: The Same Results Through Functions

18. In a new sheet (or in the free columns to the right of the data), type the labels and formulas below. All of them use the ranges in rows 2-161. Pay attention to the order of the arguments: for SLOPE, INTERCEPT, RSQ, STEYX, FORECAST and LINEST, Y (the dependent variable) is given first, then X; for CORREL and PEARSON the order does not matter.

Table L9.2. Google Sheets functions for correlation and regression (SBP as a function of BMI).

What we compute	Formula	Expected result
Pearson coefficient	=CORREL(K2:K161, M2:M161)	0,405
the same, another function	=PEARSON(K2:K161, M2:M161)	0,405
coefficient of determination	=RSQ(M2:M161, K2:K161)	0,164
slope b	=SLOPE(M2:M161, K2:K161)	1,435
intercept a	=INTERCEPT(M2:M161, K2:K161)	89,47
standard error of the estimate	=STEYX(M2:M161, K2:K161)	12,73
prediction for a BMI of 30	=FORECAST(30, M2:M161, K2:K161)	132,5
prediction for a BMI of 35	=FORECAST(35, M2:M161, K2:K161)	139,7
t statistic for r (r in cell X2)	=X2*SQRT(158)/SQRT(1-X2^2)	5,56
two-sided p value (t in cell X3)	=T.DIST.2T(X3, 158)	about 0.0000001

19. The LINEST function returns all the regression statistics in a block of 5 rows and 2 columns. Select an empty cell that has free space below it and to its right and type =LINEST(M2:M161, K2:K161, TRUE, TRUE). The resulting block is read as follows: row 1, the slope 1.435 and the intercept 89.47; row 2, their standard errors 0.258 and 7.56; row 3, R² = 0.164 and SEE = 12.73; row 4, F = 30.93 and the residual degrees of freedom

158; row 5, the regression sum of squares 5,015.2 and the residual sum of squares 25,617.0. LINEST does not display labels, so type them alongside yourselves.

20. The confidence interval for the slope: if the slope is in cell Y1 and its standard error in Y2, type =Y1-T.INV.2T(0.05, 158)*Y2 and =Y1+T.INV.2T(0.05, 158)*Y2. You should obtain 0.925 and 1.944.
21. Sheets has no dedicated function for Spearman. You proceed as in the definition: in a new column, compute the rank of each BMI with =RANK.AVG(K2, K$2:K$161, 1) and drag the formula down to row 161; do the same for SBP. Then =CORREL on the two columns of ranks gives rho = 0.372. RANK.AVG assigns tied values the mean of their ranks, exactly as in Chapter 9; the plain RANK function does not do this and gives a slightly different result.
22. For activitate and BMI, apply the same procedure and check that rho = -0.012. With many tied values (physical activity has only 5 categories), it is important to use RANK.AVG.

L9.5 MedCalc

23. Import the file (File > Import, or open the CSV file directly, with the comma as the field separator and the point as the decimal separator; check in the data sheet that the values are numerical).
24. Statistics > Correlation > Correlation coefficient: Variable Y = tas0, Variable X = imc0. MedCalc reports the sample size, the coefficient r = 0.405, the significance level P < 0.0001 and the 95% CI for r, 0.266-0.527.
25. Statistics > Correlation > Rank correlation: the same variables. Check both Spearman's rho and Kendall's tau, if the version offers both options. Expected values: rho = 0.372 (95% CI about 0.23-0.50) and tau = 0.258.
26. Statistics > Regression > Regression: Variable Y = tas0, Variable X = imc0, model "y = a + b x". In the options, check the plotting of the residuals. The output includes the coefficient of determination $R^2 = 0.164$, adjusted R^2 0.158, the Residual standard deviation 12.73, the equation y = 89.47 + 1.435 x, the table of coefficients with the standard error, 95% CI, t and P, and the ANOVA table. MedCalc also automatically displays the scatter plot with the regression line; if it does not, use Graphs > Scatter diagram, with the option of displaying the regression line and the confidence band.
27. Statistics > Regression > Regression, with Y = hba1c0 and X = glicemie0: check HbA1c = 3.379 + 0.0209 x blood glucose.

L9.6 Epi Info 7

28. From the main menu, Analyze Data (Classic Analysis). The Read (Import) command: choose the text/CSV file type and load mg2026_curat.csv. Check with List or Frequencies that the variables have been read.
29. Statistics > Linear Regression: Outcome Variable = tas0, Other Variables = imc0. The output displays the correlation coefficient in the form "r^2" = 0.16, the table of coefficients (for imc0: 1.435, standard error 0.258; for the constant: 89.47, standard error 7.56) and an ANOVA table. Instead of the t statistic, Epi Info displays an F test for each coefficient, which is t^2: for imc0, F = 30.93, and for the constant F = 140.1. The p values are the same.
30. Alternatively, in the Visual Dashboard: Add Analysis Gadget > Advanced Statistics > Linear Regression, with the same variables. The results are identical.
31. Epi Info has no separate menu for Spearman in Classic Analysis; for ordinal data, use JASP, MedCalc or the rank procedure in Sheets.

L9.7 Writing Up the Result in the a-g Format

32. Write up the solution to the problem "Is there an association between BMI and SBP at baseline?" following the model below, and compare it with your own version.

(a) Research hypothesis: patients with a higher BMI have, on average, a higher systolic blood pressure.

(b) Statistical hypotheses: H0: ρ = 0 (BMI and SBP are not linearly correlated in the population); H1: ρ ≠ 0.

(c) Method: two numerical variables on the same patients, hence correlation and simple linear regression; X = BMI (predictor), Y = SBP (response). SBP is normal (Shapiro-Wilk p = 0.303), BMI is skewed (p < 0.001), which is why we report both Pearson and Spearman.

(d) Software: JASP, Regression > Correlation and Regression > Linear Regression; check in Google Sheets.

(e) Result: r = 0.405 (95% CI 0.266-0.527), p < 0.001; rho = 0.372, p < 0.001; SBP = 89.47 + 1.435 x BMI (95% CI for the slope 0.925-1.944); $R^2 = 0.164$; SEE = 12.73 mmHg; the scatter plot with the regression line (Figure 9.4 in the chapter).

(f) Statistical conclusion: p < 0.001, H0 is rejected; the correlation is positive and weak according to Colton's rule.

(g) Research conclusion: in the patients at cardiometabolic risk in the study, each additional kg/m² is associated with an SBP higher on average by about 1.4 mmHg. BMI explains only 16.4% of the variation in SBP; 83.6% is due to other factors. The prediction for a BMI of 30 is 132.5 mmHg, and for a BMI of 35 it is 139.7 mmHg; the model should not be used for BMI values outside the range 23.5-48.3 kg/m².

33. Repeat the write-up for the height-weight pair and answer the two classic questions of the session: what proportion of the variation in weight is due to factors other than height ($1 - R^2 = 0.576,$, that is, 57.6%) and what weight is predicted for 1.78 m (93.1 kg) and for 2.35 m (the formula gives 155.0 kg, but this is an extrapolation outside the range 145-192 cm, so the prediction is not credible).

Additional Exercises

Exercises 1-4 are solved on the discipline's classic files on Moodle. We do not give the numerical results: you will compare them with your colleagues and with the instructor during the session.

1. The BD_Studenti file. Investigate the association between weight and height: (1) what type of statistical investigation this is and what its characteristics are; (2) the statistical hypotheses in words and mathematically; (3) the appropriate graphical representation; (4) the statistical processing; (5) what proportion of the variation in weight is caused by factors other than the variation in height; (6) the interpretation in relation to the chart and the hypotheses; (7) what weight a student 1.78 m tall and one 2.35 m tall should have, with a comment on extrapolation.
2. The BD_Studenti file. Compute BMI = weight (kg) / height² (m²) for each student, and the desired BMI (from the desired weight and the same height). Investigate the association between current BMI and desired BMI. Choose the coefficient after checking the normality of the two variables. Discuss separately how you would show, on the same data, whether students on average want a BMI different from their current one (hint: this is a question of paired comparison, Chapter 6, not of correlation; a high correlation does not mean that the values are equal).
3. The Probleme BPOC file (COPD problems). To what extent can the patient's age (years) predict the initial peak expiratory flow (DEM1)? Write the equation of the regression line,

interpret the slope and R², check the residuals and specify the age range in which the equation can be used.

4. The Probleme BPOC file. Is there a correlation between initial systolic blood pressure (TMAX1) and initial heart rate (FC1)? Choose between Pearson and Spearman after checking the conditions, and justify your choice.
5. Problem with data in the statement, with the complete solution. In five patients, fasting plasma glucose (mg/dL) and HbA1c (%) were measured: (90, 5.3), (100, 5.2), (110, 5.8), (125, 5.7), (140, 6.4). Compute r, test it, write the regression line of HbA1c as a function of blood glucose and predict HbA1c for a blood glucose of 120 mg/dL.

Solution. The means: $m_x = \frac{565}{5} = 113$ mg/dL, $m_y = \frac{28.4}{5} = 5.68\%$. The deviations of X: -23, -13, -3, 12, 27; those of Y: -0.38, -0.48, 0.12, 0.02, 0.72. The products: 8.74, 6.24, -0.36, 0.24, 19.44, so Sxy = 34.30. Sxx = 529 + 169 + 9 + 144 + 729 = 1,580. Syy = 0.1444 + 0.2304 + 0.0144 + 0.0004 + 0.5184 = 0.908. $r = \frac{34.30}{\sqrt{1580 \times 0.908}} = \frac{34.30}{37.88} = 0.906$, a strong correlation according to Colton. Test: $t = \frac{0.906 \times \sqrt{3}}{\sqrt{1-0.820}} = \frac{1.569}{0.424} = 3.70$; the critical value for 3 degrees of freedom is 3.182; since 3.70 > 3.182, H0 is rejected (p = 0.034). The slope $b = \frac{34.30}{1580} = 0.0217$ percentage points per mg/dL; the intercept a = 5.68 - 0.0217 x 113 = 3.23. The equation: HbA1c = 3.23 + 0.0217 x blood glucose. For 120 mg/dL: 3.23 + 0.0217 x 120 = 5.83%. $R^2 = 0.82$. Comment: with only five patients, the correlation is only borderline significant, even though r is 0.9; in mg2026 (n = 160) the correlation is smaller (0.738), but estimated much more precisely. Check the result in Sheets with =CORREL, =SLOPE, =INTERCEPT and =FORECAST.

What to submit and indicative grading scheme

Submit on Moodle, in the activity for this session, a single document (PDF or Google Docs link) containing: the BMI-SBP scatter plot with the regression line, the equation and R² (from Sheets or JASP); Table L9.1 completed with the values you obtained; a screenshot of the Coefficients table from JASP and the residual plot; the LINEST block from Sheets, with labels; the a-g write-up for BMI-SBP and for height-weight; and the solution to additional exercise 1 or 3.

Indicative grading scheme (10 points): a correct chart, with the appropriate axes (X predictor, Y response) and the line displayed, 1 point; the Pearson and Spearman coefficients with p and CI, with a justified choice, 2 points; the regression line with the interpretation of the slope and intercept, 2 points; checking the residuals, 1 point; R² and 1 - R² correctly interpreted, 1 point; the prediction and the comment on extrapolation, 1 point; a complete and correct a-g write-up, 2 points. Points are deducted for the wording "H0 is accepted", for interpreting correlation as causation and for missing units of measurement for the slope.

Offline variant

If you have no access to the internet or to the software, the session can be done entirely with a pocket calculator and a printed extract from mg2026 (the first 20 patients, the columns imc0 and tas0, provided by the instructor). The steps are those of Worked Problem 9.1: the table with the deviations, products and squares, Sxy, Sxx, Syy, then r, t (compared with the critical value from the table of the t distribution handed out in the session), b, a, a prediction and R². The correlation diagram is drawn on graph paper, with the line drawn through the point of the means (mx, my) and through a second point calculated from the equation. The Spearman coefficient is calculated with ranks, following Table 9.4. The results for 20 patients will differ from those for 160: compare them and explain why the confidence interval for r is wider.

Practical Session 10. Epidemiological Analyses: Relative Risk, Odds Ratio and the Interpretation of Observational Studies

Objective

This session applies the methods of Chapter 10 to the discipline's classic problems and to the mg2026 dataset. The emphasis is on three skills that are tested in the practical exam: turning a problem statement into the correct 2 x 2 table, choosing the measure of association appropriate to the design (RR, prevalence ratio or OR), and interpreting the confidence interval. The calculations are done by hand once, then checked in Epi Info, MedCalc, JASP and Google Sheets.

By the end, the student should be able to:

- recognize the type of study from a problem statement and build the 2 x 2 table with the a, b, c, d notation;
- compute the RR, the OR, their 95% confidence intervals and the chi-square or Fisher test, by hand and in Epi Info StatCalc;
- use the relative risk and odds ratio calculators in MedCalc and the contingency tables in JASP;
- build in Google Sheets a reusable spreadsheet for any 2 x 2 table;
- estimate the impact at the population level: the attributable risk, the attributable fraction and the NNT or NNH;
- formulate the statistical and the clinical conclusion for an observational study, with its limitations.

Tools

Software: Epi Info 7.2.7 [51], the StatCalc and Classic Analysis modules; MedCalc 24.0 [50]; JASP 0.98.1 [47]; Google Sheets with English (United States) regional settings (decimal point, "," as the argument separator) [43]. Files: mg2026_curat.csv from Moodle [52]; for the additional exercises, the Probleme BPOC file (COPD problems). Approximate time: 100 minutes.

Procedure

L10.1 From the Problem Statement to the 2 x 2 Table

1. For each problem, before any calculation, answer four questions in writing: What is the exposure (the risk factor)? What is the outcome (the disease)? Where does the study start: from the exposure, from the disease or from a "snapshot" of the population? Which totals are fixed by the design?
2. Build the table with the exposure in the rows (exposed at the top) and the disease in the columns (diseased on the left). First fill in the totals given in the statement, then the cells given directly, then the cells obtained by subtraction. Finally, check that a + b + c + d = N.
3. Practice on the four classic problem statements. Smoking-COPD: 1,200 healthy people followed for 5 years, 300 smokers, 50 smokers and 70 nonsmokers with COPD; a cohort study, so a = 50, b = 250, c = 70, d = 830. Monitor-visual acuity: 1,000 people, 200 exposed for at least 8 hours a day, of whom 90 have reduced visual acuity, and 150 with reduced visual acuity among the 800 unexposed; cross-sectional, a = 90, b = 110, c = 150, d = 650. Oral contraceptives-rheumatoid arthritis: 100 cases and 200 controls, with 40 cases and 120 controls exposed; case-control, a = 40, b = 120, c = 60, d = 80. Dust-lung

capacity: 320 retirees, 80 exposed, 30 with reduced capacity in each group; a = 30, b = 50, c = 30, d = 210.

4. Mistakes to check for: in the dust problem, "30 in both groups" does not mean equal risk (30 of 80 versus 30 of 240); in the smoking-COPD problem, the number of nonsmokers is 900, not 1,200; in the case-control problem, the fixed totals are in the columns (100 cases and 200 controls), not in the rows.

L10.2 Epi Info StatCalc: Tables (2 x 2 x n)

5. From the Epi Info main menu, choose StatCalc, then Tables (2 x 2 x n). A table appears with the rows Exposed + / Exposed - and the columns Outcome + / Outcome - (the labels may differ slightly depending on the version). For smoking-COPD, enter: 50, 250, 70, 830, in the order a, b, c, d.
6. The results appear immediately, in two blocks. The block of measures of association contains the Odds ratio 2.37 with 95% CI 1.61-3.50 (the Taylor series method, equivalent to the logarithmic method from the lectures), a maximum likelihood estimate of the OR (MLE Odds Ratio, about 2.37, with a slightly wider exact interval), the relative risk (Risk Ratio) 2.14 with 95% CI 1.53-3.01, the Risk Difference 8.89% with 95% CI about 4.3-13.5%, and the etiologic fractions (in the exposed 53.3%, in the population 22.2%). The block of tests contains the uncorrected chi-square 19.75, the Mantel-Haenszel chi-square 19.74 and the Yates-corrected chi-square 18.78, each with its p value (all below 0.0001), plus the exact tests (mid-p and Fisher).
7. The Mantel-Haenszel chi-square from a single table is the ordinary chi-square multiplied by (N - 1) / N: $\frac{19.753 \times 1199}{1200} = 19.737$. The difference matters only for small samples.
8. Enter the small version (5, 25, 7, 83). Note: RR 2.14 (95% CI 0.73-6.25), OR 2.37 (95% CI 0.69-8.13), uncorrected chi-square 1.98 (p = 0.16), Fisher exact two-sided p = 0.17. StatCalc warns that an expected frequency is below 5 (the expected frequency for cell a is 3) and recommends the exact test. Conclusion: the same measures of association, but a nonsignificant result.
9. Enter the other classic problems one by one and complete Table L10.1 in section L10.8. For the ECG-CHD problem (8, 39, 10, 134), note the borderline result: uncorrected chi-square p = 0.040, Yates p = 0.077, Fisher p = 0.0485, 95% CI for the RR 1.03-5.85.
10. Stratified analysis. StatCalc allows several strata to be entered (hence the "x n" in the name): after the first table, move on to the next stratum. Enter the obesity-diabetes association from mg2026 by sex: women 10, 19, 5, 52; men 8, 17, 6, 43. The summary result displays the crude OR (Crude) 4.32 and the Mantel-Haenszel adjusted OR 4.31, with 95% CI about 1.86-10.0. Since the adjusted OR is practically equal to the crude OR, sex is not a confounder. The stratum-specific OR values (5.47 in women, 3.37 in men) differ, but the test of homogeneity (Breslow-Day, where displayed) does not show a significant difference.

L10.3 Epi Info Classic Analysis on the Raw mg2026 Data

11. Analyze Data > Classic. The Read command: import mg2026_curat.csv.
12. Create the exposure variable. The Define command: obez (numeric type). Then Assign (or Recode): obez = 1 if imc0 ≥ 30, otherwise 0. In Classic Analysis you can write, for example, IF imc0 >= 30 THEN obez = 1 ELSE obez = 0 END. Check with Frequencies obez: 54 values of 1 and 106 values of 0.

13. Statistics > Tables: Exposure Variable = obez, Outcome Variable = diabet_ogtt. Epi Info displays the 2 x 2 table and the same statistics as StatCalc. Check the order of the rows and columns: the top left cell must be "obez = 1 and diabetes = 1", that is, 18. If the program has put the 0 categories first, the OR displayed will be the reciprocal (1 / 4.32 = 0.23); in this case, recode the variables as Yes/No or check the display options.
14. The expected values: OR 4.32 (95% CI 1.86-10.03), RR (here the prevalence ratio, since this is a cross-sectional analysis) 3.21 (95% CI 1.64-6.31), uncorrected chi-square 12.70 (p = 0.0004), Yates 11.20 (p = 0.0008), Fisher exact p = 0.0008.

L10.4 MedCalc: The Relative Risk and Odds Ratio Calculators

15. In MedCalc, the Tests menu contains calculators that work directly with the frequencies, without a data file. Choose Relative risk (in some versions "Relative risk & Number needed to treat"). For the exposed group, enter the number of events (50) and the group total (300), and for the comparison group 70 and 900.
16. MedCalc reports the relative risk 2.14, 95% CI 1.53-3.01, the statistic z = 4.41 and the significance level P < 0.0001. The z statistic is $\frac{\ln(\text{RR})}{\text{SE}[\ln(\text{RR})]} = \frac{0.762}{0.1727} = 4.41$, exactly the test built on the logarithmic scale from the lectures. If the NNT is also requested, MedCalc displays it as the number needed to produce an unfavorable effect (NNT harm or NNH): 11.25, with 95% CI about 7.4-23.1.
17. Tests > Odds ratio: for smoking-COPD, enter the numbers of exposed and unexposed among the cases and among the non-cases (a = 50, c = 70 among the diseased; b = 250, d = 830 among the healthy, in the corresponding fields of the calculator). Result: OR 2.37, 95% CI 1.61-3.50, z = 4.35, P < 0.0001. Read the field labels carefully: the OR calculator in MedCalc is designed for case-control studies and asks for the frequencies by column (cases, controls), not by row.
18. Repeat for contraceptives-arthritis (cases: 40 exposed, 60 unexposed; controls: 120 exposed, 80 unexposed): OR 0.44, 95% CI 0.27-0.73, P = 0.001. Do not use the RR calculator for this problem: the case-control design does not allow an RR.
19. For the mg2026 randomized trial, the relative risk calculator with 16 events of 54 (arm C) and 6 of 53 (arm A) gives RR 2.62 (95% CI 1.11-6.17), P = 0.028, and NNT (benefit) 5.5, with 95% CI about 3.0-29.1. Compare with the value from the lectures. Note that the P from the MedCalc calculator (the z test on the logarithmic scale) differs slightly from the p of the chi-square test (0.019); both are below 0.05.

L10.5 JASP: Contingency Tables with the Odds Ratio

20. Open mg2026_curat.csv in JASP. Create a computed variable: click the "+" sign next to the last column, name obez, nominal type, defined by the formula ifelse(imc0 >= 30, 1, 0) (the R editor or the drag-and-drop builder). Check the frequencies: 54 and 106.
21. The Frequencies > Contingency Tables menu. Rows: obez; Columns: diabet_ogtt. Under Statistics, check χ^2, "χ^2 continuity correction", "Log odds ratio (2 x 2 only)" with the 95% confidence interval and, if available, Fisher's exact test. Under Cells, check Expected counts and Row percentages.
22. Read the results: the expected frequencies 9.8, 44.2, 19.2, 86.8 (all above 5, so the chi-square test is valid); $\chi^2 = 12.70$, p < 0.001; with correction 11.20, p = 0.0008. JASP reports the logarithm of the odds ratio: ln(OR) = 1.463, with 95% CI 0.620-2.305. The conversion to the OR is done by exponentiation: $e^{1.463} = 4.32$; $e^{0.620} = 1.86$; $e^{2.305} = 10.03$.

23. Pay attention to the order of the categories. JASP orders the categories alphabetically or numerically (0 before 1), so in the table displayed the cell "obez = 0, diabetes = 0" may appear at the top left. The odds ratio stays the same if both the rows and the columns are reversed, but it becomes the reciprocal (0.23, with ln = -1.463) if only one of them is reversed. Always check the sign of the logarithm against the observed table. In JASP you can change the order of the levels in the variable window (click on the column header).
24. JASP works with individual data. For a problem given only through frequencies (for example, monitor-visual acuity), create a small file with four rows and three columns: exposure, outcome and frequency (90, 110, 150, 650), then in Contingency Tables put the frequency in the Counts box.

L10.6 Google Sheets: A Spreadsheet for Any 2 x 2 Table

25. In a new sheet, put a in B2, b in C2, c in B3, d in C3 (for example, 50, 250, 70, 830). Then type the formulas in Table L10.2. After checking the results on smoking-COPD, change only the four cells and the sheet recalculates everything for any problem.

Table L10.2. Google Sheets formulas for the analysis of a 2 x 2 table (a in B2, b in C2, c in B3, d in C3).

Cell	What it computes	Formula	Smoking-COPD
F2	risk in the exposed R1	=B2/(B2+C2)	0,1667
F3	risk in the unexposed R0	=B3/(B3+C3)	0,0778
F4	RR	=F2/F3	2,14
F5	SE[ln(RR)]	=SQRT(1/B2-1/(B2+C2)+1/B3-1/(B3+C3))	0,1727
F6	lower limit of the CI for the RR	=EXP(LN(F4)-1.96*F5)	1,53
F7	upper limit of the CI for the RR	=EXP(LN(F4)+1.96*F5)	3,01
F8	OR	=(B2*C3)/(C2*B3)	2,37
F9	SE[ln(OR)]	=SQRT(1/B2+1/C2+1/B3+1/C3)	0,1987
F10	lower limit of the CI for the OR	=EXP(LN(F8)-1.96*F9)	1,61
F11	upper limit of the CI for the OR	=EXP(LN(F8)+1.96*F9)	3,50
F12	risk difference (AR)	=F2-F3	0,0889
F13	NNT or NNH	=ROUNDUP(1/ABS(F12), 0)	12
F14	attributable fraction in the exposed	=(F4-1)/F4	0,533
F15	proportion of exposed pe	=(B2+C2)/SUM(B2:C3)	0,25
F16	population attributable fraction	=F15*(F4-1)/(1+F15*(F4-1))	0,222

26. The chi-square test in Sheets requires the table of expected frequencies. In H2:I3, type the expected frequencies: in H2 =(B2+C2)*(B2+B3)/SUM($B$2:$C$3), then the analogous formulas for the other three cells (row total x column total / N). Check: 30, 270, 90, 810. Then =CHISQ.TEST(B2:C3, H2:I3) directly returns the p value (about 0.0000088 for smoking-COPD). The value of the statistic is obtained with =CHISQ.INV.RT(p, 1) or with the formula Σ (O - E)² / E: 19.75. CHISQ.TEST does not apply the Yates correction.

27. For a case-control study, the sheet also computes the RR, but the result is meaningless. Add a warning cell (for example, a drop-down list "design" with the values cohort, cross-sectional, case-control and a formula =IF(K1="case-control", "RR is not interpreted", F4)).

L10.7 Estimating Risk at the Population Level

28. For smoking-COPD, complete the interpretation sentences with the values from the spreadsheet. Attributable risk: "smoking adds about 8.9 cases of COPD per 100 smokers over 5 years". NNH: "for every 12 smokers, one additional case occurs". Attributable fraction in the exposed: "about 53% of the COPD cases in smokers are attributable to smoking". Population attributable fraction: "about 22% of all cases in this population would be avoidable without smoking". All these sentences assume that the relationship is causal and unconfounded.
29. For mg2026, the obesity-diabetes association: the proportion of obese patients pe = 54 / 160 = 0.3375; PR = 3.21; population attributable fraction = 0.43. The prevalence of diabetes in the whole sample (18.1%) minus the prevalence in the non-obese (10.4%) gives the population attributable risk, 7.7 percentage points; divided by 18.1%, it again gives 0.43. Formulate the sentence with "if", because the data are cross-sectional.
30. For the mg2026 randomized trial (arm C vs. A, weight loss ≥ 5%): absolute difference 18.3 percentage points (95% CI 3.4-33.2), NNT 5.5, rounded up to 6 (95% CI 3.0-29.1). Interpretation: "by treating 6 patients with the full package instead of standard counseling, we obtain, on average, one more patient with relevant weight loss". Here the causal interpretation is justified by randomization.

L10.8 The Classic Problems: A Self-Check Table

31. Fill in the columns yourselves, then compare with Table L10.1. All intervals are calculated by the logarithmic method.

Table L10.1. Self-check results for the classic problems and for mg2026.

Problem (a, b, c, d)	Design	Main measure (95% CI)	OR (95% CI)	Test	Conclusion
Smoking-COPD (50, 250, 70, 830)	cohort	RR 2.14 (1.53-3.01)	2,37 (1,61-3,50)	χ^2 19.75, p < 0.001	risk factor
Smoking-COPD, small (5, 25, 7, 83)	cohort	RR 2.14 (0.73-6.25)	2,37 (0,69-8,13)	Fisher p = 0.17	nonsignificant
Monitor-visual acuity (90, 110, 150, 650)	cross-sectional	PR 2.40 (1.94-2.96)	3,55 (2,55-4,93)	χ^2 60.44, p < 0.001	risk factor
Contraceptives-RA (40, 120, 60, 80)	case-control	OR only	0,44 (0,27-0,73)	χ^2 10.71, p = 0.001	protective factor
ECG-CHD death (8, 39, 10, 134)	cohort	RR 2.45 (1.03-5.85)	2,75 (1,02-7,44)	Fisher p = 0.0485	borderline
Dust (30, 50, 30, 210)	cross-sectional	PR 3.00 (1.94-4.65)	4,20 (2,32-7,60)	χ^2 24.62, p < 0.001	risk factor
Age, unexposed (30, 210, 12, 188)	cross-sectional	PR 2.08 (1.10-3.96)	2,24 (1,11-4,50)	χ^2 5.34, p = 0.021	risk factor
Obesity-diabetes mg2026 (18, 36, 11, 95)	cross-sectional	PR 3.21 (1.64-6.31)	4,32 (1,86-10,03)	χ^2 12.70, p < 0.001	risk factor

Problem (a, b, c, d)	Design	Main measure (95% CI)	OR (95% CI)	Test	Conclusion
Arm C vs. A mg2026 (16, 38, 6, 47)	randomized	RR 2.62 (1.11-6.17)	3,30	χ^2 5.49, p = 0.019	effective intervention

32. For the monitor-visual acuity problem, also answer the classic question "what are the odds of reduced visual acuity in those who work less than 8 hours a day": 150 / 650 = 0.231. For the contraceptives problem, explain in two sentences why the RR is not calculated.

L10.9 Interpreting an Observational Study

33. Choose one of the colleagues' articles discussed in Chapter 10 (for example, the cohort of 110 female patients with RR = 2.75, 95% CI 1.34-5.64 [114], or the 1:2 matched case-control study of 435 patients [116]) and complete a reading sheet with the following headings: the design and the argument in the text that identifies it; the exposure and the outcome; the measure of association and whether it is appropriate to the design; the confidence interval and its interpretation; the confounders taken into account (if the abstract mentions them); which Bradford Hill criteria are supported; and what else should be reported according to STROBE [25].
34. Discuss in the group: can an adjusted OR of 7.20 (95% CI 4.26-12.17), obtained in a cross-sectional study in which the outcome has a prevalence of 24.1% [4], be translated as "a 7 times higher risk"? The answer is in Table 10.4 of the chapter.

Additional Exercises

1. The Probleme BPOC file (COPD problems; no numerical results given). Do patients over 50 years of age have a higher risk of dyspnea at the start of the study (DISPN = 1) than those under 50? Create the age group variable, build the 2 x 2 table, specify the design (the baseline data are cross-sectional), compute the PR and OR with 95% CIs in Epi Info and in Sheets, apply the appropriate test and formulate the conclusions in the a-g format.
2. The Probleme BPOC file (no numerical results given). Is the presence of rales associated with the age group? Compute the OR and check whether the interval contains 1. Compare the result of the chi-square test with that of Fisher's test and explain any differences.
3. Problem with data in the statement, with the complete solution. During an influenza season, a family practice follows 1,000 patients with chronic diseases: 400 vaccinated against influenza, of whom 20 develop confirmed influenza, and 600 unvaccinated, of whom 72 develop influenza. Analyze the data.

Solution. Design: cohort (observational; the patients chose whether to be vaccinated). The table: a = 20 (vaccinated, diseased), b = 380, c = 72 (unvaccinated, diseased), d = 528. The risks: $R_1 = \frac{20}{400} = 5.0\%$; $R_0 = \frac{72}{600} = 12.0\%$. $\text{RR} = \frac{5.0}{12.0} = 0.42$; $\text{SE}[\ln(\text{RR})] = \sqrt{\frac{1}{20} - \frac{1}{400} + \frac{1}{72} - \frac{1}{600}} = \sqrt{0.0597} = 0.244$; $\text{CI}\,95\% = \exp(-0.875 \pm 0.479) = 0.26 - 0.67$. $\text{OR} = \frac{20 \times 528}{380 \times 72} = \frac{10\,560}{27\,360} = 0.39(;\ \text{CI}\,95\%\ 0.23 - 0.64)$. $\chi^2 = 14.08$, $p < 0.001$. Vaccination appears as a protective factor: the risk of influenza is 58% lower in the vaccinated (estimated vaccine effectiveness = 1 - RR = 58%). The absolute risk difference is 7.0 percentage points (95% CI 3.6-10.4), and $\text{NNT} = \frac{1}{0.07} = 14.3$, rounded up to 15 (95% CI about 10-28): 15 patients must be vaccinated to prevent one case of influenza in one season. The main limitation: the study is observational, and patients who choose to be vaccinated may also have other protective behaviors (the "healthy vaccinee" effect), so the estimate may be optimistic.

What to submit and indicative grading scheme

Submit on Moodle a document containing: Table L10.1 completed with your own values (from Epi Info or Sheets) for at least five problems; a screenshot of StatCalc for smoking-COPD and for the small version; the Google Sheets sheet from L10.6 (a link with view rights); the JASP output for obesity-diabetes, with the conversion of the log OR into the OR; the interpretation sentences from L10.7; the reading sheet from L10.9 for one article; and the solution to additional exercise 1.

Indicative grading scheme (10 points): 2 x 2 tables built correctly and the design identified, 2 points; RR and OR calculated and chosen correctly according to the design, 2 points; the confidence intervals and the appropriate test (chi-square or Fisher), 2 points; the impact measures (AR, NNT/NNH, attributable fraction) with interpretation, 1 point; a functional spreadsheet, 1 point; the critical reading sheet, 1 point; a cautiously worded clinical conclusion, with the limitations of the design, 1 point. Points are deducted for calculating the RR in a case-control study, for using the grand total as the denominator of the risk in the unexposed, and for the conclusion "the factor has no effect" drawn from a nonsignificant result.

Offline variant

All the classic problems can be solved with a pocket calculator, if it has the ln and e^x functions. The steps are: the 2 x 2 table and checking the totals; R1, R0, RR and OR; the SE on the logarithmic scale (the formulas in Chapter 10, section 10.7); the interval on the logarithmic scale and the exponentiation of its limits; the expected frequencies and $\chi^2 = \frac{\sum(O-E)^2}{E}$, compared with the critical value 3.84 for 1 degree of freedom and α = 0.05; the AR, the NNT or NNH, and the attributable fractions. When an expected frequency is below 5, Fisher's exact test cannot practically be done by hand; note that it would be the appropriate test and base the conclusion on the confidence interval, mentioning that it is approximate. At the next session, compare the results obtained by hand with those from Epi Info.

Practical Session 11. Bayesian Analysis: Probabilistic Reasoning in Diagnosis and Screening

Objective

This session turns the theory in Chapter 11 into tools you can use after the exam as well: a reusable Bayesian calculator in Google Sheets, a natural frequency tree, a differential diagnosis table and your first Bayesian analyses in JASP. All calculations are done first by hand, on paper, then in the software, so that you can check yourselves.

By the end, the student should be able to:

- calculate the post-test probability from the pre-test probability, sensitivity and specificity, both through Bayes' theorem and through the odds form;
- build in Google Sheets a Bayesian calculator and a natural frequency tree that update automatically when the prevalence changes;
- use the Fagan nomogram for a quick graphical estimate;
- solve a differential diagnosis table with three diseases and three clinical signs;
- obtain from the mg2026 dataset the 2 x 2 table of the FINDRISC score and calculate the post-test probabilities for different screening contexts;
- run in JASP a Bayesian t-test and a Bayesian contingency table and interpret the BF10 factor.

Tools

Google Sheets (English (United States) regional settings: decimal point and "," as the argument separator), JASP 0.98.1 [47], optionally MedCalc 24.0 [50]. The file mg2026_curat.csv from Moodle (160 rows). A printed Fagan nomogram (Figure 11.3 in the chapter) and a ruler. Approximate time: 100 minutes.

In the file mg2026_curat.csv, imported into Google Sheets with row 1 as the header, the columns used in this session are: C (sex), K (imc0), M (tas0), S (glicemie0), W (findrisc) and X (diabet_ogtt, 1 = diabetes confirmed by the OGTT, 0 = no diabetes). The data occupy rows 2-161.

Procedure

L11.1 The Bayesian Calculator in Google Sheets (20 Minutes)

Open a new sheet and name it "Bayes". Build the calculator in columns A (labels) and B (values), exactly as in the table below. Cells B1-B3 are inputs; the rest are formulas.

Table L11.1. Structure of the Bayesian calculator.

Cell	Label (column A)	Content of column B
B1	Pre-test probability	0,181
B2	Sensitivity	0,759
B3	Specificity	0,748
B4	LR+	=B2/(1-B3)
B5	LR-	=(1-B2)/B3
B6	Pre-test odds	=B1/(1-B1)
B7	Post-test odds (positive test)	=B6*B4
B8	Post-test probability (positive test)	=B7/(1+B7)
B9	Post-test odds (negative test)	=B6*B5
B10	Post-test probability (negative test)	=B9/(1+B9)

Cell	Label (column A)	Content of column B
B11	PPV by Bayes' formula (check)	=B2*B1/(B2*B1+(1-B3)*(1-B1))
B12	NPV by Bayes' formula (check)	=B3*(1-B1)/(B3*(1-B1)+(1-B2)*B1)

Format B8, B10, B11 and B12 as a percentage with one decimal place (Format > Number > Percent). With the values for FINDRISC ≥ 12 from the mg2026 dataset, you should obtain: LR+ 3.01; LR- 0.32; pre-test odds 0.221; post-test probability 40.0% after a positive score and 6.7% after a negative one. B11 must equal B8 (40.0%), and B12 must equal 1 - B10 (93.3%). If the two routes do not give the same result, there is a mistake in one of your formulas.

Now use the calculator as a tool. Change only B1 and complete Table L11.2 (the correct results are those in Table 11.7 of the chapter).

Table L11.2. Post-test probabilities for FINDRISC ≥ 12, to be completed.

Pre-test probability (B1)	Post-test, positive (B8)	Post-test, negative (B10)
0,02	5,8%	0,7%
0,05		
0,10		
0,30		
0,50		

For automatic tabulation, type "pre-test" in D1, "post +" in E1 and "post -" in F1. In D2:D7 enter 0.02, 0.05, 0.10, 0.181, 0.30, 0.50. In E2 type =D2/(1-D2)*$B$4/(1+D2/(1-D2)*$B$4), and in F2 =D2/(1-D2)*$B$5/(1+D2/(1-D2)*$B$5), then drag the formulas down to row 7. The $ sign fixes the reference to the LR, so that it does not shift when copied. Select D1:F7 and insert a line chart (Insert > Chart, Line chart): you have reproduced part of Figure 11.2.

Check by hand (compulsory, in your notebook): for a pre-test probability of 0.10, the odds are 0.111; 0.111 x 3.01 = 0.334; 0.334/1.334 = 25.1%.

L11.2 The Natural Frequency Tree (15 Minutes)

In the same sheet, starting at cell A15, build a natural frequency tree for a population of 1,000 people, using B1-B3.

Table L11.3. The natural frequency tree in Google Sheets.

Cell	Label	Formula
B15	Population	1000
B16	Diseased	=B15*B1
B17	Non-diseased	=B15-B16
B18	Diseased with a positive test (TP)	=B16*B2
B19	Diseased with a negative test (FN)	=B16-B18
B20	Non-diseased with a positive test (FP)	=B17*(1-B3)
B21	Non-diseased with a negative test (TN)	=B17-B20
B22	Total positive tests	=B18+B20
B23	PPV from the tree	=B18/B22
B24	NPV from the tree	=B21/(B19+B21)

Step 1. Set B1 = 0.10, B2 = 0.80, B3 = 0.90. You should obtain 100 diseased, 80 TP, 20 FN, 900 non-diseased, 90 FP, 810 TN, 170 positive tests, PPV 47.1% and NPV 97.6%, that is, exactly Figure 11.1 in the chapter.

Step 2. Keep Se and Sp and change the prevalence to 0.01 and then to 0.50. Write down the PPV and NPV. At 1% you should obtain 8 TP and 99 FP (PPV 7.5%); at 50%, 400 TP and 50 FP (PPV 88.9%).

Step 3. Write in your notebook, in two sentences, how you would explain a positive result to a patient in each of the three situations, using whole numbers ("of 170 people with a positive test,

80 have the disease"). The exercise does not have a single correct answer, but it will be discussed at the end.

L11.3 The Fagan Nomogram (10 Minutes)

Use the printed nomogram (Figure 11.3 in the chapter) and a ruler. For each situation, draw a line from the pre-test point through the LR value, read the post-test probability and then check it with the calculator from L11.1 (enter the LR value directly in B4 or B5, instead of the formula, or calculate it separately).

Table L11.4. Exercises with the Fagan nomogram.

Situation	Pre-test	LR	Post-test (nomogram)	Post-test (calculation)
FINDRISC ≥ 12, mg2026 sample	18,1%	3,01	about 40%	40,0%
FINDRISC < 12, mg2026 sample	18,1%	0,32		6,7%
Highly specific confirmation test	10%	10		52,6%
Rule-out test, high suspicion	50%	0,1		9,1%
Elastography above the threshold for confirming cirrhosis [3]	28,1%	19,7		88,5%

The differences between the reading and the calculation should be a few percentage points. Notice on the nomogram why the same LR shifts the probability a lot in the middle range and only a little at the extremes.

L11.4 The Differential Diagnosis Table with Three Diseases (20 Minutes)

Scenario (fictitious but plausible data). A 34-year-old female patient has mild microcytic anemia on her complete blood count. The family physician considers three diagnoses: iron deficiency anemia (prior probability 0.60), thalassemia minor (0.15) and anemia of chronic disease (0.25). Three results are available: ferritin below 30 ng/mL (F), increased red cell distribution width (RDW) (R) and a Mentzer index (MCV/red blood cell count) below 13 (M). The frequency of each result in each disease is given below.

Table L11.5. Prior probabilities and frequencies of the results (fictitious data).

Diagnosis	p(D)	$p(F/D)$	$p(R/D)$	$p(M/D)$
Iron deficiency anemia	0,60	0,90	0,80	0,15
Thalassemia minor	0,15	0,05	0,20	0,80
Anemia of chronic disease	0,25	0,10	0,40	0,20

The patient has normal ferritin (F-), normal RDW (R-) and a Mentzer index below 13 (M+).

Step 1. In your notebook, calculate the likelihood of the observed combination in each disease (for an absent result, use 1 - p): iron deficiency 0.10 x 0.20 x 0.15 = 0.003; thalassemia 0.95 x 0.80 x 0.80 = 0.608; chronic disease 0.90 x 0.60 x 0.20 = 0.108.

Step 2. In Google Sheets, in a new sheet "Differential", enter Table L11.5 in A1:E4 (header in row 1, the diseases in rows 2-4). In G1, H1 and I1, type the observed results coded as 1 (present) or 0 (absent): G1 = 0 (F), H1 = 0 (R), I1 = 1 (M).

Step 3. In F2, type the formula for the combined likelihood:

=IF($G$1=1, C2, 1-C2)*IF($H$1=1, D2, 1-D2)*IF($I$1=1, E2, 1-E2)

In G2, the product with the prior probability: =B2*F2. In H2, the posterior probability: =G2/SUM($G$2:$G$4). Copy F2:H2 down to row 4.

Check results: products 0.0018, 0.0912, 0.0270; sum 0.1200; posterior probabilities: iron deficiency anemia 1.5%, thalassemia minor 76.0%, anemia of chronic disease 22.5%.

Step 4. Change the coding in G1:I1 for a patient with low ferritin, increased RDW and a Mentzer index above 13 (1, 1, 0). Which diagnosis becomes the most likely? Write down the result. (You should obtain about 97.8% for iron deficiency anemia.)

Step 5. Discussion: what would happen if the patient actually had sideroblastic anemia, which is not in the table? Which assumption of the model would be violated if ferritin and RDW were correlated within the same disease?

The table you have obtained is a working "naive Bayes" model. You can reuse it for any differential diagnosis problem by changing the names of the diseases, the prior probabilities and the frequencies of the signs; Worked Problem 11.5 (chest pain) and exercise 7 in Chapter 11 (headache) can be checked immediately.

L11.5 Screening with FINDRISC on the mg2026 Dataset (15 Minutes)

Import mg2026_curat.csv into a new sheet "Data" (File > Import > Upload, with "Replace current sheet" or "Insert new sheet"). Check that W contains the FINDRISC score and X the OGTT diagnosis.

Step 1. Obtain the 2 x 2 table for the threshold ≥ 12 with the COUNTIFS function, in a sheet "Screening":

- TP: =COUNTIFS(Data!X2:X161, 1, Data!W2:W161, ">=12") should give 22
- FN: =COUNTIFS(Data!X2:X161, 1, Data!W2:W161, "<12") should give 7
- FP: =COUNTIFS(Data!X2:X161, 0, Data!W2:W161, ">=12") should give 33
- TN: =COUNTIFS(Data!X2:X161, 0, Data!W2:W161, "<12") should give 98

Step 2. Calculate the prevalence (29/160 = 18.1%), Se (75.9%), Sp (74.8%), PPV (40.0%) and NPV (93.3%), then copy Se and Sp into the calculator from L11.1.

Step 3. Make the threshold variable: type the threshold in a cell (for example, B1 = 12) and replace ">=12" with ">="&B1 and "<12" with "<"&B1. Try the thresholds 10, 14 and 16. You should find the values in Table 12.5 of Chapter 12 (for example, at ≥ 14: TP 17, FP 16, FN 12, TN 115).

Step 4. For the threshold ≥ 12, calculate the post-test probability for three fictitious patients: a 30-year-old woman without risk factors (pre-test 2%), an overweight 55-year-old man (10%) and an obese female patient with known prediabetes (30%). Formulate the recommendation for each: reassurance, fasting plasma glucose or an OGTT.

Step 5 (sequential testing). For a patient with FINDRISC ≥ 12 and fasting plasma glucose ≥ 110 mg/dL, multiply the likelihood ratios successively (FINDRISC LR+ 3.01; blood glucose LR+ 2.78). You obtain a post-test probability of about 64.9%. Then count directly in the data: =COUNTIFS(Data!X2:X161, 1, Data!W2:W161, ">=12", Data!S2:S161, ">=110") gives 20, and the same formula with X = 0 gives 19. The observed probability is 20/39 = 51.3%. Explain the difference in two sentences (hint: section 11.5.2).

L11.6 Bayesian Analyses in JASP (20 Minutes)

Open mg2026_curat.csv in JASP. Check the variable types: sex nominal, tas0 and imc0 continuous (scale), diabet_ogtt nominal.

Step 1. The Bayesian independent-samples t-test. From the T-Tests menu, in the Bayesian section, choose Independent Samples T-Test. Move tas0 to Variables and sex to Grouping Variable. Leave the alternative hypothesis two-sided and the default prior (Cauchy, width 0.707). Under Plots, check Prior and posterior and Bayes factor robustness check. Expected result: BF10 about 0.18 (if you choose to display BF01, about 5.7). Interpretation: moderate evidence for the absence of a difference in SBP between the sexes. Compare with the frequentist t-test from the practical session on t-tests (p = 0.81): the p value did not allow us to claim the absence of a difference, whereas the Bayes factor moderately supports it.

Step 2. The robustness plot shows how BF10 changes when the width of the prior varies. If the conclusion (evidence for H0) remains the same for all reasonable widths, the result is robust. Write down how BF10 changes for the "wide" and "ultrawide" priors.

Step 3. The Bayesian contingency table. First create the variable obez: in JASP, use the "+" button next to the column names to add a computed column (Computed column), with the formula imc0 >= 30, of nominal type. (Alternatively, create the column in Google Sheets with =IF(K2>=30, 1, 0) and re-import the file.) Then, from the Frequencies menu, Bayesian section, choose Contingency Tables. Put obez under Rows and diabet_ogtt under Columns. Under Sampling, choose Joint multinomial, with the default prior concentration (1). Expected result: BF10 about 85, that is, very strong evidence for the association between obesity and diabetes. If you choose Independent multinomial, rows fixed, BF10 drops to about 64; the conclusion remains the same. Note that the sampling type is a design choice that must be justified and reported.

Step 4 (optional). From the Frequencies menu, Bayesian section, choose Binomial Test for the variable diabet_ogtt, with the default prior Beta(1, 1). Check the Prior and posterior plot. The 95% central credible interval for the proportion with diabetes should be about 12.9-24.8%, as in Figure 11.4 of the chapter. Change the prior to Beta(10, 40) and notice how little the interval changes (about 13.6-24.1%).

Step 5. Report the results in the format of the discipline's assignments: (a) the research hypothesis; (b) the statistical hypotheses; (c) the method and its justification (why a Bayesian test); (d) the software and the settings (the prior, the sampling type); (e) the result, with the chart; (f) the statistical conclusion, using the scale in Table 11.8 of the chapter; (g) the research conclusion in words.

Additional Exercises

1. In the BPOC file (COPD data) on Moodle, consider dyspnea (DISPN) as the "disease" and rales (RAL) as the "symptom". Build the disease-symptom table with COUNTIFS and calculate p(B+), p(S+ / B+), p(S+ / B-) and p(B+ / S+). Check p(B+ / S+) with Bayes' theorem. Comment on whether rales would be a useful "test" for dyspnea in this sample.
2. In the Date_HbA1c file on Moodle, choose a clinical threshold for HbA1c (for example, 6.5%) and a variable that can serve as a reference, if there is one in the file. Build the 2 x 2 table, calculate Se, Sp, LR+ and LR-, and use the calculator from L11.1 for three different pre-test probabilities. Discuss the limitations of the chosen reference.
3. Problem with data (complete solution). A rapid streptococcal antigen test has Se 86% and Sp 95%. In a child with pharyngitis, the physician estimates a pre-test probability of 40%. Calculate the post-test probability after a positive test and after a negative test.

Solution. $LR^+ = \frac{0.86}{0.05} = 17.2$; $LR^- = \frac{0.14}{0.95} = 0.147$. Pre-test odds = 0.40/0.60 = 0.667. Positive test: 0.667 x 17.2 = 11.47, probability 11.47/12.47 = 92.0%. Negative test: 0.667 x 0.147 = 0.098, probability 0.098/1.098 = 8.9%. With a treatment threshold of 25% (Worked Problem 11.6), the positive test justifies treatment, and the negative one brings the probability below the threshold.

4. Problem with data (complete solution). Using the differential diagnosis table from L11.4, a patient has low ferritin (F+), normal RDW (R-) and a Mentzer index above 13 (M-). Calculate the posterior probabilities.

Solution. Likelihoods: iron deficiency 0.90 x 0.20 x 0.85 = 0.153; thalassemia 0.05 x 0.80 x 0.20 = 0.008; chronic disease 0.10 x 0.60 x 0.80 = 0.048. Products with the priors: 0.60 x 0.153 = 0.0918; 0.15 x 0.008 = 0.0012; 0.25 x 0.048 = 0.0120. Sum 0.1050. Posterior probabilities: iron deficiency 87.4%, thalassemia 1.1%, chronic disease 11.4%.

What to submit and indicative grading scheme

Submit on Moodle a single Google Sheets file (a link with view rights) or one exported as .xlsx, with the sheets "Bayes", "Differential" and "Screening", plus a short document (one page at most) with the results from JASP (screenshots of the tables with BF10) and the interpretation in the a-g format.

Indicative grading scheme (10 points): a correct Bayesian calculator, with the cross-check B8 = B11 (2 points); a working natural frequency tree for three prevalences (1.5 points); the nomogram exercises, with the reading and the calculation (1 point); the differential diagnosis table with the correct formulas and the two scenarios (2 points); the 2 x 2 table of FINDRISC obtained with COUNTIFS and a variable threshold, plus the explanation of sequential testing (1.5 points); the JASP analyses correctly reported, with the interpretation of BF10 on the scale from the lectures (2 points).

Offline variant

If you do not have access to Google Sheets or JASP, all the calculations in L11.1-L11.5 can be done with a pocket calculator and on paper, using the tables in this session as templates: convert the probability to odds, multiply by the LR, convert back. The natural frequency tree is drawn by hand for 1,000 people, and the printed Fagan nomogram replaces the calculator. For L11.5, the 2 x 2 table of FINDRISC is given in Chapter 12 (Table 12.3), and for L11.6 use the reference results given above (BF10 0.18 for SBP and sex; about 85 for obesity and diabetes) and write only the interpretation. The same logic can be applied in LibreOffice Calc or Microsoft Excel, with the same formulas (in a configuration with a decimal comma, the argument separator becomes ";").

Practical Session 12. Performance of Diagnostic Tests: Sensitivity, Specificity, Predictive Values and the ROC Curve

Objective

In this session you evaluate a diagnostic test from start to finish: you build the 2 x 2 table, calculate all the indicators (first by hand, then in Google Sheets and MedCalc), choose a threshold, build the ROC curve and compare two tests. The main test is the FINDRISC score from the mg2026 dataset, with the OGTT as the reference standard; the laboratory test analyzed at the end is fasting plasma glucose.

By the end, the student should be able to:

- calculate from a 2 x 2 table the prevalence, Se, Sp, PPV, NPV, accuracy, the error rate (b + c)/N, the Youden index, LR+, LR- and DOR, with confidence intervals;
- use the MedCalc procedure for evaluating a diagnostic test and recalculate the predictive values for a different prevalence;
- build in Google Sheets a table of thresholds and the corresponding ROC curve, and estimate the AUC by the trapezoidal method;
- run the ROC analysis in MedCalc (AUC with 95% CI, the criterion table, the Youden index) and the comparison of two ROC curves;
- obtain the same results in JASP, where the version allows it, and read a 2 x 2 table in Epi Info;
- write up the evaluation of the performance of a laboratory test in the a-g format of the discipline's assignments.

Tools

Google Sheets (English (United States) regional settings: decimal point, "," as the separator), MedCalc 24.0 [50], JASP 0.98.1 [47], Epi Info 7.2.7 [51]. The file mg2026_curat.csv from Moodle. Approximate time: 100 minutes.

The columns of mg2026_curat.csv used here: S (glicemie0, mg/dL), V (hba1c0, %), W (findrisc, 0-26), X (diabet_ogtt: 1 diabetes confirmed by the OGTT, 0 no diabetes); the data occupy rows 2-161. The mg2026 dataset contains 29 people with diabetes and 131 without diabetes.

Procedure

L12.1 The Indicators of a 2 x 2 Table in Google Sheets (15 Minutes)

We start from the classic problem from the lectures: a test applied to 440 patients, with a = 150, b = 30, c = 4, d = 256 (the diseased in the first row, the positive test in the first column).

Step 1. In a new sheet "2x2", enter the table: B1 "T+", C1 "T-", A2 "B+", A3 "B-", then B2 = 150, C2 = 30, B3 = 4, C3 = 256. In D2 type =B2+C2, in D3 =B3+C3, in B4 =B2+B3, in C4 =C2+C3, in D4 =D2+D3.

Step 2. Starting at A6, build the list of indicators (the label in column A, the formula in column B):

Table L12.1. Formulas for the indicators in Google Sheets.

Indicator	Formula in column B	Expected result
Prevalence	=D2/D4	40,9%
Sensitivity	=B2/D2	83,3%
Specificity	=C3/D3	98,5%
PPV	=B2/B4	97,4%
NPV	=C3/C4	89,5%
Accuracy	=(B2+C3)/D4	92,3%
Error rate	=(C2+B3)/D4	7,7%
Youden index	=B2/D2+C3/D3-1	0,818
LR+	=(B2/D2)/(1-C3/D3)	54,2
LR-	=(1-B2/D2)/(C3/D3)	0,17
DOR	=(B2*C3)/(C2*B3)	320

Watch the error rate: the formula contains cells C2 (false negatives, b) and B3 (false positives, c). A formula that uses C3 (true negatives) is wrong; check that accuracy plus error rate equals 1.

Step 3. The Wilson confidence interval for sensitivity. If the numerator is in B2 and the denominator in D2, the lower limit is:

=(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)

and the upper limit is obtained by replacing "-1.96*SQRT" with "+1.96*SQRT". For Se = 150/180 you should obtain about 77.2-88.1%. Adapt the formula for specificity (C3 and D3): about 96.1-99.4%.

Step 4. Predictive values at a different prevalence. In a cell (for example, F2), type a new prevalence, 0.10, and calculate the PPV with Bayes' theorem: =(B2/D2)*F2/((B2/D2)*F2+(1-C3/D3)*(1-F2)), and the NPV with =(C3/D3)*(1-F2)/((C3/D3)*(1-F2)+(1-B2/D2)*F2). At a prevalence of 10% you should obtain a PPV of 85.8% and an NPV of 98.2%. Compare with the values at the prevalence of 40.9%.

Step 5. Redo in your notebook, without software, the two reconstruction problems from Chapter 12 (Worked Problems 12.2 and 12.3), then check them by entering the tables you obtained in the "2x2" sheet. All formulas are recalculated automatically.

L12.2 Evaluating the Test in MedCalc (10 Minutes)

MedCalc has a dedicated procedure for the 2 x 2 table of a diagnostic test, in the Tests menu, the option Diagnostic test (2 x 2 table). If the installed version organizes the menus differently, look for the procedure by name; the same procedure is available free of charge online, on the MedCalc website, under the name "Diagnostic test evaluation calculator".

Step 1. Pay attention to the orientation of the table: in MedCalc, the disease is in the columns (Disease present / Disease absent) and the test is in the rows (Test positive / Test negative). Enter: Test positive and Disease present = 150; Test positive and Disease absent = 4; Test negative and Disease present = 30; Test negative and Disease absent = 256. If you enter 30 instead of 4, all the indicators will be wrong, although the program will not warn you.

Step 2. Read the results. The program displays the sensitivity, specificity, LR+, LR-, prevalence, PPV, NPV and accuracy, each with a 95% CI. For Se and Sp, MedCalc uses exact binomial intervals (Clopper-Pearson), slightly wider than the Wilson ones: Se 83.3% (95% CI

77.1-88.5), Sp 98.5% (95% CI 96.1-99.6), accuracy 92.3% (95% CI 89.4-94.6), LR+ 54.2 (95% CI about 20.4-143.6), LR- 0.17 (95% CI about 0.12-0.23).

Step 3. In the prevalence field (Disease prevalence), enter 10%. The PPV and NPV are recalculated for the new prevalence; you should find the values from L12.1, step 4. Note how wide the confidence interval for the LR+ is: with only 4 false positives, the estimate is very imprecise.

L12.3 The Threshold Table and the ROC Curve for FINDRISC in Google Sheets (20 Minutes)

Import mg2026_curat.csv into a sheet "Data". In a new sheet "ROC", build the threshold table.

Step 1. In A1:I1, type the header: Threshold, TP, FP, FN, TN, Se, Sp, 1-Sp, J. In A2:A8, type the thresholds 16, 15, 14, 12, 11, 10, 8 (ordered from the strictest to the most permissive).

Step 2. In B2 (TP): =COUNTIFS(Data!$X$2:$X$161, 1, Data!$W$2:$W$161, ">="&A2). In C2 (FP): =COUNTIFS(Data!$X$2:$X$161, 0, Data!$W$2:$W$161, ">="&A2). In D2 (FN): =29-B2. In E2 (TN): =131-C2. In F2: =B2/29. In G2: =E2/131. In H2: =1-G2. In I2: =F2+G2-1. Copy row 2 down to row 8.

Step 3. Check the results against Table 12.5 in the chapter. At the threshold ≥ 12 you should obtain TP 22, FP 33, FN 7, TN 98, Se 0.759, Sp 0.748; at ≥ 11, the maximum J = 0.595.

Step 4. Add the point (0, 0) above the thresholds and the point (1, 1) below them, in a separate area with two columns (1-Sp and Se), for example K1:L9, in increasing order of 1-Sp. Select the area and insert a chart of the Scatter chart type; in Chart editor > Customize > Series, check "Line" to join the points. Add the diagonal series, with the points (0, 0) and (1, 1). Axis titles: "1 - specificity" and "sensitivity". The chart should resemble Figure 12.2 in the chapter.

Step 5. Calculate the AUC by the trapezoidal method. If 1-Sp is in K2:K10 and Se in L2:L10 (with (0, 0) in the first row and (1, 1) in the last), in M3 type =(K3-K2)*(L3+L2)/2, copy down to M10, then =SUM(M3:M10). With the seven thresholds above you will obtain an area of about 0.845; with all the possible thresholds of the score (0-25), the area becomes 0.852. Explain in your notebook why the area with few thresholds is smaller.

Step 6 (optional). Repeat for fasting plasma glucose (column S), with the thresholds 126, 120, 115, 110, 105, 100. Control results: at ≥ 110, TP 24, FP 39 (Se 82.8%, Sp 70.2%); the trapezoidal area with these six thresholds is 0.836 (Worked Problem 12.7).

L12.4 ROC Analysis and Comparison of Curves in MedCalc (20 Minutes)

Open mg2026_curat.csv in MedCalc (File > Open, or import the CSV file and check that the decimal separator has been recognized correctly).

Step 1. From the Statistics menu, choose ROC curves > ROC curve analysis. Under Variable, select findrisc; under Classification variable, select diabet_ogtt (coding 1 = disease present). Under Methodology, you can choose the DeLong or the Hanley and McNeil method for the standard error. Check the option to display the criterion values (Criterion values and coordinates of the ROC curve) and the ROC plot.

Step 2. Expected results: AUC 0.852; the test of AUC = 0.5 with $p < 0.001$. The confidence interval depends on the method chosen: 0.762-0.943 with the Hanley-McNeil formula, about 0.782-0.923 with the DeLong standard error; MedCalc may display an "exact binomial" interval by default, which is slightly different. Note in the report which method the program used.

Step 3. The Youden index. MedCalc displays the criterion associated with the maximum Youden index: J = 0.595, with the criterion "> 10". For a score with integer values, "> 10" means

exactly "≥ 11", the threshold in the chapter (Se 93.1%, Sp 66.4%). This notation confuses many students in the exam: always read the sign in front of the criterion.

Step 4. In the criterion table, find the row "> 11" (that is, ≥ 12) and check Se 75.86% and Sp 74.81%, with +LR 3.01 and -LR 0.32. The criterion table is, in fact, the threshold table you built manually in L12.3.

Step 5. Comparing two ROC curves. From the Statistics menu, choose ROC curves > Comparison of ROC curves. Select the variables findrisc and glicemie0, with diabet_ogtt as the classification variable. Expected results: AUC FINDRISC 0.852, AUC blood glucose 0.844; difference 0.008 (95% CI about -0.060 to 0.077), z about 0.24, p about 0.81 (the DeLong method for correlated curves, because both tests were applied to the same patients). Conclusion: no significant difference was found between the discriminative abilities of the two tests. The chart shows two curves that cross.

Step 6 (optional). Add hba1c0 to the comparison. The AUC for HbA1c is 0.764; compare it visually and numerically with the other two tests.

L12.5 The ROC Curve in JASP (10 Minutes)

JASP has no separate menu called "ROC" in its basic modules, but the ROC curve can be obtained through logistic regression and, in recent versions, also through the Machine Learning module. The available options differ between versions; the description below corresponds to the usual layout, and if you cannot find an option, use MedCalc (L12.4) or the calculation in Sheets (L12.3).

Step 1. Open mg2026_curat.csv. Set diabet_ogtt as a nominal variable.

Step 2. From the Regression menu, choose Logistic Regression. Put diabet_ogtt under Dependent Variable and findrisc under Covariates.

Step 3. In the statistics or model performance section (Statistics / Performance Diagnostics), check Confusion matrix and, under Performance metrics, AUC, Sensitivity, Specificity, Precision and F-measure. In the performance plots section, check ROC plot.

Step 4. Interpretation. The AUC reported for the model with a single predictor is 0.852, the same as the AUC of the FINDRISC score, because the logistic model transforms the score monotonically into a probability, without changing the order of the patients. However, the confusion matrix and the sensitivity displayed by JASP use, by default, a threshold of 0.5 for the predicted probability, not a threshold of the score; with a prevalence of 18.1%, very few patients exceed a predicted probability of 0.5, so the sensitivity displayed will be low and the specificity high. This is a good example of "the same ROC curve, a different operating point". If the version allows the threshold (cutoff) to be changed, try 0.18 and compare with the threshold table.

Step 5. The alternative in the Machine Learning module (if it is activated, from the "+" button of the module bar): a classification procedure also reports the confusion matrix, the AUC and the ROC curve, usually on a separate test set. The results will differ from those above because the evaluation is done on part of the data; it is, in fact, a simple internal validation.

L12.6 The 2 x 2 Table in Epi Info (5 Minutes)

Epi Info has no procedure dedicated to diagnostic tests in the form we use, but the 2 x 2 table is easy to obtain, and the row and column percentages contain the indicators we are looking for.

Step 1. In Classic Analysis or in the Visual Dashboard, import the file and create the variable findrisc12 (1 if findrisc ≥ 12, otherwise 0), for example with the DEFINE and RECODE commands or with a computed variable.

Step 2. Run a table (Tables) with findrisc12 as the exposure (in the rows) and diabet_ogtt as the outcome (in the columns). Epi Info displays the frequencies, the row and column percentages, the chi-square test and the measures of association (OR, RR).

Step 3. Read: in the row findrisc12 = 1, the row percentage of those with diabetes is PPV $\left(\frac{22}{55} = 40.0\%\right)$; in the row findrisc12 = 0, the row percentage of those without diabetes is NPV $\left(\frac{98}{105} = 93.3\%\right)$. In the column diabet_ogtt = 1, the column percentage of those with findrisc12 = 1 is the sensitivity (22/29 = 75.9%); in the column diabet_ogtt = 0, the percentage of those with findrisc12 = 0 is the specificity (98/131 = 74.8%). The OR displayed (9.33) is the diagnostic odds ratio.

Step 4. Note that the significance of the chi-square test shows only that the test and the disease are associated, not how good the test is. A test with an AUC of 0.6 can have $p < 0.001$ in a large sample.

L12.7 Analyzing the Performance of a Laboratory Test: Fasting Plasma Glucose for Diabetes Confirmed by the OGTT (15 Minutes)

Use the results obtained so far (or run the same procedures for glicemie0) and write up a complete analysis in the format of the discipline's assignments.

(a) Research hypothesis: fasting plasma glucose discriminates between adults with and without diabetes confirmed by the OGTT in the population at cardiometabolic risk seen in family medicine.

(b) Statistical hypotheses: H0: AUC = 0.5 (the test does not discriminate); H1: AUC ≠ 0.5.

(c) Method: ROC analysis, with the OGTT as the reference standard; 95% CI for the AUC; choice of threshold by the Youden index and comparison with the clinical threshold of 126 mg/dL.

(d) Software: MedCalc (ROC curve analysis), checked in Google Sheets.

(e) Expected results: AUC 0.844 (95% CI 0.752-0.937, Hanley-McNeil), $p < 0.001$. The Youden threshold ≥ 110 mg/dL (MedCalc displays "> 109" or the next lower value present in the data): Se 82.8%, Sp 70.2%, PPV 38.1%, NPV 94.8%, J = 0.530. At the clinical threshold ≥ 126 mg/dL: Se 37.9%, Sp 94.7%, PPV 61.1%, NPV 87.3%, LR+ about 7.1, LR- about 0.66. Include the ROC plot.

(f) Statistical conclusion: H0 is rejected; fasting plasma glucose has a very good discriminative ability on the scale from the lectures (0.81-0.90), but the lower limit of the confidence interval (0.752) corresponds to a "good" performance.

(g) Research conclusion: in this population, fasting plasma glucose is a useful triage test at a threshold of about 110 mg/dL (it misses few diabetic patients, but sends about 4 in 10 patients without diabetes for an OGTT). The 126 mg/dL threshold confirms diabetes with high specificity, but it misses almost two thirds of the cases identified by the OGTT, so a result below 126 mg/dL does not rule out diabetes. The performance of blood glucose does not differ significantly from that of the FINDRISC score (p = 0.81).

Add a paragraph on limitations: a small sample of diseased patients (29), hence wide intervals for sensitivity; a threshold chosen on the same data (optimistic overestimation); simulated data; and STARD 2015 also requires the participant flow diagram [140].

Additional Exercises

1. In the Date_HbA1c file on Moodle, choose a numerical laboratory variable and a dichotomous variable that can serve as a reference (if there is one). Build in Google

Sheets the threshold table, the ROC curve and the AUC by the trapezoidal method; check in MedCalc. Report in the a-g format.

2. In the BPOC file (COPD data) on Moodle, evaluate how well a spirometric index at the start of the study (for example, IPB1) "predicts" the presence of rales (RAL). Run the ROC analysis in MedCalc, report the AUC with its 95% CI and comment on whether such an analysis makes clinical sense.
3. Problem with data (complete solution). A test has Sp = 0.90, PPV = 0.60, TN = 270 and FN = 10. Reconstruct the table and calculate Se, NPV and the error rate.

Solution. d = 270 and $\text{Sp} = \frac{d}{c+d} = 0.90$, so c + d = 300 and c = 30. From $\text{PPV} = \frac{a}{a+30} = 0.60$: a = 0.60a + 18, so 0.40a = 18 and a = 45. b = 10. Table: a = 45, b = 10, c = 30, d = 270; N = 355. $\text{Se} = \frac{45}{55} = 81.8\%$; $\text{NPV} = \frac{270}{280} = 96.4\%$; $\text{ER} = \frac{10+30}{355} = 11.3\%$; $\text{Acc} = \frac{315}{355} = 88.7\%$.

4. Problem with data (complete solution). For IL-6 in neonatal sepsis, Se 80.1% and Sp 85.7% have been reported [126]. Using the calculator from L12.1 (step 4), calculate the PPV and NPV at prevalences of 5% and 30%.

Solution. $\text{LR}^+ = \frac{0.801}{0.143} = 5.60$; $\text{LR}^- = \frac{0.199}{0.857} = 0.232$. At 5%: odds 0.0526; positive 0.0526 x 5.60 = 0.295, PPV 22.8%; negative 0.0526 x 0.232 = 0.0122, post-test probability 1.2%, NPV 98.8%. At 30%: odds 0.429; positive 0.429 x 5.60 = 2.40, PPV 70.6%; negative 0.429 x 0.232 = 0.0995, post-test probability 9.1%, NPV 90.9%.

What to submit and indicative grading scheme

Submit on Moodle the Google Sheets file (the sheets "2x2", "ROC" and, if applicable, "Glucose"), the screenshots from MedCalc (the ROC analysis for FINDRISC and the comparison of curves) and the one-page document with the analysis of fasting plasma glucose in the a-g format.

Indicative grading scheme (10 points): the "2x2" sheet with all indicators correct, including the error rate (b + c)/N and the Wilson CI (2 points); the evaluation in MedCalc with the correct orientation of the table and the recalculation for a prevalence of 10% (1 point); the threshold table, the ROC curve and the trapezoidal AUC in Sheets (2 points); the ROC analysis and the comparison of curves in MedCalc, with the interpretation of the criterion "> 10" (2 points); JASP or Epi Info, with an explanation of the difference between the default operating point and the thresholds of the score (1 point); the analysis of blood glucose in the a-g format, with limitations (2 points).

Offline variant

Without access to software, the session can be done entirely on paper. The indicators in L12.1 are calculated with a pocket calculator; the threshold tables for FINDRISC and for blood glucose are given in Chapter 12 (Table 12.5 and Worked Problem 12.7); the ROC curve is drawn on graph paper, and the AUC is estimated by the trapezoidal method. For the ROC analysis and the comparison of curves, use the reference results from L12.4 (AUC 0.852 and 0.844, nonsignificant difference, p = 0.81) and concentrate on the interpretation and on the write-up in the a-g format. MedCalc also offers a free online version of the calculator for the 2 x 2 table, usable from a phone when the network comes back.

Practical Session 13. Designing a Mini Clinical Study and the Integrated Analysis of a Complete Dataset

Objective

This practical session connects the two ends of a study: the one where the data are born and the one where they are turned into conclusions. In the first part you will design a mini observational study, build the data collection form in Google Forms and simulate the collection and compilation of the data in Google Sheets. In the second part you will fully analyze the mg2026 dataset, from description to the ROC curve and risk analysis, and write a final report in academic format. The session prepares you directly for the practical exam in Practical Session 14 and applies the concepts from Chapter 13.

By the end, the student should be able to:

- turn a clinical question into a project sheet (PICO, design, variables, sample size);
- build a Google Forms form with a consent page, sections, branching and validation rules, linked to a Google Sheets sheet;
- check, clean and document a raw dataset, with a change log;
- choose and apply the appropriate tests to the mg2026 dataset in JASP and MedCalc and check their results against the control figures in the worksheet;
- write a 2-3 page IMRaD report, with a table, a chart, correctly reported results and limitations.

Tools

- Google Forms and Google Sheets (institutional account), with English (United States) regional settings: the decimal separator is the point, and the argument separator in formulas is "," [43,45,46,148].
- JASP 0.98.1 [47], MedCalc 24.0 [50], optionally Epi Info 7.2.7 (StatCalc) [51]; the JASP working guides and the discipline's collection of applications [14,48,49].
- The files mg2026_curat.csv and mg2026_brut.csv, from the discipline's Moodle platform [52].
- Chapter 13 of this book, for the data dictionary and the analysis plan.
- Approximate time: 100 minutes (about 40 minutes for the design and form part, 50 for the analysis, and the rest for the report, which is finished at home).

Procedure

L13.1 The Project Sheet of the Mini-Study

Work in teams of 3-4 students. Each team designs a mini cross-sectional observational study in a family practice. The proposed topic, which you may adapt: "Physical activity, body mass index and blood pressure in adults attending the family practice". In a shared Google Docs document, complete a project sheet with the following headings.

1. The research question in PICO format [24]. Example: P, adults aged 18-80 years attending the practice during one month; I (the exposure), low physical activity (under 150 minutes per week); C, physical activity of at least 150 minutes per week; O, hypertension (SBP ≥ 140 mmHg or DBP ≥ 90 mmHg) and the SBP value.
2. The design: cross-sectional, a single time point of measurement. State explicitly what this design does not allow (temporal sequence, and therefore a causal conclusion).

3. The statistical hypotheses for the main outcome, in words and mathematically. Example: H0, the mean SBP is the same in those with low activity and in those who are active (μ1 = μ2); H1, the means differ (μ1 ≠ μ2).
4. The variables, with the type of each (numerical, ordinal, nominal) and the test you will use. At least 10 variables: code, age, sex, area of residence, education, smoking (with the number of cigarettes per day for smokers), physical activity in minutes per week, height, weight, SBP, DBP, and a Likert item ("How satisfied are you with your state of health?", 1-5).
5. Sample size. If you want to estimate the prevalence of hypertension, assumed to be 30%, with a margin of ±10 percentage points, $n = \frac{1.96^2 \times 0.30 \times 0.70}{0.10^2} = 80.7$, so 81 patients. Also calculate it for a margin of ±5 percentage points and comment on the difference (the correct result is 323).
6. Ethical aspects: information and consent, anonymization (a study code instead of the name), data minimization (age in years, not the date of birth), and the approval of the ethics committee for a real study [32,35].

Expected result: a one-page project sheet, with the data dictionary in table form (the name of the variable, the description, the type, the unit, the permitted values), following the model of Table 13.2.

L13.2 Building the Form in Google Forms

Open Google Forms from the institutional account and create a new form, with the title of the study. Build it in sections, in the order below. The button for a new question is the "+" sign in the side bar; the button for a new section is the symbol with two rectangles in the same bar.

Section 1, Information and consent. Add a text block (title and description) that explains the aim of the study, what data are collected, that participation is voluntary, that the data are anonymous and that the person can withdraw at any time. Then add a "Multiple choice" question, marked "Required": "Do you agree to take part in this study?", with the options "Yes, I agree" and "I do not agree". From the question's three-dot menu, choose "Go to section based on answer". For "Yes", choose "Continue to next section"; for "No", choose "Submit form" [148]. In this way, anyone who refuses does not see a single question about their health.

Section 2, General data. The participant code ("Short answer", required, with the validation "Regular expression", "Matches", the pattern P[0-9][0-9][0-9], and the error message "The code has the form P001"); age ("Short answer", the validation "Number", "Between", 18 and 80, the message "Enter your age in completed years, between 18 and 80"); sex ("Multiple choice": F, M); area of residence ("Multiple choice": urban, rural); education ("Multiple choice", in order: lower secondary school, high school or vocational school, university, postgraduate). The validation rules are activated from the question's three-dot menu, the option "Response validation" [45].

Section 3, Measurements. Height in cm (validation "Number", "Between" 120 and 220, message "Height in centimeters, for example 168"); weight in kg (between 30 and 250); SBP (between 70 and 250 mmHg) and DBP (between 40 and 150 mmHg). Do not add a question for BMI: it is calculated in the spreadsheet from weight and height. Notice how each rule would have stopped one of the errors in the file mg2026_brut.csv (height 1.68, weight 8.6, SBP 1450).

Section 4, Lifestyle. The question "Do you currently smoke?" ("Multiple choice": "Yes", "No, I have quit", "No, I have never smoked"), with branching: "Yes" leads to section 4a, the other options to section 5. Section 4a contains a single question: "How many cigarettes do you smoke, on average, per day?" (a number between 1 and 80). Next come the minutes of physical activity per week ("Short answer", a number between 0 and 3,000) and the Likert item ("Linear scale", from 1 to 5, with the labels "not at all satisfied" and "very satisfied").

Section 5, End. A thank-you text. In the settings (the gear icon), turn off the collection of e-mail addresses: for an anonymous study, the e-mail address would be an unnecessary piece of identifying data.

Check the form with the "Preview" button (the eye icon): try to enter an age of 15 years, a height of 1.75 and the answer "I do not agree" to the consent question. The form must reject the first two and end immediately at the third. Expected result: a working form, with 5 sections (plus 4a), at least 3 validation rules and 2 branchings.

L13.3 Simulating Data Collection and Compiling the Data in Google Sheets

In the "Responses" tab of the form, choose "Link to Sheets" and create a new spreadsheet [46]. Each team member fills in the form 5 times, with fictitious but plausible data, as if they were different patients; at least one of them refuses to participate, and at least two are smokers. The team thus obtains 15-20 records.

Open the response sheet and carry out the following checks, in a new tab named "checks".

1. Rename the columns with short names, without spaces and without diacritics (cod, varsta, sex, mediu, educatie, inaltime, greutate, tas, tad, fumat, tigari, activitate, satisfactie). The first column, "Timestamp", is kept for traceability.
2. Count the complete records: =COUNTA(B2:B21). Count the refusals: =COUNTIF(B2:B21, "I do not agree").
3. Check for duplicate codes: =COUNTIF(B$2:B$21, B2) in a helper column; any value greater than 1 signals a duplicate.
4. Calculate BMI in a new column: =ROUND(G2/(F2/100)^2, 1), assuming weight is in column G and height in F.
5. Calculate the descriptive statistics for SBP: =AVERAGE(H2:H21), =STDEV(H2:H21), =MEDIAN(H2:H21), =MIN(H2:H21), =MAX(H2:H21).
6. Build a pivot table ("Insert" > "Pivot table") with sex in the rows and smoking in the columns, with the values counted [44].
7. Create a tab "changes", with the columns: date, who, row, variable, old value, new value, reason, source. Deliberately introduce an error (for example, manually change an SBP to 1500), then correct it and record the correction in the log.

Discuss in the team: which validation rules worked, which errors still got through (for example, an SBP of 95 with a DBP of 110, both within the permitted range but clinically incompatible) and which logical check should be added. Expected result: the spreadsheet with the tabs "responses", "checks" and "changes", shared with the instructor as read-only.

L13.4 Cleaning the File mg2026_brut.csv

Import the file mg2026_brut.csv into a new Google Sheets spreadsheet ("File" > "Import" > "Upload"). Without looking at Chapter 13, look for the errors using only the tools below, then compare your list with Table 13.3.

- The number of rows with data: you should obtain 161 (one row is duplicated).
- Duplicates: sort by the id column or use =COUNTIF(A$2:A$162, A2); patient P091 appears twice.
- Categorical variables: for sex, mediu and fumat, use "Data" > "Create a filter" and look at the list of distinct values. You should find: for sex, "f" and "Feminin" in addition to F and M; for mediu, "Urban"; for fumat, "Fumator " (with a space).
- Numerical variables: =MIN() and =MAX() on each column. You should find tas0 = 1540, a height of 1.72 and a weight of 7.26. Use =COUNT() compared with =COUNTA(): the difference shows the values stored as text (blood glucose "87 mg/dl", HbA1c "5,2").

- Missing values: =COUNTBLANK() on each column; age is missing for P042 and blood glucose at 6 months for P134.

Record in the change log how you would handle each error. Recall the rule: formatting errors are corrected directly, impossible values are corrected only from the source document. In this exercise, the "source document" is the file mg2026_curat.csv: check the values of patients P034, P057 and P102 there and note, for each one, what correction you would have guessed and what the source shows. For height and weight, also use the BMI kept in the table as an independent check. Expected result: 12 problems identified, each with its detection method and correction decision.

L13.5 The Complete Analysis of the mg2026 Dataset

Work on mg2026_curat.csv. For each analysis, write in the report elements (a)-(g) of the format of the discipline's assignments: the research hypothesis, the statistical hypotheses, the method and its justification, the software and the way of solving, the result with a chart, the statistical conclusion and the research conclusion. The control figures below allow you to check yourselves.

1. Description. In JASP: "Descriptives" > "Descriptive Statistics", the variables varsta, imc0, tas0, glicemie0, trigliceride; under "Statistics", check the mean, standard deviation, median, quartiles, skewness, kurtosis and "Shapiro-Wilk test"; under "Plots", "Distribution plots" and "Boxplots". Control figures: age 56.29 ± 8.77 years (Shapiro-Wilk p = 0.687); BMI median 28.25, IQR 26.2-31.4, skewness 1.42, p < 0.001; SBP 131.13 ± 13.88 mmHg$(p = 0.303)$; triglycerides median 151.5 mg/dL, skewness 0.95. For the categorical variables ("Frequencies" > "Frequency tables"): F 86, M 74; A 53, B 53, C 54; current smokers 53, former smokers 33, never smokers 74.
2. Table 1 by arm. Repeat the description with "Split" by brat. Compare the results with Table 13.4 and explain why you do not add p values.
3. The confidence interval. In JASP, "T-Tests" > "One Sample T-Test" on tas0, with "Location parameter" and "Confidence interval" checked: 95% CI 128.96-133.30 mmHg. In Google Sheets: =AVERAGE(M2:M161) - T.INV.2T(0.05, 159)*STDEV(M2:M161)/SQRT(160), assuming tas0 is in column M.
4. The primary outcome. Create the variable dglic = glicemie0 - glicemie6 (in JASP, the "+" button in the header of the data table, "Computed column"). "ANOVA" > "ANOVA", dependent dglic, factor brat; under "Assumption Checks", check "Homogeneity tests"; under "Post Hoc Tests", move brat and check "Tukey"; under "Additional Options", "Estimates of effect size" (eta²). Control figures: means 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). In the "Nonparametrics" section of the same menu, Kruskal-Wallis: H = 36.63, p < 0.001.
5. Paired data and repeated measures. "T-Tests" > "Paired Samples T-Test", glicemie0 and glicemie6, with "Normality" checked: mean difference 10.19 mg/dL (95% CI 8.38-11.99), t = 11.14, df 159, p < 0.001, Shapiro-Wilk on the differences p = 0.725. "ANOVA" > "Repeated Measures ANOVA" with glicemie0, glicemie3, glicemie6 as the levels of the time factor: F = 64.97, p < 0.001; under "Nonparametrics", Friedman $\chi^2 = 74.67$, p < 0.001.
6. Proportions and risk. "Frequencies" > "Contingency Tables", rows brat, columns scadere5; under "Statistics", check χ^2, and under "Cells", the expected frequencies and the row percentages: A 6/53, B 16/53, C 16/54; $\chi^2 = 6.77$, df 2, p = 0.034; the smallest expected frequency 12.6. For C vs. A, in MedCalc: "Tests" > "Relative risk", with 16 events of 54 and 6 of 53: RR 2.62 (95% CI 1.11-6.17), NNT 5.5 (that is, 6). The same analysis in Epi Info, "StatCalc" > "2 x 2 tables".
7. Ordinal data. "ANOVA" > "ANOVA" > "Nonparametrics", satisfactie by brat: Kruskal-Wallis H = 32.16, p < 0.001. "T-Tests" > "Independent Samples T-Test" with "Mann-Whitney" checked, for arms A and C only (filter the data): U = 2,265.5, p < 0.001, rank-biserial correlation 0.58.
8. Correlation and regression. "Regression" > "Correlation", imc0 and tas0, with Pearson and Spearman: r = 0.405, rho = 0.372, both p < 0.001; the scatter plot from "Plots". "Regression" >

"Linear Regression", dependent tas0, covariate imc0: SBP = 89.47 + 1.435 x BMI, $R^2 = 0.164$. Add varsta, sex and fumat: $R^2 = 0.308$, BMI coefficient 1.389.

9. The obesity-diabetes association. Create the variable obez (imc0 ≥ 30). "Contingency Tables" obez x diabet_ogtt, with "Log odds ratio" checked: 18/54 vs. 11/106; χ^2 = 12.70, p < 0.001; OR 4.32 (95% CI 1.86-10.03); the prevalence ratio, calculated in MedCalc, 3.21 (95% CI 1.64-6.31). "Regression" > "Logistic Regression", dependent diabet_ogtt, covariates obez, varsta, sex, with "Odds ratios" checked: OR for obesity 6.01, for age 1.11 per year.
10. The diagnostic test. In MedCalc: "Statistics" > "ROC curves" > "ROC curve analysis", the variable findrisc, the classification variable diabet_ogtt; check "Criterion values and coordinates of the ROC curve". Control figures: AUC 0.852 (95% CI 0.762-0.943), Youden threshold > 10, that is, ≥ 11 (Se 93.1%, Sp 66.4%). At the threshold ≥ 12, build the 2 x 2 table by hand (TP 22, FP 33, FN 7, TN 98) and calculate Se 75.9%, Sp 74.8%, PPV 40.0%, NPV 93.3%, LR+ 3.01, LR- 0.32.

Watch the last analysis: MedCalc reports the threshold in the form "> 10", which, for a score with integer values, means "≥ 11". Many students write the wrong threshold in the report precisely for this reason.

L13.6 The Final Report in IMRaD Format

The individual report, 2-3 pages long, follows the structure of a scientific article: Introduction, Methods, Results and Discussion (IMRaD). It is based on the analysis in L13.5, not on the team's fictitious data.

The Introduction (one paragraph) presents the problem (cardiometabolic risk in family medicine), what is known and the study question, in PICO format. The Methods describe the design (a simulated pragmatic randomized trial, 3 arms, 12 practices), the participants and the criteria, the interventions, the primary and secondary outcomes, the statistical analysis (the tests, the level $\alpha = 0.05$, the software with its version) and state that the data are simulated. The Results begin with the flow of participants (Figure 13.3) and with Table 1 (a table with 8-10 rows is enough), continue with the primary outcome (description, test, effect size, CI) and with two secondary outcomes of your choice, then with an observational analysis (obesity-diabetes or FINDRISC). Include a single chart, chosen for the primary outcome (a box plot by arm or a plot of the differences with CIs). The Discussion summarizes the main result in one sentence, interprets it clinically (for example, the NNT), discusses the limitations (lack of blinding, a surrogate outcome, short follow-up, simulated data without losses to follow-up) and formulates the conclusion.

Follow the reporting rules in section 14.4: exact p values, the effect size with a 95% CI, the units of measurement, "H0 is not rejected" instead of "H0 is accepted", and no causal conclusions from the observational analyses. A model sentence for the main result: "At 6 months, the mean decrease in fasting plasma glucose was 3.0 mg/dL in arm A, 11.1 mg/dL in arm B and 16.3 mg/dL in arm C (ANOVA, $F(2;\ 157) = 22.93,\ p < 0.001;\ \eta^2 = 0.23$)."

Additional Exercises

1. Using the BD_Studenti file on Moodle, design the Google Forms form that would have produced this file: for each variable, choose the type of question and the validation rule. Identify the variables that were probably collected in class intervals and propose the question that would have allowed the exact value to be collected.
2. Using the BPOC file (a crossover study with measurements at the start of the study and after medication), write the analysis plan for the question "does heart rate change after treatment?" and for "does vital capacity after treatment differ according to the medication?". For each one, state why the test chosen is one for paired data or for independent groups. Apply the tests and interpret them.

3. Using the Date_HbA1c file, write Table 1 for two groups of your choice, choosing correctly between mean ± SD and median (IQR) for each variable, based on the normality test.
4. A team wants to estimate the prevalence of smoking among the patients of a practice, assumed to be 25%, with a margin of ±5 percentage points and 95% confidence. Calculate n. Solution: $n = \frac{1.96^2 \times 0.25 \times 0.75}{0.05^2} = \frac{3.8416 \times 0.1875}{0.0025} = 288.1$, so 289 patients. With an estimated response rate of 80%, 289/0.8 = 361.3, that is, 362 patients must be invited.
5. In a form, the question "How often do you do sports? rarely / sometimes / often" produced answers that were hard to interpret. Rewrite it so that it produces a numerical variable and explain what you gain in the analysis. Model solution: "In the last 7 days, how many minutes in total did you spend on moderate or vigorous physical activity? ___ minutes"; the numerical variable allows the mean, the median, the correlation with BMI and later dichotomization at a threshold (150 minutes), whereas the options "rarely / sometimes / often" are interpreted differently by each respondent.

What to submit and indicative grading scheme

Submit on Moodle, in the section for Practical Session 13: the project sheet (with the data dictionary), the link to the form (with view rights), the spreadsheet with the tabs "responses", "checks" and "changes", the list of errors in mg2026_brut.csv and the individual IMRaD report, in PDF format.

Table L13.1. Indicative grading scheme for Practical Session 13 (100 points).

Component	Criteria	Points
Project sheet	correct PICO; the design named, with its limitations; statistical hypotheses in words and mathematically; correctly calculated sample size; ethical aspects	15
Data dictionary	all variables, with type, unit, permitted values and codes	5
The Google Forms form	a consent page with exit branching; at least 3 validation rules with clear messages; branching for smokers; appropriate question types; no e-mail collected	15
Compilation in Sheets	renamed columns; checks (duplicates, BMI, descriptive statistics); pivot table; change log	10
Cleaning mg2026_brut	errors identified (at least 10 of 12), with the detection method and the correct decision	10
Analysis of mg2026	tests correctly chosen and justified; assumptions checked; results consistent with the control figures; the (a)-(g) format	25
IMRaD report	structure; Table 1; one appropriate chart; results with effect size and CI; limitations; a conclusion without causal overstatement	20
Total		100

Errors that reduce the score regardless of the component: the wording "H0 is accepted"; a test for independent groups applied to paired data, or vice versa; mean ± SD for visibly skewed variables without discussion; an OR presented as a relative risk; causal conclusions from the obesity-diabetes analysis; the FINDRISC threshold reported incorrectly ("> 10" transcribed as "≥ 10").

Offline variant

If you have no internet access, the design part is done on paper: draw the form section by section, with arrows for the branchings and with the validation rule written next to each field, and the "collection" is simulated by each student filling in 5 printed forms. The compilation is done in an Excel or LibreOffice Calc file, with the same formulas (with English regional settings the argument separator remains ","). The cleaning of the file mg2026_brut.csv and the analysis of mg2026 are done in JASP and MedCalc, installed locally on the laboratory computers; Epi Info can replace MedCalc for the 2 x 2 tables. For the ROC curve without MedCalc, build by hand the table with the thresholds 8, 10, 12, 14, 15 and 16 of the FINDRISC score (the table in Chapter 12) and plot the points (1 - Sp, Se) on graph paper; the area can be approximated by the trapezoidal rule.

Practical Session 14. The Practical Exam: Sample Paper, Grading Scheme and Preparation Plan

Objective

The last practical session is dedicated to the practical exam. You will find out how it is organized, work through two sample papers in full, with complete solutions and the expected figures, see how each requirement is scored and which mistakes cost the most points. The session ends with a two-week preparation plan and a checklist of software skills. A student who can solve both sample papers alone, within the time limit, is ready for the exam.

By the end, the student should be able to:

- describe the structure of the practical exam and its weight in the final grade;
- solve, within the time allowed, a paper with 6-8 tasks covering description, estimation, testing, correlation and regression, risk and the diagnostic test;
- write up each solution in the discipline's (a)-(g) format;
- check their results and avoid the common mistakes listed in the grading scheme;
- organize their preparation on the basis of the list of skills.

Tools

- JASP 0.98.1 [47], MedCalc 24.0 [50], Epi Info 7.2.7 [51], Google Sheets with English (United States) regional settings (argument separator ",") [43].
- The file mg2026_curat.csv and the discipline's classic files from Moodle [52]; the collections of practical applications [14,149] and the JASP working guide [49].
- Chapters 2-12 and 14 of this book; the summary table of formulas (Table 14.2).
- Approximate time: 100 minutes, of which 60 for solving a sample paper under exam conditions and 40 for correcting it on the basis of the grading scheme.

Procedure

L14.1 How the Practical Exam Is Conducted

The final grade in Biostatistics is made up of three parts, according to the course syllabus: the 50-question multiple-choice exam, marked in the format of the Romanian residency examination (50%), the activity and assignments during the semester (20%) and the practical exam in the laboratory (30%). The practical exam tests exactly what you practiced in the sessions: opening a data file, choosing and applying the appropriate method in a program, interpreting the result and writing it up correctly.

The exam is taken individually, at the computer, in the practical session room. You receive a paper with 6-8 tasks, on a familiar dataset (of the mg2026 type or one of the files from the sessions) or on a small, new dataset whose data are given in the statement. You have at your disposal the software used during the semester. The answers are written in a document (or on the exam sheet), in the format required for the assignments: (a) the research hypothesis; (b) the statistical hypotheses in words and mathematically; (c) the method chosen and its justification; (d) the way of solving and the software; (e) the result, with a chart where required; (f) the statistical conclusion; (g) the research conclusion in words. Not all tasks need all seven elements: a descriptive statistics task has no statistical hypotheses, and an estimation task has a confidence interval, not a test.

What is scored: the correct choice of method and its justification (usually the largest share of the score), the correctness of the figures, the statistical interpretation and the medical interpretation. A correct figure with a wrong method earns little; a correct method with a small calculation error, well interpreted, earns almost everything. In the exam, the most frequent loss of points does not come from the calculation but from wrongly worded conclusions ("H0 is accepted", "the correlation proves that...").

L14.2 Sample Paper 1 (the mg2026 Dataset)

Open the file mg2026_curat.csv. Consider the sample representative of adults at cardiometabolic risk in family practices in the west of the country. Solve the tasks, with α = 0.05.

1. Describe the variables HDL cholesterol (hdl) and waist circumference (talie). Choose the appropriate indicators, justify the choice by the shape of the distribution and produce an appropriate chart for each.
2. Estimate the mean HDL in the population, with 95% confidence. In what interval do about 95% of the individual HDL values lie?
3. Does HDL differ between women and men?
4. Study the relationship between waist circumference and fasting plasma glucose at baseline (glicemie0). Produce the scatter plot with the regression line and estimate the blood glucose of a patient with a waist of 100 cm.
5. Does systolic blood pressure at baseline (tas0) differ among current smokers, former smokers and never smokers?
6. Is current smoking associated with hypertension (tas0 ≥ 140 mmHg) at baseline? Calculate the appropriate measure of association.
7. In arm C, did SBP change between baseline and 6 months (tas0 vs. tas6)?
8. Evaluate fasting plasma glucose ≥ 110 mg/dL as a test for diabetes confirmed by the OGTT (diabet_ogtt): build the 2 x 2 table and calculate Se, Sp, PPV, NPV, accuracy, the error rate and the likelihood ratios.

L14.3 Solution to Paper 1

Task 1. HDL: n = 160, mean 49.52 mg/dL, SD 11.78, standard error 0.931, median 49, IQR 41-59, minimum 20, maximum 78, skewness 0.04; Shapiro-Wilk W = 0.991, p = 0.371. We do not reject normality, and the mean and the median are practically equal, so we describe the variable by mean ± SD: 49.52 ± 11.78 mg/dL. The appropriate chart is the histogram (with the normal curve superimposed). Waist: mean 88.54 cm, SD 12.11, median 86, IQR 80.75-95, minimum 65, maximum 140, skewness 0.86; Shapiro-Wilk W = 0.961, p < 0.001 (p = 0.0002). The distribution is skewed to the right and the mean is larger than the median, so we describe it by the median and IQR: 86 (80.75-95) cm. Chart: a box plot or a histogram. In JASP: "Descriptives" > "Descriptive Statistics", with Shapiro-Wilk checked.

Task 2. (c) Estimation by a confidence interval for the mean, with the t distribution (n = 160, df 159, t = 1.975). $(d)\ \text{SE} = \frac{11.78}{\sqrt{160}} = 0.931;\ \text{CI}\,95\% = 49.52 \pm 1.975 \times 0.931 = 49.52 \pm 1.84.$ (e) 95% CI for the mean: 47.68-51.36 mg/dL. The interval for the individual values is $m \pm 1.96 \times s = 49.52 \pm 23.09$, that is, 26.4-72.6 mg/dL. (g) With 95% confidence, the population mean HDL lies between 47.7 and 51.4 mg/dL; about 95% of patients have HDL between 26 and 73 mg/dL. The two intervals answer different questions, and confusing them is an error penalized in the grading scheme.

Task 3. (a) HDL differs between the sexes. (b) H0: μF = μM (there is no difference between the mean HDL values in women and men); H1: μF ≠ μM. (c) A numerical variable, two independent groups; normality is not rejected in either group (Shapiro-Wilk p = 0.413 in women,

p = 0.477 in men), but Levene's test rejects the equality of variances (p = 0.034), so we use Welch's t-test. (d) JASP: "T-Tests" > "Independent Samples T-Test", with "Welch" and "Equality of variances" checked. (e) Women 53.92 ± 11.70 mg/dL(n 86), men 44.41 ± 9.66 mg/dL(n 74); t = 5.63, df 157.7, p < 0.001; difference 9.51 mg/dL (95% CI 6.18-12.85); Cohen's d 0.88. (f) We reject H0. (g) Women have HDL about 9.5 mg/dL higher than men on average, a large difference (d 0.88), consistent with known physiology. The Student variant would have given t = 5.55, with the same conclusion; the choice of Welch is justified by the result of Levene's test.

Task 4. (a) A larger waist is associated with higher blood glucose. (b) H0: ρ = 0 (the variables are not correlated); H1: ρ ≠ 0. (c) Both variables are numerical, but neither is normally distributed (talie p = 0.0002; glicemie0 p = 0.037), so the main correlation is Spearman's; we also report Pearson and the regression line, which is explicitly required. (d) JASP: "Regression" > "Correlation" (Pearson and Spearman), then "Regression" > "Linear Regression", dependent glicemie0, covariate talie. (e) rho = 0.430, p < 0.001; r = 0.488, p < 0.001; $R^2 = 0.238$. The line: glicemie0 = 49.17 + 0.636 x talie (95% CI for the slope 0.457-0.814); the standard error of the estimate 13.80 mg/dL. For a waist of 100 cm: 49.17 + 63.55 = 112.7 mg/dL. (f) We reject H0. (g) There is a moderate positive correlation between waist and blood glucose: each additional centimeter is associated with a blood glucose about 0.6 mg/dL higher; waist explains about 24% of the variability in blood glucose. The individual prediction is imprecise (the standard error of the estimate is almost 14 mg/dL), and the association does not prove causation.

Task 5. (a) SBP differs by smoking status. (b) H0: μcurrent = μformer = μnever; H1: at least one mean differs. (c) A numerical variable, three independent groups; Shapiro-Wilk in each group: p = 0.37 (former smokers), p = 0.86 (current smokers), p = 0.23 (never smokers); Levene p = 0.114: one-way ANOVA with Tukey post hoc tests. (d) JASP: "ANOVA" > "ANOVA", dependent tas0, factor fumat, with "Homogeneity tests", "Tukey" and eta². (e) Former smokers 127.5 ± 14.4 mmHg(n 33), current smokers 137.0 ± 14.6(n 53), never smokers 128.5 ± 11.8(n 74); F(2, 157) = 7.85, p < 0.001 (p = 0.0006); $\eta^2 = 0.091$. Tukey: current vs. former smokers +9.5 mmHg (95% CI 2.5-16.5; p = 0.005); current vs. never smokers +8.5 mmHg (95% CI 2.9-14.2; p = 0.001); former vs. never smokers 1.0 mmHg (95% CI -5.6 to 7.6; p = 0.94). Kruskal-Wallis, as a check, H = 12.77, p = 0.002. (f) We reject H0. (g) Current smokers have an SBP 8-10 mmHg higher than former smokers and never smokers, who do not differ from each other. The difference is clinically relevant. Since the data are cross-sectional, we cannot claim that smoking produces this difference; the multiple regression in Chapter 9 shows, however, that it persists at the same BMI, age and sex (+6.7 mmHg vs. former smokers).

Task 6. (a) Smoking is associated with hypertension. (b) H0: the proportion with hypertension is the same in smokers and nonsmokers (or PR = 1); H1: the proportions differ. (c) Two dichotomous variables (smoker yes/no, hypertension yes/no), cross-sectional design: chi-square; the measure of association is the prevalence ratio (PR), with the OR reported in addition. (d) Create the variables fum (fumat = "fumator") and hta (tas0 ≥ 140) in JASP ("Computed column"), then "Frequencies" > "Contingency Tables", with χ², the expected frequencies and "Log odds ratio"; the PR in MedCalc, "Tests" > "Relative risk". (e) Hypertension: 21/53 (39.6%) in smokers, 20/107 (18.7%) in the others; the smallest expected frequency 13.6; χ² = 8.15, df 1, p = 0.004; PR = 2.12 (95% CI 1.26-3.55); OR = 2.85 (95% CI 1.37-5.95). (f) We reject H0. (g) The prevalence of hypertension is about 2 times higher in current smokers; the OR overestimates this relationship, because hypertension is common (25.6% in the sample).

Task 7. (a) SBP changed in arm C. (b) H0: μd = 0 (the mean of the differences tas0 - tas6 is 0); H1: μd ≠ 0. (c) Paired data (the same patients, two time points), n = 54; the normality of the differences is not rejected (Shapiro-Wilk p = 0.196): paired t-test. (d) Filter to arm C in JASP (the

funnel icon, brat = "C"), then "T-Tests" > "Paired Samples T-Test", tas0 and tas6, with "Normality". (e) SBP 129.1 ± 14.0 mmHg at baseline and 122.1 ± 16.8 mmHg at 6 months; mean difference 6.96 mmHg (SD 8.18; 95% CI 4.73-9.20); t = 6.26, df 53, p < 0.001; dz = 0.85; Wilcoxon, as a check, p < 0.001. (f) We reject H0. (g) In arm C, SBP decreased on average by about 7 mmHg over 6 months. Be careful with the interpretation: a before-after comparison in a single arm does not show the effect of the intervention relative to control; for that, the decreases are compared between the arms (in mg2026, ANOVA on the decrease in SBP gives p = 0.056).

Task 8. (c) A diagnostic test against the OGTT reference standard, with a fixed threshold. (d) Create the variable test = glicemie0 ≥ 110 and build the table test x diabet_ogtt. (e) The table:

Table L14.1. Fasting plasma glucose ≥ 110 mg/dL versus diabetes confirmed by the OGTT (mg2026, n = 160).

	OGTT diabetes yes	OGTT diabetes no	Total
Glucose ≥ 110	24	39	63
Glucose < 110	5	92	97
Total	29	131	160

Se $= \frac{24}{29} = 82.8\%$; Sp $= \frac{92}{131} = 70.2\%$; PPV $= \frac{24}{63} = 38.1\%$; NPV $= \frac{92}{97} = 94.8\%$; accuracy = (24 + 92)/160 = 72.5%; error rate = (39 + 5)/160 = 27.5%; $LR^+ = \frac{0.828}{0.298} = 2.78$; $LR^- = \frac{0.172}{0.702} = 0.25$. (g) At this threshold, fasting plasma glucose misses about one in six patients with diabetes and gives a false positive result in almost 30% of those without diabetes. A negative result makes diabetes unlikely (NPV 94.8%), while a positive result raises the probability only to 38%. The overall performance of blood glucose (AUC 0.844) is close to that of the FINDRISC score (AUC 0.852).

L14.4 Sample Paper 2 (a Small, New Dataset)

A family physician tests an 8-week program of dietary sodium reduction in 12 hypertensive patients. The physician measures SBP before and after the program and records BMI at the start of the program.

Table L14.2. The data for paper 2 (SBP in mmHg, BMI in kg/m²).

Patient	SBP before	SBP after	BMI
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
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

Tasks (α = 0.05):

1. Describe SBP before the program (mean, SD, median, minimum, maximum), with the hand calculation of the mean and the standard deviation.

2. Estimate the mean SBP before the program in the population of similar hypertensive patients, with 95% confidence (t(11, 0.975) = 2.201).
3. Did SBP decrease after the program? Apply the appropriate test, with the hand calculation of the test statistic.
4. Is BMI correlated with SBP before the program? Calculate r by hand, test it (t(10, 0.975) = 2.228) and write the regression line.
5. In a cohort of 400 hypertensive patients followed for 5 years, 150 are smokers. Stroke occurred in 18 smokers and 10 nonsmokers. Calculate the RR with 95% CI, the OR, the risk difference, the NNH and the attributable fraction in the exposed; test the association.
6. A rapid test for left ventricular hypertrophy has Se 90% and Sp 80%. It is applied to 1,000 hypertensive patients, with a prevalence of hypertrophy of 15%. Reconstruct the 2 x 2 table and calculate the PPV, NPV, accuracy, error rate, LR+ and LR-. What would the PPV be at a prevalence of 2%?
7. Which test do you choose for: a) comparing SBP after the program between men and women (6 and 6 patients, skewed distributions); b) comparing the proportion of patients with SBP < 140 mmHg before and after the program in the same patients; c) comparing satisfaction (1-5) among three different programs; d) the course of SBP at 0, 4 and 8 weeks in the same patients, with normal distributions?

L14.5 Solution to Paper 2

Task 1. The sum of the SBP values before is 1,833, so $m = \frac{1833}{12} = 152.75$ mmHg. The sum of the squared deviations from the mean is 574.25, so $s^2 = \frac{574.25}{11} = 52.20$ and s = 7.23 mmHg. The ordered values: 142, 145, 147, 148, 149, 150, 152, 155, 158, 160, 162, 165; the median is the mean of the values in positions 6 and 7: (150 + 152)/2 = 151 mmHg; minimum 142, maximum 165. Shapiro-Wilk p = 0.805: we do not reject normality, so mean ± SD is an appropriate description: 152.75 ± 7.23 mmHg.

Task 2. $\text{SE} = \frac{7.23}{\sqrt{12}} = 2.086$; $\text{CI}\,95\% = 152.75 \pm 2.201 \times 2.086 = 152.75 \pm 4.59$, that is, 148.16-157.34 mmHg. With a sample of 12 patients we use the t distribution with 11 degrees of freedom, not the value 1.96.

Task 3. (a) The program lowers SBP. (b) H0: μd = 0; H1: μd ≠ 0 (or, one-sided, μd > 0, if this was decided in advance). (c) Paired data, n = 12; the differences (before - after) are 12, 7, 10, 5, 13, 8, 15, 4, 9, 6, 11, 2; Shapiro-Wilk on the differences p > 0.99: paired t-test. (d) The mean of the differences $\text{md} = \frac{102}{12} = 8.5$ mmHg; the sum of the squared deviations of the differences is 167, so $\text{sd} = \sqrt{\frac{167}{11}} = \sqrt{15.18} = 3.90$; $\text{SE} = \frac{3.90}{\sqrt{12}} = 1.125$; $t = \frac{8.5}{1.125} = 7.56$, df 11. The two-sided critical value is 2.201; 7.56 is much larger, so p < 0.001 (the software gives p = 0.00001). 95% CI for md: $8.5 \pm 2.201 \times 1.125 = 8.5 \pm 2.48$, that is, 6.0-11.0 mmHg; $\text{dz} = \frac{8.5}{3.90} = 2.18$. The Wilcoxon test gives, as a check, p < 0.001 (all 12 differences are positive). (f) We reject H0. (g) SBP decreased on average by 8.5 mmHg (95% CI 6.0-11.0). Without a control group, the decrease cannot be attributed with certainty to the program: regression to the mean and the effect of being followed up are alternative explanations.

Task 4. The mean BMI is 30.02 kg/m². The working sums: $S_{\text{xy}} = \sum(x - m_x)\left(y - m_y\right) = 121.85$; $S_{\text{xx}} = \sum(x - m_x)^2 = 70.06$; Syy = 574.25.

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \frac{121.85}{\sqrt{70.06 \times 574.25}} = \frac{121.85}{200.57} = 0.608 t = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.608 \times 3.162}{\sqrt{0.630}} = \frac{1.922}{0.794} = 2.42$$

Since 2.42 > 2.228, we reject H0: ρ = 0 (p = 0.036). $R^2 = 0.369$. The slope $b = \frac{S_{xy}}{S_{xx}} = \frac{121.85}{70.06} = 1.739$; the intercept a = 152.75 - 1.739 x 30.02 = 100.54. The line: SBP = 100.54 + 1.739 x BMI; for a BMI of 30, the predicted SBP is 152.7 mmHg. The Spearman correlation (rho = 0.629, p = 0.028) leads to the same conclusion. Interpretation: a moderate-to-good positive correlation; BMI explains about 37% of the variability in SBP in this small group. With 12 patients, the confidence interval for the slope is very wide (0.14-3.34 mmHg per kg/m²), so the estimate is imprecise.

Task 5. The table: smokers, 18 with stroke, 132 without; nonsmokers, 10 with stroke, 240 without. Risks: 18/150 = 0.12 and 10/250 = 0.04.

$$\text{RR} = \frac{0.12}{0.04} = 3.0;\ \text{SE}(\ln \text{RR}) = \sqrt{\frac{1}{18} - \frac{1}{150} + \frac{1}{10} - \frac{1}{250}} = \sqrt{0.0556 - 0.0067 + 0.1000 - 0.0040} = \sqrt{0.1449} = 0.381;\ \text{CI}\,95\%\,\text{RR} = \exp(1.099 \pm 1.96 \times 0.381) = \exp(0.352) \text{ to } \exp(1.845) = 1.42 \text{ to } 6.33;\ \text{OR} = \frac{18 \times 240}{132 \times 10} = \frac{4320}{1320} = 3.27(;\ \text{CI}\,95\%\ 1.47 - 7.30);\ \text{RD} = 0.12 - 0.04 = 0.08(;\ \text{CI}\,95\%\ 0.023 - 0.137);\ \text{NNH} = \frac{1}{0.08} = 12.5$$, that is, 13; AF in the exposed = (3 - 1)/3 = 0.667

The expected frequencies (the smallest 10.5) allow the chi-square test: χ^2 = 9.22, df 1, p = 0.002 (with the Yates correction 8.03, p = 0.005). Conclusion: in this cohort, hypertensive smokers have a 3 times higher risk of stroke; for every 13 smokers, one additional stroke occurs over 5 years compared with nonsmokers; if the relationship is causal, two thirds of the strokes in smokers would be attributable to smoking. Since this is a cohort design, the RR is the appropriate measure; the OR (3.27) is close, because stroke is relatively rare.

Task 6. Diseased: 150; healthy: 850. TP = 0.90 x 150 = 135; FN = 15; TN = 0.80 x 850 = 680; FP = 170.

Table L14.3. The reconstructed table for the rapid test (n = 1,000, prevalence 15%).

	Hypertrophy yes	**Hypertrophy no**	**Total**
Test positive	135	170	305
Test negative	15	680	695
Total	150	850	1000

PPV = $\frac{135}{305} = 44.3\%$; NPV = $\frac{680}{695} = 97.8\%$; accuracy = 815/1,000 = 81.5%; error rate = (170 + 15)/1,000 = 18.5%; $\text{LR}^+ = \frac{0.90}{0.20} = 4.5$; $\text{LR}^- = \frac{0.10}{0.80} = 0.125$. At a prevalence of 2% (20 diseased per 1,000): TP 18, FP 196, so PPV = $\frac{18}{214} = 8.4\%$, and the NPV rises to 99.7%. Interpretation: a negative result practically rules out hypertrophy; a positive result must be confirmed by echocardiography, especially at low prevalences.

Task 7. a) The Mann-Whitney test (two small, skewed independent groups). b) McNemar's test (proportions on paired data). c) The Kruskal-Wallis test (an ordinal variable, three independent groups). d) Repeated measures ANOVA (three time points, the same patients, normal distributions); the nonparametric alternative would be the Friedman test.

L14.6 The Grading Scheme and Common Mistakes

Each paper is scored out of 100 points, distributed among the tasks in proportion to their difficulty. Within a testing task, the indicative scoring is as follows.

Table L14.4. Indicative distribution of the score for a statistical testing task (10 points).

Element	What is required	Points
(a)-(b) Hypotheses	the research hypothesis; H0 and H1 in words and mathematically	2
(c) Method	the correct test, justified by the type of variable, the number of groups, independence and the checking of assumptions	3
(d)-(e) Solution and result	the software and the menu; the test statistic, the degrees of freedom, p, the effect size or CI; the chart requested	3
(f) Statistical conclusion	"H0 is rejected" or "H0 is not rejected", with the significance level	1
(g) Research conclusion	the medical interpretation, with the direction and size of the effect, without exaggeration	1

For description tasks, points are awarded for choosing the indicators according to the shape of the distribution, for the figures and for the chart. For estimation tasks, for the formula or the menu, the correct interval and its interpretation. For risk and diagnostic test tasks, for a correctly oriented 2 x 2 table, the measure appropriate to the design and the interpretation.

The mistakes that cost the most points, from the experience of previous exam sessions:

- A test for independent groups applied to paired data (or vice versa). This is the most costly mistake, because it invalidates the whole task.
- Failing to check normality before a parametric test, or checking it on the wrong variable (for the paired t-test, normality is checked on the differences, not at each time point).
- "H0 is accepted" instead of "H0 is not rejected"; "the result proves that there is no difference".
- Confusing the confidence interval for the mean with the interval for the individual values.
- An OR reported in a cohort study when the RR is required, or an RR calculated in a case-control study.
- A wrongly oriented 2 x 2 table (the diseased in the rows and the test in the columns, then the formulas applied as if it were the other way round), which swaps Se with PPV.
- The error rate calculated as (a + d)/N; the correct formula is (b + c)/N.
- Causal conclusions from a correlation or from a cross-sectional study.
- p values written as "p = 0.000" (write p < 0.001) and excess decimals.
- The score threshold from MedCalc transcribed incorrectly ("> 10" means "≥ 11" for integer values).
- An inappropriate chart: a pie chart for a numerical variable, a histogram for a nominal variable.

L14.7 The Two-Week Preparation Plan

The plan assumes about an hour and a half a day, five days a week, plus an exam simulation at the end of each week.

Week 1, consolidation.

- Day 1: types of variables, descriptive statistics and charts (Chapter 2, sessions 1-2). On mg2026, describe all the variables and choose the indicators; redo task 1 of paper 1.

- Day 2: the normal distribution, normality tests, the z score, the logarithmic transformation (Chapter 3). On mg2026: SBP, BMI, triglycerides before and after log transformation.
- Day 3: confidence intervals and sample size (Chapter 4). Calculate by hand and in the software the CIs for SBP, HDL and the proportion with diabetes (Wald and Wilson).
- Day 4: t-tests, ANOVA and their nonparametric equivalents (Chapters 6-8). Redo tasks 3, 5 and 7 of paper 1 and task 3 of paper 2, without looking at the solution.
- Day 5: chi-square, relative risk, odds ratio, NNT (Chapters 7 and 10). The smoking-COPD, monitor-visual acuity and contraceptives-rheumatoid arthritis problems from the sessions; task 6 of paper 1 and task 5 of paper 2.
- End of the week: sample paper 1 in 60 minutes, timed, then correction with the grading scheme.

Week 2, integration.

- Day 6: correlation and regression (Chapter 9). Calculate r by hand for paper 2; on mg2026, BMI-SBP, age-SBP, waist-blood glucose; multiple regression.
- Day 7: Bayesian reasoning and diagnostic tests (Chapters 11-12). Reconstruct 2 x 2 tables from Se, Sp and prevalence or from Se and PPV; the ROC curve of the FINDRISC score in MedCalc.
- Day 8: choosing the test (Chapter 8 and Table 14.1). Take 20 questions from the chapter exercises and, for each one, write only the test and its justification, in 1 minute.
- Day 9: writing up (section 14.4). Rewrite the conclusions of all the tasks solved so far in the (f)-(g) format, checking for the forbidden wordings.
- Day 10: review on the files from the sessions (BD_Studenti, BPOC, Date_HbA1c), one question of each type.
- End of the week: sample paper 2 in 60 minutes, then a version of your own: change the variables in paper 1 (for example, cholesterol instead of HDL, area of residence instead of sex) and solve it.

L14.8 Checklist of Software Skills

Check off each skill only after you have performed it on your own, without the worksheet.

Google Sheets:

- ☐ importing a CSV, English (United States) regional settings, the "," separator;
- ☐ =AVERAGE, =MEDIAN, =STDEV, =MIN, =MAX, =QUARTILE, =COUNTIF, =COUNTBLANK;
- ☐ the CI for the mean with =T.INV.2T and =SQRT;
- ☐ a pivot table for frequencies and contingency tables;
- ☐ a histogram and a box plot from the "Insert" > "Chart" menu.

JASP:

- ☐ opening a CSV and checking the type of each variable (nominal, ordinal, scale);
- ☐ a computed column (differences, BMI, dichotomization at a threshold) and filtering to a subgroup;
- ☐ Descriptive Statistics with Shapiro-Wilk, histograms, box plots, "Split" by group;
- ☐ One Sample, Independent Samples (with Welch and Mann-Whitney) and Paired Samples T-Test (with Wilcoxon);
- ☐ ANOVA with Levene, Tukey and eta²; Kruskal-Wallis; Repeated Measures ANOVA and Friedman;
- ☐ Contingency Tables with χ^2, expected frequencies, row percentages and the log odds ratio;
- ☐ Correlation (Pearson, Spearman) with the scatter plot; simple and multiple Linear Regression; Logistic Regression with ORs.

MedCalc:

- ☐ Tests > Relative risk and Odds ratio from directly entered frequencies;
- ☐ Statistics > ROC curves > ROC curve analysis, with the threshold table and the Youden index;
- ☐ the sample size calculator for a mean and for a proportion.

Epi Info:

- ☐ StatCalc > 2 x 2 tables (RR, OR, chi-square, Fisher);
- ☐ StatCalc > Sample size for a cross-sectional study.

Calculation by hand (with a pocket calculator):

- ☐ the mean, SD, SE, and the 95% CI for a mean and for a proportion;
- ☐ the paired t from the differences; r from the sums of products;
- ☐ RR, OR, RD, NNT with CIs; Se, Sp, PPV, NPV, LR, the error rate; the post-test probability via odds.

Additional Exercises

1. Using the BD_Studenti file, put together an exam paper with 6 tasks following the model of paper 1 (a description, a CI, a test for two independent groups, a test for paired data, an association between categorical variables, a correlation) and exchange it with a colleague; each of you solves the other's paper and marks it using the grading scheme in L14.6.
2. Using the BPOC file, solve in the (a)-(g) format: "Does heart rate change significantly after treatment?" and "Is an age over 50 years a risk factor for an initial SBP above 140 mmHg?". State the design of each question and why the appropriate measure for the second one is the prevalence ratio or the OR, not the RR.
3. Using the Date_HbA1c file, choose one skewed and one symmetric variable and write their complete description, with a chart.
4. Problem with data in the statement. A test has Se 80% and PPV 50% in a sample of 400 people with a prevalence of 10%. Reconstruct the table. Solution: diseased 40, TP 32, FN 8; from the PPV of 50%, TP + FP = 64, so FP 32; TN = 360 - 32 = 328. Sp $= \frac{328}{360} = 91.1\%$; NPV $= \frac{328}{336} = 97.6\%$; accuracy (32 + 328)/400 = 90.0%; error rate (32 + 8)/400 = 10.0%; $LR^+ = \frac{0.80}{0.089} = 9.0$.
5. Problem with data in the statement. In 10 patients, blood glucose before and after a treatment has the differences (mg/dL): 15, 8, 22, -3, 10, 12, 5, 18, 9, 14. Calculate the mean of the differences, the SD, t and the conclusion (t(9, 0.975) = 2.262). Solution: $\text{md} = \frac{110}{10} = 11.0$; the sum of the squared deviations from the mean is 442, $\text{sd} = \sqrt{\frac{442}{9}} = \sqrt{49.11} = 7.01$; $\text{SE} = \frac{7.01}{\sqrt{10}} = 2.22$; $t = \frac{11.0}{2.22} = 4.96 > 2.262$, so we reject H0 ($p < 0.001$); $\text{CI}\,95\%\ 11.0 \pm 2.262 \times 2.22 = 6.0 - 16.0$ mg/dL. The normality of the differences is not rejected (Shapiro-Wilk p = 0.97).

What to submit and indicative grading scheme

At the lab simulation, submit on Moodle the document with the solution to sample paper 1, in the (a)-(g) format, with the charts inserted. The instructor marks it using the grading scheme in L14.6 and tells you, for each task, which element was missing. At the actual exam, the score obtained is converted into the grade for the practical exam, which accounts for 30% of the final grade.

Table L14.5. Indicative distribution of the 100 points among the tasks of paper 1.

Task	Content	Points
1	description, choice of indicators, charts	12
2	CI for the mean and the interval for the individual values	10
3	test for two independent groups	14
4	correlation and regression, prediction	14

Task	Content	Points
5	ANOVA with post hoc tests	14
6	chi-square and the measure of association	12
7	test for paired data	12
8	diagnostic test, 2 x 2 table	12
Total		100

Offline variant

If the software cannot be used, the exam can be taken in the "paper" version, with a type 2 paper: the small dataset is given in the statement, and the calculations are done with a pocket calculator, based on the summary table of formulas (Table 14.2) and the tables of critical values of t, χ^2 and z handed out with the paper. The charts are sketched by hand (histogram, box plot, scatter plot with the regression line). To prepare, solve paper 2 entirely by hand and then check the figures in JASP or MedCalc when you have access to them.

Appendix A. Answers to the Self-Assessment Questions and Proposed Exercises

The answers are given by chapter. For multiple-choice questions, all correct options are listed; for exercises with calculations, the result and the method are indicated, and the complete solution can be reworked following the model of the worked problems in the chapter.

Chapter 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

Explanations: In question 3, the trap is option d (0.621), which is $P(\text{obese} \mid \text{diabetes}) = \frac{18}{29}$, not $P(\text{diabetes} \mid \text{obese}) = \frac{18}{54} = 0.333$; the condition indicates the total to which the cell is related. In question 8, two mutually exclusive events with nonzero probabilities cannot be independent: if B has occurred, the probability of A becomes zero, so P(A|B) = 0 ≠ P(A); and P(A) + P(B) = P(A or B) cannot exceed 1. In question 7, a narrow confidence interval shows a small random error, but it says nothing about systematic error; a very large study can be very precise and still biased.

Proposed exercises: 1. $P(D \mid \text{urban}) = \frac{19}{95} = 0.200$; $P(D \mid \text{rural}) = \frac{10}{65} = 0.154$; $P(\text{urban and } D) = \frac{19}{160} = 0.119$; $P(\text{urban}) \times P(D) = 0.594 \times 0.181 = 0.108$. The values are close but not identical; the difference is small and may be due to chance, so the events are approximately independent in the sample (a formal decision requires a chi-square test). 2. $P(\text{smoker} \mid \text{woman}) = \frac{30}{86} = 0.349$; $P(\text{woman} \mid \text{smoker}) = \frac{30}{53} = 0.566$; $P(\text{woman or smoker}) = \frac{86+53-30}{160} = \frac{109}{160} = 0.681$. The first two differ because they are related to different totals (86 women and 53 smokers, respectively). 3. Binomial 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. $k = \frac{2\,400}{120} = 20$; records 9, 29, 49, 69, 89, ..., the last one 9 + 20 x 119 = 2,389. 5. Sampling fraction 90/1,500 = 6%; 600 x 0.06 = 36 (18-39 years), 36 (40-64 years), 300 x 0.06 = 18 (≥ 65 years). 6. (a) information bias (reporting bias, social desirability: consumption is underreported); (b) confounding (physical activity is related both to supplements and to the risk of fracture); (c) selection bias through losses to follow-up (those who remain no longer represent the randomized group); (d) spectrum-type selection bias (patients with advanced disease versus healthy volunteers; the performance of the test will be overestimated). 7. The statistical unit: the prescription; n = 1,498; the target population: prescriptions (and, indirectly, the treated older adults) in urban areas of western Romania; cross-sectional study; numerical variables: age, the number of drugs per prescription; qualitative variables: sex, the presence of a potentially inappropriate drug (yes/no); if a patient has several prescriptions, the observations are no longer independent, and patients with many prescriptions weigh more in the results. 8. 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. Total probability: $P(\text{COPD}) = 0.30 \times 0.15 + 0.70 \times 0.03 = 0.045 + 0.021 = 0.066 (6.6\%)$; Bayes: $P(\text{smoker} \mid \text{COPD}) = \frac{0.045}{0.066} = 0.682 (68.2\%)$. 10. (a) P: adults at cardiometabolic risk in family practices; I: counseling + app + monthly teleconsultation (arm C); C: standard counseling (arm A); O: the change in systolic blood pressure from baseline to 6 months (tas0 - tas6); T: 6 months. (b) The data are pseudonymized (there is a correspondence

list that allows re-identification); they remain personal data, in the special category of data concerning health, subject to the GDPR and Romanian Law 190/2018.

Chapter 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

Explanations: In question 2, the standard deviation (7.0) is larger than half the mean (11.6) for a variable that cannot be negative, so the distribution is skewed to the right and the median better describes the typical value. In question 5, n = 8 is even, and the median is the mean of the values in positions 4 and 5, (4 + 5)/2 = 4.5. In question 7, option c is false: the standard deviation describes the variability of the patients and does not decrease systematically as n increases; only the standard error, $\frac{s}{\sqrt{n}}$, decreases. In question 10, option c is false: a biologically plausible and verified extreme value is not removed; the 1.5 x IQR rule flags values to be checked.

Proposed exercises: 1. Blood group: nominal; Apgar score: ordinal (a discrete score 0-10); number of pregnancies: discrete numerical, ratio scale; body temperature: continuous numerical, interval scale; NYHA class: ordinal; HbA1c: continuous numerical, ratio scale; Likert agreement: ordinal; patient code: an identifier, not analyzed; length of hospital stay: numerical (discrete when whole days are counted), ratio scale; smoking yes/no: dichotomous nominal. 2. A = 0.35 x 40 = 14; B = 15.0 + 35.0 = 50.0%; $C = \frac{10}{40} = 25.0\%$ (check 50.0 + 25.0 = 75.0); D = 0.125 x 40 = 5; E = 40 - 6 - 14 - 10 - 5 = 5 (12.5%); blood glucose ≥ 126 mg/dL: 12.5 + 12.5 = 25.0% (10 patients). 3. $m = \frac{800}{6} = 133.33$ mmHg; median (130 + 136)/2 = 133 mmHg; MAD $= \frac{56}{6} = 9.33$ mmHg; sum of squared deviations 693.33; $s^2 = \frac{693.33}{5} = 138.67$ mmHg^2; s = 11.78 mmHg; CV = 8.8%; SE $= \frac{11.78}{\sqrt{6}} = 4.81$ mmHg. 4. Mean 85/10 = 8.5 days; median (5 + 6)/2 = 5.5 days; two modes, 4 and 5 (a bimodal distribution in the sample); Q1 = 4, Q3 = 8 (the medians of the halves; software gives 4.25 and 7.75), IQR = 4; the fences are -2 and 14, so 31 days is an extreme value (12 is not); without it the mean becomes 54/9 = 6.0 days; report the median 5.5 (Q1-Q3 4-8). 5. k = 1 + 3.322 x log10(250) = 1 + 3.322 x 2.398 = 8.97, that is, 9 classes; $h = \frac{412-68}{9} = 38.2$ mg/dL, rounded to 40 mg/dL (for example, classes 60-99, 100-139, ..., 400-439). 6. If 3 is the SE, SD = 3 x √64 = 24 mg/dL (CV 21%, plausible); if 3 were the SD, the CV would be 2.7%, a biologically implausible homogeneity, so the interpretation as an SE is the plausible one; the SE for SBP in women = 14.47/√86 = 1.56 mmHg. 7. The weighted mean = $\frac{120\times45.2+60\times61.8+20\times70.4}{200} = \frac{10\,540}{200} = 52.7$ years; the simple mean of the means, 59.13 years, overestimates age because it gives the small ward (20 patients) the same weight as the large ward. 8. $\frac{(a)\,26}{53} = 49.1\%$; (b) BMI, triglycerides and the FINDRISC score are reported as median (Q1-Q3) because they are skewed or are scores; (c) after randomization, baseline differences between the arms are by definition due to chance, so a test would check a null hypothesis that we know to be true; the clinical size of the imbalances is judged, not the p value. 9. (a) a bar chart (a pie chart with 4 categories is acceptable); (b) side-by-side box plots by arm; (c) a scatter plot; (d) a line chart; (e) a histogram; (f) 100% stacked bars, one for each sex, with the categories in the order of the scale. 10. The arithmetic mean = 1,111,000/4 = 277,750 copies/mL; the common logarithms are 3, 4, 5, 6, with a mean of 4.5, so the geometric mean = 10 to the power 4.5 = 31,623 copies/mL; viral load varies multiplicatively (an approximately log-normal distribution), and on the logarithmic scale a decrease of "1 log" means a 10-fold reduction, easy to compare between patients.

Chapter 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

Explanations: In question 4, m - 2s = 11.6 - 14.0 = -2.4 days; a variable that cannot be negative and has m - 2s below zero cannot be normal, and lengths of hospital stay almost always have a long right tail. In question 7, a p of 0.30 means only that the hypothesis of normality is not rejected; with 12 observations the test has low power, so it does not "prove" anything and does not require a parametric test. In question 10, option c is false: by construction, 5% of healthy people have values outside the interval, and 1 - 0.95^20 = 0.64 explains option e.

Proposed exercises:

1. z(125) = (125 - 131.13) / 13.88 = -0.44, Φ = 0.3300, 33rd percentile; z(150) = 1.36, Φ = 0.9131, 91st percentile. In Sheets: =NORM.S.DIST(-0.44, TRUE) and =NORM.S.DIST(1.36, TRUE).
2. z(240) = 0.83, theoretical proportion 1 - 0.7967 = 20.3% (observed 23.8%); $z(190) = -0.53$, theoretical proportion 29.8% (observed 30.0%). A good fit; the right tail is slightly heavier than the model predicts, a difference of the size of sampling fluctuations.
3. $49.52 \pm 1.96 \times 11.78 = 49.52 \pm 23.09$, that is, 26.4-72.6 mg/dL; theoretically 95% (observed 94.4%). HDL < 40: z = -0.81, Φ = 0.2090, that is, 20.9% (observed 19.4%).
4. Binomial n = 6, p = 0.33: mean 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. 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. $4.3 \pm 1.96 \times 0.35 = 4.3 \pm 0.69$, that is, 3.61-4.99 mmol/L; $z = \frac{5.1-4.3}{0.35} = 2.29$, Φ = 0.9890, 98.9th percentile; the value is outside the interval and is repeated (hemolysis of the sample can falsely raise serum potassium) before any decision.
7. m - 2s = 9.1 - 11.6 = -2.5 days: the distribution is skewed to the right; the median and IQR (possibly the minimum and maximum) would have been more appropriate, and the comparison should have used a nonparametric test or the logarithmic scale.
8. 0.39 / 0.192 = 2.03 and 0.74 / 0.381 = 1.94: a borderline departure, confirmed by Shapiro-Wilk p = 0.034; the score is discrete, with many tied values. Report the median (IQR), possibly also the mean ± SD; for comparisons with n = 160 the t-test is robust, but a nonparametric test (Mann-Whitney, Kruskal-Wallis) is a safe choice; ideally, report that both lead to the same conclusion.
9. Geometric mean $e^{4.97} = 144.0$; factor $e^{0.49} = 1.63$; the 95% interval: $e^{4.97-0.96} = e^{4.01} = 55.1$ to $e^{4.97+0.96} = e^{5.93} = 376.3$ (about 55-376, asymmetric around 144).
10. IQR = 23; the fences are 93 - 34.5 = 58.5 and 116 + 34.5 = 150.5, so 154 is an extreme value according to Tukey's rule; $z = \frac{154-105.44}{15.76} = 3.08$, so also according to the z criterion. The value is plausible in a population with undiagnosed diabetes: it is checked against the laboratory report, kept, and a sensitivity analysis is performed.
11.
    (a) points on the line; (b) a "U"-shaped curve: points above the line at both ends, rising steeply on the right (positive skewness); (c) a reclining "S" shape: points below the line on the left and above the line on the right; (d) points on the line, with a single point far above the line in the top right corner.

Chapter 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

Explanations: In question 3, the interval 0.90-2.18 includes the null value 1, so p ≥ 0.05; this does not prove the absence of an association (option c), but shows that the data are compatible

both with a decrease and with a doubling of the odds. In question 4, the rule of three applies: 3 / 60 = 0.05, that is, 5%. In question 9, option d is false: a more permissive α level (0.10) reduces the critical value and therefore the required sample, at the cost of more false positive results. In question 10, option e confuses precision with effect size: a wide interval shows an imprecise estimate.

Proposed exercises:

1. SE = $\frac{36.91}{\sqrt{160}} = 2.918$ mg/dL; margin 1.975 x 2.918 = 5.76; 95% CI 203.63-215.15 mg/dL.
2. SE = $\frac{11.78}{\sqrt{160}} = 0.931$ mg/dL. Individual values: 68.3% 37.74-61.30; 95.4% 25.96-73.08; 99.7% 14.18-84.86 mg/dL. Mean: 68.3% 48.59-50.45; 95.4% 47.66-51.38; 99.7% 46.73-52.31 mg/dL.
3. SE = $\frac{1.5}{\sqrt{25}} = 0.30\%$; t(24, 0.975) = 2.064 (small n, σ estimated from the sample); margin 0.62; 95% CI 6.58-7.82%.
4. $p = \frac{41}{160} = 25.6\%$; SE = 0.0345; Wald 18.9-32.4%; Wilson 19.5-32.9%.
5. Rule of three: 3 / 120 = 2.5% (the exact method gives 3.0%).
6. Difference 8 mmHg; SE = $\sqrt{\frac{15^2}{40} + \frac{14^2}{45}} = \sqrt{5.625 + 4.356} = 3.16$ mmHg; 95% CI about 8 ± 6.19, that is, 1.8-14.2 mmHg (with Welch's t, about 80 df, 1.7-14.3). The interval excludes 0, so the difference is significant at the 0.05 level, but the estimate is imprecise.
7. OR = $\frac{20\times40}{30\times10} = 2.67$; ln OR = 0.981; SE = $\sqrt{\frac{1}{20} + \frac{1}{30} + \frac{1}{10} + \frac{1}{40}} = 0.456$; 0.981 ± 0.895 gives 0.086-1.876; 95% CI 1.09-6.52. It excludes 1 (a significant association), but the lower limit is close to 1.
8. (a) $n = \left(\frac{1.96\times10}{2}\right)^2 = 96.04$, so 97; $(b)\, n = \frac{1.96^2\times0.25}{0.04^2} = 600.25$, so 601.
9. $n = \frac{2\times(1.96+0.84)^2}{0.4^2} = \frac{15.68}{0.16} = 98.0$, so 99 per group (G*Power: 100); with 20% losses, 99 / 0.8 = 123.75, so 124 per group (125 starting from 100).
10. The value 1.1 is almost certainly the standard error; SD = $1.1 \times \sqrt{160} = 13.9$ mmHg. OR 1.75 (95% CI 0.95-3.10): the interval includes 1, so the association is not significant at the 0.05 level.
11. The width decreases by $\sqrt{n}$: to halve it, n must be increased 4-fold, so about 160 patients.

Chapter 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

Explanations: In question 3, the value 132 mmHg lies within the 95% CI (128.96-133.30), so the two-sided test at α = 0.05 does not reject H0; option c is wrong because "accepting H0" is not a permitted conclusion, and the population mean is not known for certain. In question 7, switching from a justified one-sided test to a two-sided one splits α between two tails and reduces power; an increase in the standard deviation increases the noise and reduces power. In question 8, option d overgeneralizes: a nonsignificant result does not demonstrate the absence of an effect, and the confidence interval shows only which differences are excluded in the population studied.

Proposed exercises: 1. The t-test for one mean: SE = $\frac{0.45}{\sqrt{160}} = 0.0356$; $t = \frac{5.58-5.7}{0.0356} = -3.37$ (with the exact mean 5.5775, t = -3.48), df = 159, p < 0.001; H0 is rejected. 95% CI for the mean 5.51-5.65%. The mean HbA1c in the population studied is significantly below 5.7%. 2. SE = $\frac{4.27}{12.649} = 0.338$; $t = \frac{10.37-12}{0.338} = -4.83$, p < 0.001; 95% CI 9.70-11.04; $d = -\frac{1.63}{4.27} = -0.38$ (a small-to-medium effect). H0 is rejected; the mean FINDRISC is below 12. 3. SE = $\frac{9.30}{12.649} =$

0.735; $t = \frac{2.71}{0.735} = 3.68$, p < 0.001 (about 0.0003); 95% CI 73.26-76.16 beats/min; d = 0.29. The result is extremely statistically significant, but a difference of about 3 beats/min from 72 has no clinical relevance. 4. The z test: SE $= \frac{1.2}{\sqrt{36}} = 0.2$ g/dL; $z = \frac{11.2-11.8}{0.2} = -3.0$; |z| > 1.96 (and > 2.576); p = 0.0027. H0 is rejected, a highly significant difference: the mean hemoglobin of the children in the community is lower than the mean of the general population. 5. (a) a type I error (false positive); (b) a type II error (false negative), favored by the small sample; (c) a correct decision (H0 true, not rejected). 6. $(a)\ 1 - 0.95^8 = 0.337$; $\frac{(b)\ 0.05}{8} = 0.00625$; $(c)\ 1 - 0.95^4 = 0.185$, Bonferroni threshold 0.05 / 4 = 0.0125. 7. Without correction: 4 (0.002, 0.009, 0.018, 0.035). Bonferroni (threshold 0.01): 2. Holm: 0.002 < 0.01; 0.009 < 0.0125; 0.018 > 0.0167, stop: 2. Benjamini-Hochberg (thresholds 0.01, 0.02, 0.03, 0.04, 0.05): 0.035 ≤ 0.04, so 4. 8. (a) T.DIST.2T(2.5, 12) = 0.028 (between the critical values 2.179 and 3.055); (b) two-sided 0.097, one-sided 0.048; (c) T.DIST.2T(2.9, 159) = 0.004 (the function takes the absolute value of t). 9. (a) The 95% CI contains 0, so p > 0.05. (b) An indeterminate effect: the interval includes both 0 and differences above the clinical threshold of 5 mmHg (up to 9 mmHg). (c) A larger study, sized for a difference of 5 mmHg (about 123-125 patients per group at σ = 14, 80% power). 10. (a) SE $= \frac{3.91}{12.649} = 0.309$; $t = \frac{29.04-30}{0.309} = -3.11$, p = 0.002. (b) The distribution of BMI is skewed (1.42) and non-normal; the Wilcoxon signed-rank test for the median against 30 (p < 0.001) or an analysis on log-transformed data is preferred. (c) Median 28.25 kg/m² (IQR 26.2-31.4), Wilcoxon test, p < 0.001: the typical BMI of the patients is below the threshold of 30 kg/m². 11. $n = \frac{2\times(1.96+0.84)^2\times14^2}{10^2} = \frac{2\times7.84\times196}{100} = 30.7$, that is, 31 patients per group (the exact calculation with the t distribution gives 32); $d = \frac{10}{14} = 0.71$.

Chapter 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

Explanations: In question 2, the degrees of freedom of the paired test are calculated from the number of pairs (25 - 1 = 24), not from the total number of measurements. In question 8, option c is false: ANOVA shows only that at least one mean differs; the C - B pair is identified by the post hoc test. Option e is false because $\eta^2 = 0.226$ shows a large effect, which does not depend on n. In question 10, the nonparametric alternative for repeated measures is the Friedman test; Kruskal-Wallis applies to independent groups.

Proposed exercises: 1. The differences: 0.6, 0.3, 0.1, 0.9, 0.6, 0.3, 0.8, 0.4; mD = 0.50%; sum of squared deviations 0.52; $s_D = \sqrt{\frac{0.52}{7}} = 0.273$; SED = 0.096; t = 5.19, df = 7, p = 0.0013 (critical t 2.365); 95% CI 0.27-0.73%; dz = 1.83. H0 is rejected: HbA1c decreased extremely significantly, on average by 0.5 percentage points (a study without a control group). 2. $s^2_{\text{pooled}} = \frac{24\times144+29\times196}{53} = 172.45$; spooled = 13.13; SE $= 13.13 \times \sqrt{\frac{1}{25}+\frac{1}{30}} = 3.56$; $t = \frac{7}{3.56} = 1.97$, df = 53, p = 0.054 (critical t 2.006); 95% CI -0.13 to 14.13 mmHg; d = 0.53. H0 is not rejected (borderline); the interval includes both 0 and large differences, so the result is indeterminate, and the study is underpowered for a difference of 7 mmHg. 3. Student: $s^2_{\text{pooled}} = \frac{11\times1.21+39\times10.24}{50} = 8.25$; SE = 0.946; t = -1.37, df = 50, p = 0.175. Welch: SE $= \sqrt{\frac{1.21}{12}+\frac{10.24}{40}} = 0.597$; t = -2.18, df = 48.9, p = 0.034. The pooled variance is dominated by the large group, which has a much larger variance (an SD ratio of almost 3); for the small group, with the small variance, the standard error is overestimated, and the Student test becomes too conservative. Welch's test is the correct one;

conclusion: a significant difference. 4. M = 126.33; SS between = $10 \times [(120-126.33)^2 + (128-126.33)^2 + (131-126.33)^2] = 646.7$(df 2, MS 323.3); SS within = 9 × (100 + 144 + 121) = 3,285 (df 27, MS 121.7); F = 2.66 < 3.35; p = 0.088; $\eta^2 = 0.16$. H0 is not rejected. 5. df between = 3, df within = 36, df total = 39; MS between = 120, MS within = 40; F = 3.0 > 2.87; p = 0.043; $\eta^2 = \frac{360}{1800} = 0.20$. H0 is rejected (a significant difference); the post hoc tests follow. 6. Welch: SE = $\sqrt{\frac{6.49^2}{52} + \frac{4.60^2}{84}} = \sqrt{0.810 + 0.252} = 1.030$; $t = \frac{3.68}{1.030} = 3.57$, df ≈ 82.7, p ≈ 0.0006. spooled = 5.40, d ≈ 0.68 (a medium-to-large effect). The Student variant would give t ≈ 3.86. Both values are compatible with the reported p = 0.001; the differences may come from the test used (possibly a nonparametric one) or from rounding. 7. Welch: SE = $\sqrt{\frac{7.0^2}{224} + \frac{5.8^2}{62}} = \sqrt{0.219 + 0.543} = 0.873$; $t = \frac{2.5}{0.873} = 2.87$, df ≈ 115, p ≈ 0.005 (a teaching calculation, not a result reported by the authors). The SD represents about 60% of the mean, and the mean minus 2 SD would be negative: the distribution of length of hospital stay is skewed to the right. On the individual data, the median with IQR and the Mann-Whitney test, or an analysis on log-transformed data, are preferred. 8. SED = $\frac{9.05}{12.649} = 0.716$; $t = \frac{4.78}{0.716} = 6.68$, df = 159, p < 0.001; 95% CI 3.36-6.19 mmHg; dz = 0.53. The mean decrease of about 5 mmHg is relevant at the population level, but the analysis of the whole sample has no control group; the effect of the interventions is assessed by comparing the arms. 9. (a) paired t / Wilcoxon signed-rank; (b) independent t (Student or Welch) / Mann-Whitney; (c) repeated measures ANOVA / Friedman; (d) independent t / Mann-Whitney; (e) one-way ANOVA / Kruskal-Wallis; (f) t for a single mean / Wilcoxon for a median. 10. HSD = $3.51 \times \sqrt{\frac{121.7}{10}} = 3.51 \times 3.49 = 12.2$ mmHg. The differences: 8, 11 and 3 mmHg, all below 12.2; no pair differs, consistent with the nonsignificant F. 11. The paired test uses the standard deviation of the individual differences, which depends on the correlation between the two measurements and cannot be deduced from the SDs at each time point. What would be needed: the number of patients (not stated in the abstract) and the mean difference with the SD of the differences or with its 95% CI (or the median of the differences with IQR, if the Wilcoxon test was used).

Chapter 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 Explanations: in question 2, the three values of 7 occupy positions 3, 4 and 5 and each receives the mean rank 4, and the sum of the ranks (21) confirms the calculation. In question 7, an observed frequency of 0 is not a problem if the expected frequencies are large (b), and unequal groups do not invalidate the test (e); what matters are the expected frequencies, the independence of the observations and the use of absolute numbers. In question 8, satisfaction is not a measure of glycemic control (d). In question 10, a small p value does not measure the strength of the association (c); the strength is read from φ = 0.28 or from the prevalence ratio. Proposed exercises: 1. The ranks: 4, 6, 6, 1.5, 8, 6, 3, 1.5; sum 36 = $\frac{8\times9}{2}$. 2. R1 = 11, R2 = 34; U1 = 11 - 10 = 1, U2 = 34 - 15 = 19; U = 1 ≤ 1, so H0 is rejected (exact p 0.032); anxiety is lower in the group with video information (median 4 vs. 8). 3. The differences: 8, 2, -2, 12, 6, 0, 7, 9; patient 6 has a difference of 0 and is excluded, n = 7; the ranks of the absolute values: 2 and 2 → 1.5; 6 → 3; 7 → 4; 8 → 5; 9 → 6; 12 → 7; W+ = 26.5, W- = 1.5; T = 1.5 ≤ 2, H0 is rejected (exact p 0.047), blood pressure decreased (median of the differences 7 mmHg). 4. Sum $40 + 60 + 71 = 171 = \frac{18\times19}{2}$; $H = \frac{\left[\frac{12}{18\times19}\right]\times(40^2+60^2+71^2)}{6} - 57 = 2.89 < 5.99$, H0 is not rejected (p = 0.24). 5. $\chi^2 F =$

$\left[\frac{12}{10\times3\times4}\right] \times (14^2 + 20^2 + 26^2) - 120 = 7.2 > 5.99$, H0 is rejected (p = 0.027); $W = \frac{7.2}{10\times2} = 0.36$. 6. Vaccinated: 10 with influenza / 190 without; unvaccinated: 24 / 126; expected 19.43 / 180.57 / 14.57 / 135.43; χ^2 = 11.83, p < 0.001 (Yates 10.60, p = 0.001); RR $= \frac{5.0\%}{16.0\%} = 0.31$ (vaccination reduces the risk by about 69% in this sample); φ = 0.18. 7. N = 70; column totals 20, 36, 14; expected in row 3: 2.57, 4.63, 1.80 (the cell in row 1, column 3 has exactly 5.0); 3 of 9 cells (33%) have E < 5, above the 20% limit, so the chi-square test is not applicable as such; solutions: merging row 3 with a neighboring category if this makes clinical sense, an exact test (Fisher-Freeman-Halton) or more data. 8. McNemar (paired data): $\chi^2 = \frac{(12-4)^2}{16} = 4.00$, p = 0.046; with correction (8 - 1)²/16 = 3.06, p = 0.080; exact variant p = 0.077; with only 16 discordant pairs the corrected or exact variant is preferred, so the result is borderline, nonsignificant. 9. Expected 70, 80, 34, 16; χ^2 = 0 + 0.31 + 0.47 + 0.06 = 0.85, df 3, p = 0.84; the observed distribution is compatible with the hypothesized proportions. 10. Expected 1.5 and 10.5 in each row, so Fisher's exact test; p = 0.217, H0 is not rejected; with 3 events in 24 patients the power is very low, and the result does not prove that the drug is safe. 11. Medians 4 (IQR 3-4) in B and 3 (IQR 2-3) in A; U = 2,000 (or 809, if the software reports the U of group A); p < 0.001; $r_{\text{rb}} = 1 - \frac{2\times809}{2809} = 0.42$ (a medium-to-large effect).

Chapter 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 Explanations: in question 4, the overall risk is 1 - 0.95³ = 0.143, that is, about 14%, not 15% (3 x 5% is only an approximation). In question 7, a significant Shapiro-Wilk test at n = 300 does not require a nonparametric test (c), and choosing the test according to the p value is a procedural error (e). In question 8, changes within the arms do not prove efficacy (d); only the comparison between arms shows the superiority of an intervention. Proposed exercises: 1. The variable: number of visits (discrete numerical, strongly skewed); two independent groups; Mann-Whitney U (alternative: the t-test, only if the groups are large and the skewness moderate; at an advanced level, Poisson or negative binomial regression); H0: the distributions of the number of visits are identical, H1: they differ. 2. SpO2 (numerical), the same 25 patients at two time points, paired data; the paired t-test if the differences are normal, otherwise Wilcoxon (SpO2 is bounded above and often skewed); H0: μd = 0. 3. The variable vaccinated yes/no (nominal), three independent groups; chi-square on a 3 x 2 table, df = 2 (the expected frequencies are large); H0: p1 = p2 = p3. 4. The score 0-21 (numerical or ordinal), the same 30 residents under three conditions, repeated measures; repeated measures ANOVA if the distribution allows it, otherwise Friedman, followed by post hoc tests with correction. 5. Agreement between two raters on a binary variable: Cohen's kappa coefficient (McNemar's test shows only whether one radiologist says "pneumonia" systematically more often than the other, and the chi-square test shows association, not agreement). 6. The aim is to predict a binary outcome over time: if all patients were followed up completely for 10 years, logistic regression (OR); if the duration of follow-up varies or there are losses to follow-up, Cox regression (HR) on the time to diabetes. 7. 1 - 0.95³ = 0.143 (14.3%); Bonferroni threshold 0.05/3 = 0.0167; no comparison remains significant (0.03 and 0.04 > 0.0167); the correct approach was ANOVA followed by Tukey. 8. The smallest expected frequency = $\frac{12\times10}{60} = 2.0 < 5$, so Fisher's exact test. 9. Women 208.08 ± 35.37 mg/dL$(n\,86)$, men $210.91 \pm 38.82(n\,74)$; SW p = 0.61 and 0.93; Levene p = 0.65; Student t = -0.48, df 158, p = 0.63 (Welch p = 0.63); the F - M difference -2.82 mg/dL (95% CI -14.50 to 8.85); d = -0.08; Mann-Whitney p = 0.76; we do not reject H0, cholesterol does not differ between the sexes. 10. Rural 10/65 (15.4%), urban 19/95 (20.0%); the smallest expected frequency 11.8; χ^2 = 0.55, df 1, p = 0.46 (Yates p = 0.59, Fisher p

= 0.53); we do not reject H0, diabetes is not associated with the area of residence in this sample. 11. Open answer; check the consistency between the design, the type of variable and the test.

Practical Session 8 - Expected Answers (for the Instructor)

L1. Waist circumference by sex. Women 86.34 ± 11.67 cm (median 84, IQR 78-93), men 91.09 ± 12.19 cm (median 90, IQR 83-97). SW: women p = 0.008, men p = 0.005 (mild skewness), but the groups are large (86 and 74), so both tests are acceptable if the choice is justified. Levene p = 0.80. Student's t-test t = -2.52, df 158, p = 0.013 (Welch t = -2.51, p = 0.013); the F - M difference -4.76 cm (95% CI -8.50 to -1.01); d = -0.40. Mann-Whitney p = 0.0096, r_rb = 0.24. Conclusion: we reject H0; men have a waist about 5 cm larger on average (a small-to-medium effect); the clinical thresholds for waist circumference differ by sex.

L2. Triglycerides by diabetes status. Without diabetes: median 154 mg/dL (IQR 104.5-209.5), SW p < 0.001; with diabetes: median 143 (IQR 112-179), SW p = 0.11. Appropriate test: Mann-Whitney (skewed distribution, unequal groups of 131 and 29); alternative: the t-test on log-transformed triglycerides. U = 2,033.5 (reported for the group without diabetes), p = 0.55; r_rb = 0.07; Welch on the raw values p = 0.30. We do not reject H0. In (g), the comment that triglycerides are usually higher in diabetic patients in real populations is accepted; with 29 cases and simulated data, the difference does not appear.

L3. Baseline HbA1c by arm. 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. We do not reject H0. The key point of the problem is the discussion: in a randomized trial, baseline differences are due to chance by definition; the baseline table is described, not tested (CONSORT). What matters is the clinical size of any difference (negligible here: 0.06 percentage points).

L4. Baseline cholesterol by arm. 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, $\eta^2 = 0.021$; Kruskal-Wallis p = 0.19. We do not reject H0. The same discussion as for L3; the largest difference between the arms (13 mg/dL) is not clinically important for the glycemic outcomes.

L5. Weight loss by arm. A 0.71 ± 2.48 kg (median 1.1); B 2.84 ± 2.36 (median 2.3); C 3.37 ± 2.39 (median 3.55); SW by arm p = 0.70, 0.26, 0.17; Levene p = 0.79. ANOVA F(2, 157) = 18.25, p < 0.001, $\eta^2 = 0.189$. Tukey: B - A 2.13 kg (95% CI 1.02-3.24; p < 0.001); C - A 2.66 kg (95% CI 1.56-3.77; p < 0.001); C - B 0.53 kg (95% CI -0.57 to 1.63; p = 0.49). Kruskal-Wallis H = 27.50, p < 0.001. Conclusion: both arms with the app led to a greater weight loss than counseling alone (by 2.1-2.7 kg), and teleconsultation did not bring a significant additional benefit over the app alone.

L6. BMI by smoking status. Former smokers median 28.0 (IQR 26.5-30.9); current smokers 28.3 (26.3-31.4); never smokers 28.35 (26.1-31.08). SW p < 0.05 in all groups (0.029, < 0.001, 0.003); Levene p = 0.048. Appropriate test: Kruskal-Wallis (skewness in all groups, and the group of former smokers has only 33 patients); H = 0.46, df 2, p = 0.80. ANOVA p = 0.35 and Welch p = 0.38 lead to the same conclusion. We do not reject H0.

L7. Satisfaction by sex. Median 4 (IQR 3-4) in both sexes; Mann-Whitney U = 3,221, p = 0.89; r_rb = 0.01. We do not reject H0. Points are deducted for reporting the means (3.56 and 3.49) as the main description and for using the t-test.

L8. Age by area of residence. Rural 57.38 ± 9.08 years (n 65), urban $55.55 \pm 8.52 (n\ 95)$; SW p = 0.87 and 0.69; Levene p = 0.64; Student t = 1.30, df 158, p = 0.19 (Welch t = 1.29, p = 0.20); the rural - urban difference 1.84 years (95% CI -0.98 to 4.66); d = 0.21; Mann-Whitney p = 0.17. We do not reject H0.

L9. Weight loss ≥ 5% by sex. Women 23/86 (26.7%), men 15/74 (20.3%). The expected frequencies: women 20.4 with weight loss and 65.6 without; men 17.6 and 56.4. $\chi^2 = 0.92$, df 1, $p = 0.34$ (Yates $p = 0.44$, Fisher $p = 0.36$); $\varphi = 0.08$. We do not reject H0.

L10. Smoking by area of residence. Rural: former smokers 14 (21.5%), current smokers 16 (24.6%), never smokers 35 (53.8%); urban: 19 (20.0%), 37 (38.9%), 39 (41.1%). The smallest expected frequency 13.4. $\chi^2 = 3.80$, df 2, $p = 0.15$; Cramér's V = 0.15. We do not reject H0; the trend (more smokers in urban areas) does not reach the significance level.

L11. Smoking by arm. A: former smokers 5 (9.4%), current smokers 18 (34.0%), never smokers 30 (56.6%); B: 11 (20.8%), 18 (34.0%), 24 (45.3%); C: 17 (31.5%), 17 (31.5%), 20 (37.0%). The smallest expected frequency 10.9. $\chi^2 = 8.60$, df 4, $p = 0.072$; Cramér's V = 0.16. We do not reject H0. The expected discussion: the imbalance (for example, 9.4% former smokers in A vs. 31.5% in C) is due to chance, a result of randomization; it does not "invalidate" the study; if smoking influences the primary outcome, it can be included as an adjustment variable in the analysis, a decision that should be made in advance.

L12. Blood glucose ≥ 100 at 0 and 6 months, McNemar in each arm. Arm A: ≥ 100 at both time points 32, only at baseline 3, only at 6 months 1, at neither 17; proportion 35/53 (66.0%) at baseline and 33/53 (62.3%) at 6 months; McNemar $\chi^2 = 1.00$, $p = 0.32$ (with correction 0.25, $p = 0.62$; exact $p = 0.63$); we do not reject H0. Arm B: at both 19, only at baseline 15, only at 6 months 0, at neither 19; 34/53 (64.2%) vs. 19/53 (35.8%); $\chi^2 = 15.00$, $p < 0.001$ (with correction 13.07, $p < 0.001$; exact $p < 0.001$); we reject H0. The discussion: the within-arm results describe the change in each arm; the comparison of efficacy is made between the arms (for example, chi-square on the proportion of those who fell below 100, or an analysis of the decrease in blood glucose by arm). Points are deducted for the ordinary chi-square test on the table "time point x glucose ≥ 100".

L13. Blood glucose and HbA1c at baseline. The scatter plot shows a linear relationship, without important extreme values; HbA1c is normally distributed (SW $p = 0.30$), blood glucose shows a slight departure (SW $p = 0.037$ at n = 160, skewness 0.37). Pearson correlation $r = 0.738$, $p < 0.001$ (Spearman rho = 0.727, $p < 0.001$); $R^2 = 0.54$. We reject H0: $\rho = 0$. A strong, positive association, physiologically expected (both measure glycemia, in the short term and in the long term); it is not a measure of agreement between the two, since they are different quantities.

L14. Blood glucose at 0, 3 and 6 months in arm A. Means 107.8 / 105.7 / 104.8 mg/dL; medians 111 / 104 / 105. SW at each time point $p = 0.20$, 0.62, 0.71 (normal); Mauchly's test $p = 0.020$ (sphericity violated, Greenhouse-Geisser ε 0.88). The justified main analysis: repeated measures ANOVA with the Greenhouse-Geisser correction: $F(2, 104) = 2.73$, $p = 0.078$ (without correction $p = 0.070$); we do not reject H0. The alternative, Friedman: $\chi^2 = 6.04$, df 2, $p = 0.049$; the Wilcoxon comparisons with Bonferroni do not remain significant (the smallest adjusted p 0.075). The complete answer recognizes the discordance, reports the test decided in advance and concludes that in arm A there is no convincing evidence of a decrease in blood glucose over time. Points are deducted for choosing the Friedman test justified only by the fact that it gives $p < 0.05$.

L15. Age and FINDRISC. The FINDRISC score is a score with integer values and mild skewness (SW $p = 0.034$); Spearman (rho = 0.456, $p < 0.001$) or Pearson ($r = 0.442$, $p < 0.001$) is accepted, with justification. We reject H0. The expected discussion: age is one of the components of the FINDRISC score (patients receive points for age groups), so part of the correlation is "built in" by the definition of the score; the correlation does not show that age adds information independent of the other components of the score.

Chapter 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

Explanations: In question 3, $R^2 = 0.60^2 = 0.36$, so the proportion due to other factors is 1 - 0.36 = 0.64 (64%); the option "40%" appears when r, not r², is subtracted from 1. In question 6, option c is false because r measures only the linear relationship (a U-shaped relationship can have an r close to zero), and option d is false because r does not depend on the units of measurement. In question 8, $R^2 = 0.088^2 = 0.0077$, that is, below 1%, not 8.8%; the correlation is significant only because of the large sample and does not demonstrate a causal relationship. In question 9, normality is required of the residuals, not of the X variable.

Proposed exercises: 1. mx = 53, my = 130.33; Sxy = 660, Sxx = 868, Syy = 553.33; $r = \frac{660}{\sqrt{868 \times 553.33}} = 0.952$ (a strong direct correlation); t = 6.24, df 4, critical t 2.776, p = 0.003, H0 is rejected; $b = \frac{0.760 \text{ mmHg}}{\text{year}}$, a = 90.03; SBP = 90.03 + 0.760 x age; at 60 years, 135.7 mmHg; $R^2 = 0.907$. 2. n = 40: $t = \frac{0.30 \times \sqrt{38}}{\sqrt{0.91}} = 1.94$, df 38, critical t 2.024, p = 0.060, H0 is not rejected. n = 200: t = 4.43, p < 0.001, H0 is rejected. The same r becomes significant in a larger sample; according to Colton, the correlation is weak in both cases. 3. $z = 0.5 \times \ln\left(\frac{1.62}{0.38}\right) = 0.725$; SE = $\frac{1}{\sqrt{47}} = 0.146$; CI on the z scale 0.439-1.011; 95% CI for ρ: 0.41-0.77. The upper limit exceeds 0.75, so the correlation in the population could be strong, but the lower limit (0.41) is also compatible with a weak correlation. 4. VAS ranks: 1, 4, 2, 6, 5, 7, 3; anxiety ranks: 2, 4, 1, 7, 6, 5, 3; di: -1, 0, 1, -1, -1, 2, 0; $\sum {d_i}^2 = 8$; $r_s = 1 - \frac{6 \times 8}{7 \times 48} = 0.857$, p = 0.014; a strong direct correlation. 5. $R^2 = 0.738^2 = 0.545$: about 54.5% of the variation in HbA1c is linearly associated with the variation in fasting plasma glucose; $1 - R^2 = 0.455$: 45.5% of the variation in HbA1c is due to other factors. According to Colton, a moderate correlation, very close to the 0.75 threshold. 6. 0.81 strong direct; -0.573 moderate inverse; 0.256 weak direct; -0.088 no correlation; 0.729 moderate direct; -0.832 strong inverse; 0.107 no correlation; 0.50 moderate direct (at the boundary between weak and moderate). 7. The mean of the differences (bias) = 3.0 mmHg; SD = 3.46 mmHg; limits of agreement $3.0 \pm 1.96 \times 3.46 = -3.8$ to 9.8 mmHg. The correlation coefficient would show only whether the patients are ranked similarly; a systematic difference of a few mmHg would not lower r. 8. 5 years: 1.11^5 = 1.69; 10 years: 1.11^10 = 2.84 (with the unrounded coefficient, about 2.9). It is not correct: the OR refers to odds (p / (1 - p)), not to probabilities; the two are close only for rare events. 9. A nonsmoking woman, 45 years old, BMI 27: 64.51 + 1.389 x 27 + 0.449 x 45 - 1.119 = 121.1 mmHg. A male smoker, 65 years old, BMI 35: 64.51 + 1.389 x 35 + 0.449 x 65 - 1.481 + 6.738 = 147.6 mmHg. The difference of 26.5 mmHg: BMI contributes 8 x 1.389 = 11.1 mmHg (the most), age 20 x 0.449 = 9.0 mmHg, smoking 6.738 + 1.119 = 7.9 mmHg, sex -1.5 mmHg. 10. (a) Each additional centimeter is associated with a weight higher on average by 0.95 kg. (b) 1 - 0.38 = 0.62, that is, 62%. (c) At 170 cm: -95 + 0.95 x 170 = 66.5 kg. At 120 cm the formula gives 19 kg, an absurd extrapolation outside the range 155-195 cm; the model does not apply.

Chapter 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

Explanations: In question 1, the study starts from the disease (cases and controls), so it is a case-control study and only the OR is calculated. In question 2, the risk difference is 0.20 - 0.10 = 0.10, and NNH $= \frac{1}{0.10} = 10$; the option "5" appears when you divide by the risk in the exposed. In question 3, RR = 0.70 means a relative risk reduction of 30% (1 - 0.70), not of 70%. In question 4,

the disease is common (prevalences of 45% and 18.75%), so the OR (3.55) is larger than the prevalence ratio (2.40). In question 8, a confounder does not lie on the causal pathway; an intermediate variable (mediator) is not adjusted for as a confounder. In question 10, the p value is not one of the Bradford Hill criteria.

Proposed exercises: 1. A case-control study (selection by infertility). a = 89 (infertile women with an IUD), b = 640 (controls with an IUD), c = 194 (infertile women without an IUD), d = 3,193 (controls without an IUD). $OR = \frac{89\times3193}{640\times194} = 2.29(;\ CI\,95\%\ 1.76 - 2.98)$; χ^2 = 39.35, p < 0.001. The odds of having used an IUD are 2.3 times higher in infertile women; a significant association. 2. a = 23, b = 84, c = 117, d = 800; $OR = \frac{23\times800}{84\times117} = 1.87(;\ CI\,95\%\ 1.14 - 3.09)$; χ^2 = 6.20, p = 0.013. Down syndrome is rare in the population, so, if the sample was built as a case-control study, the OR approximates the RR; the RR cannot be calculated directly from this design. 3. Arm B: 16 / 53 = 30.2%; arm A: 6 / 53 = 11.3%; RR = 2.67 (95% CI 1.13-6.29); absolute difference 18.9 percentage points (95% CI 3.9-33.9); NNT = 5.3, rounded up to 6. A result very close to that for arm C vs. A (RR 2.62; NNT 6). 4. a = 10, b = 43, c = 19, d = 88. $PR = \frac{18.9\%}{17.8\%} = 1.06(;\ CI\,95\%\ 0.53 - 2.12)$; OR = 1.08 (95% CI 0.46-2.52); χ^2 = 0.03, p = 0.86. H0 is not rejected; in mg2026, smoking is not associated with diabetes. The clinical conclusion: the data do not show an association, but the wide interval does not exclude a moderate effect either. 5. a = 21, b = 33, c = 20, d = 86. $PR = \frac{38.9\%}{18.9\%} = 2.06(;\ CI\,95\%\ 1.23 - 3.46)$; OR = 2.74 (95% CI 1.32-5.69); χ^2 = 7.52, p = 0.006. The OR exceeds the PR because the outcome is common (25.6% have SBP ≥ 140 mmHg). 6. Exposed: 10 / 400 = 2.5 cases per 100 person-years; unexposed: 8 / 800 = 1.0 per 100 person-years; IRR = 2.5. 7. P ≈ I x D = 0.005 x 4 = 0.02, that is, 2%. If the duration doubles to 8 years, at the same incidence the prevalence rises to about 4%, even though the treatment is an advance. 8. Risk difference = 6% - 2% = 4 percentage points; RR = 3; $NNH = \frac{1}{0.04} = 25$: for every 25 patients treated, one additional adverse effect occurs compared with placebo. 9. AFe = (3 - 1) / 3 = 0.667 (66.7%); $PAF = \frac{0.40\times2}{1+0.40\times2} = \frac{0.8}{1.8} = 0.444(44.4\%)$. 10. Stratum $OR\ 1 = \frac{20\times60}{80\times10} = 1.5$; stratum $OR\ 2 = \frac{30\times60}{20\times60} = 1.5$. The combined table: a = 50, b = 100, c = 70, d = 120, crude OR = 0.86. $OR(MH) = \frac{\frac{20\times60}{170}+\frac{30\times60}{170}}{\frac{80\times10}{170}+\frac{20\times60}{170}} = \frac{17.65}{11.76} = 1.5$. The stratum-specific measures are equal to each other and differ greatly from the crude one: the stratification factor is a confounder, which in the crude analysis even reverses the direction of the association. 11. a = 10, b = 55, c = 19, d = 76; OR = 0.73 (95% CI 0.31-1.69); χ^2 = 0.55, p = 0.46. The interval contains 1: it cannot be claimed that rural residence is protective; the result is nonsignificant.

Chapter 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

Explanations: In question 2, out of 1,000 patients, 100 are diseased (80 positive tests) and 900 healthy (90 false positives), so $PPV = \frac{80}{170} = 47\%$; option a) confuses the PPV with sensitivity. In question 5, BF10 = 0.05 means BF01 = 20, that is, strong evidence for H0; the Bayes factor is not a p value. In question 7, option c) is false: likelihood ratios are derived from sensitivity and specificity, which in principle do not depend on prevalence (but can change with the spectrum of patients). In question 9, verification bias is a design bias of diagnostic studies (not all patients receive the reference standard), not a cognitive error at the bedside.

Proposed exercises:

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: $\frac{0.75 \times 0.12}{0.75 \times 0.12 + 0.20 \times 0.88} = \frac{0.090}{0.266} = 0.338$.
2. Out of 10,000 women: 80 with cancer (72 positive, 8 false negatives), 9,920 without cancer (about 694 false positives, 9,226 negatives). PPV $= \frac{72}{766} = 9.4\%$; NPV $= \frac{9226}{9234} = 99.9\%$. For each cancer found, about 766/72 = 10.6, that is, roughly 11 women with a positive mammogram undergo further investigation.
3. 0.40 -> 0.40/0.60 = 0.667; 0.05 -> 0.0526; odds 3 -> 3/4 = 0.75; odds 0.25 -> 0.25/1.25 = 0.20.
4. (a) $LR^+ = \frac{0.703}{0.340} = 2.07$; $LR^- = \frac{0.297}{0.660} = 0.45$. (b) Pre-test odds 0.429; after a dNLR above the threshold 0.429 x 2.07 = 0.886, probability 47.0%; below the threshold 0.429 x 0.45 = 0.193, probability 16.2%. The test shifts the probability only a little, in agreement with the AUC of 0.624 (satisfactory on the scale from the lectures). (c) Three decimal places (70.312%) suggest a precision that a sample of 336 patients does not provide; one decimal place (70.3%) and a confidence interval would have been appropriate.
5. $LR^- = \frac{0.07}{0.86} = 0.081$. At 33.8%: 0.512 x 0.081 = 0.042, post-test probability 4.0%. At 5%: 0.0526 x 0.081 = 0.0043, post-test probability 0.43%.
6. In series: Se = 0.90 x 0.85 = 0.765; Sp = 1 - 0.20 x 0.05 = 0.99. In parallel: Se = 1 - 0.10 x 0.15 = 0.985; Sp = 0.80 x 0.95 = 0.76. Triage in the emergency department: parallel (we do not miss diseased patients). Confirmation before a toxic treatment: series (few false positives).
7. Products: migraine 0.70 x 0.05 x 0.05 x 0.70 = 0.001225; tension-type 0.25 x 0.02 x 0.10 x 0.10 = 0.00005; SAH 0.05 x 0.80 x 0.60 x 0.70 = 0.0168. Sum 0.018075. Posterior probabilities: migraine 6.8%; tension-type 0.3%; SAH 92.9%. Decision: emergency imaging (head CT), regardless of reservations about the independence of the signs.
8. Odds at the threshold 0.10/0.90 = 0.111. The testing threshold: 0.111/4 = 0.0278, that is, 2.7%. The test-treatment threshold: 0.111/0.25 = 0.444, that is, 30.8%.
9. (a) SOFA: 0.771 good, 0.873 very good; NEWS2: 0.892 very good, 0.729 good; APACHE II: 0.698 satisfactory, 0.806 at the boundary between good and very good (0.81 after rounding). (b) NEWS2 on admission, SOFA on day five; the performance of a score depends on when it is applied. (c) Logistic regression produces odds ratios (ORs); HRs come from a Cox model, which requires time-to-event data. Reporting HRs from a logistic regression is a point to raise in critical reading.
10. (a) $BF_{01} = \frac{1}{0.16} = 6.25$: moderate evidence for the absence of a correlation. (b) Posterior odds 1 x 0.16 = 0.16; probability 0.16/1.16 = 13.8%.
11. P: previously uninvestigated adults at cardiometabolic risk in family medicine; I: HbA1c (for example, a threshold of ≥ 6.5%); C: the OGTT as the reference standard; O: sensitivity, specificity and likelihood ratios for type 2 diabetes. Appropriate design: a cross-sectional diagnostic accuracy study in consecutive patients, all of whom receive both tests, reported according to STARD 2015.

Chapter 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

Explanations: In question 1, option b) is the wrong formula that sometimes appears in notes; d contains the true negatives, that is, people classified correctly. In question 4, the AUC is neither the accuracy nor the sensitivity; it is the probability that, in a randomly chosen diseased-healthy pair, the diseased person has the higher value. In question 6, option c) describes verification bias,

and option e) is false because a larger sample with the same inappropriate spectrum only gives a more precise estimate of a biased value. In question 8, the error rate is (7 + 33)/160 = 0.25, and the specificity is 98/131 (98/105 is the NPV).

Proposed exercises:

1. a = 45, b = 5, c = 30, d = 120. Prevalence 50/200 = 0.25; Se 0.90; $\frac{\text{Sp}\,120}{150} = 0.80$; $\frac{\text{PPV}\,45}{75} = 0.60$; $\frac{\text{NPV}\,120}{125} = 0.96$; $\frac{\text{Acc}\,165}{200} = 0.825$; $\frac{\text{ER}\,35}{200} = 0.175$; J 0.70; LR+ 4.5; LR- 0.125; $\frac{\text{DOR}(45\times120)}{5\times30} = 36$.
2. a = 80, b = 20; from $\frac{80}{80+c} = \frac{4}{5}$, c = 20; from $\frac{d}{20+d} = \frac{9}{10}$, d = 180. Se 0.80; $\frac{\text{Sp}\,180}{200} = 0.90$; N = 300.
3. d = 150 and Sp = $\frac{3}{4}$, so c + d = 200 and c = 50; from $\frac{150}{b+150} = \frac{15}{16}$, b = 10; a = 40. $\frac{\text{Se}\,40}{50} = 0.80$; $\frac{\text{PPV}\,40}{90} = 0.444$; N = 250.
4. 100 diseased: 75 TP, 25 FN; 900 non-diseased: 855 TN, 45 FP. $\frac{\text{PPV}\,75}{120} = 62.5\%$; $\frac{\text{NPV}\,855}{880} = 97.2\%$.
5. HPV: a = 17, b = 1, c = 19, d = 27; $\frac{\text{Acc}\,44}{64} = 68.8\%$; J 0.531. p16/Ki-67: a = 12, b = 6, c = 7, d = 39; $\frac{\text{Acc}\,51}{64} = 79.7\%$; J 0.514. HPV is better suited for ruling out (Se 94.4%, LR- 0.095), p16/Ki-67 for confirmation (Sp 84.8%, LR+ 4.38). The almost equal Youden indices hide these different roles.
6. Trapezoids: 0.010 + 0.025 + 0.070 + 0.170 + 0.282 + 0.297 = 0.854; very good (0.81-0.90).
7. Po = $\frac{175}{200} = 0.875$; A "pneumonia" 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, substantial agreement.
8.
   (a) Deaths 159 x 0.2515 = 40, survivors 119; TP 0.878 x 40 = 35, FN 5; TN 0.7372 x 119 = 88, FP 31. (b) PPV = $\frac{35}{66} = 53.0\%$, close to the reported 53.73%. (c) Wilson for 35/40: about 73.9-94.5%, almost 80 times wider than the reported interval; the interval in the abstract cannot come from a usual method for a proportion estimated from 40 deaths.
9. $\frac{\text{Se}\,40}{50} = 80\%$; $\frac{\text{Sp}\,360}{450} = 80\%$; precision 40/130 = 30.8%; $F_1 = \frac{2\times0.308\times0.80}{0.308+0.80} = 0.44$; $\frac{\text{Acc}\,400}{500} = 80\%$. A classifier that declares everything benign would have an accuracy of 450/500 = 90%, higher than the algorithm, which shows why accuracy is not used on its own with imbalanced classes; the low precision reflects the prevalence of 10%.
10. 0.624 satisfactory; 0.752 good. The test of H0: AUC = 0.5 shows only that the marker discriminates better than chance; in 336 patients, even modest discrimination becomes significant. Clinical usefulness is judged by the size of the AUC, the likelihood ratios at the threshold and the consequences of the decisions.
11. Pre-test odds 0.111; $\text{LR}^+ = \frac{0.801}{0.143} = 5.60$, so PPV = $\frac{0.622}{1.622} = 38.4\%$; $\text{LR}^- = \frac{0.199}{0.857} = 0.232$, post-test probability 2.5%, so NPV 97.5%.
12. $\text{LR}^+ = \frac{0.379}{0.053} = 7.1$; $\text{LR}^- = \frac{0.621}{0.947} = 0.66$. A result ≥ 126 mg/dL increases the odds about 7-fold (good for confirmation), but a result below the threshold hardly changes the probability (LR- 0.66), so the threshold misses most of the diabetic patients identified by the OGTT and is not suitable for screening.

Chapter 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

Explanations: In question 4, NNT = $\frac{1}{0.183} = 5.5$ and it is always rounded up, to 6; the values 3 and 29 are the limits of the confidence interval, not the point estimate. In question 6, option c is

false: ITT dilutes the effect with nonadherent patients, so it tends to underestimate the biological effect, not to overestimate it; option e describes the per-protocol analysis. In question 9, SBP 1540 and a height of 1.72 are errors in numerical variables and are detected through the minimum, the maximum and the box plot, not through a frequency table of the categorical variables. In question 10, Bartlett's test must be significant (the items are correlated), and alpha measures internal consistency, not validity.

Proposed exercises: 1. $n = \frac{2\times(1.96+1.28)^2\times12^2}{8^2} = \frac{2\times10.50\times144}{64} = 47.2$, so 48 patients per group; with 10% losses, 48/0.9 = 53.3, so 54 per group. 2. p1 = 0.35, p2 = 0.20, $\bar{p} = 0.275$; $n = \frac{\left[1.96\times\sqrt{2\times0.275\times0.725}+0.84\times\sqrt{0.35\times0.65+0.20\times0.80}\right]^2}{0.15^2} = 137.9$, so 138 patients per group. 3. Design effect = 1 + (20 - 1) x 0.02 = 1.38; 64 x 1.38 = 88.3, so 89 patients per arm (about 5 practices of 20 patients per arm). 4. $\alpha = \frac{5}{4}\times\left(1-\frac{5.5}{18.0}\right) = 1.25\times0.694 = 0.868$; good internal consistency. 5. Risk 30/150 = 20% with the program, 45/150 = 30% in the control group; RR = 0.67 (relative risk reduction 33%); absolute risk reduction 10 percentage points; NNT $= \frac{1}{0.10} = 10$. 6. 1 - 0.95^8 = 0.337 (about 34%); the Bonferroni threshold is 0.05/8 = 0.00625 for each test. 7. a) the lower limit -2.4 is above -3: non-inferiority demonstrated; b) the lower limit -4.1 falls below -3: an inconclusive result, non-inferiority is not demonstrated (but neither is inferiority); c) the whole interval is above 0: non-inferiority and even superiority, if testing for superiority was planned in the protocol. 8. a) crossover, analysis for paired data; b) cluster randomized (schools); c) a single-arm pilot, repeated measures (repeated measures ANOVA or Friedman); d) 2 x 2 factorial; e) observational cohort (the exposure was not assigned by the investigator). 9. 1.75: wrong unit (meters); 1680: an extra digit; "1 m 70": text in a numerical column. All are checked against the source document and corrected only from there; if the source is not available, the value is declared missing. Prevention: validation in the form, "number between 120 and 220", with the message "in centimeters". The values 168, 172, 159, 165, 181 are plausible. 10. 3/300 = 0.01, that is, the upper limit of the 95% CI is about 1%; the adverse effect could occur in up to about 1 case per 100 patients without having been observed. Rarer effects can be detected only in phase IV, in thousands or tens of thousands of patients. 11. $p = \frac{16}{53} = 30.2\%$; SE $= \sqrt{\frac{0.302\times0.698}{53}} = 0.063$; Wald 95% CI 17.8-42.5%; Wilson 95% CI 19.5-43.5% (asymmetric, recommended for small samples or proportions far from 50%). 12. RR $= \frac{\frac{16}{53}}{\frac{6}{53}} = 2.67$(; CI 95% $1.13 - 6.29$); risk difference 18.9 percentage points (95% CI 3.9-33.9); NNT = 5.3, that is, 6. The results are practically identical to those for C vs. A (RR 2.62, NNT 6): on this outcome, teleconsultation adds no benefit over the app (RR C vs. B 0.98, 95% CI 0.55-1.75).

Chapter 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

Explanations: In question 2, options a and b are the classic traps: the confidence interval concerns the population mean, not the individual values, and the sample mean always lies within the interval (it is its center). In question 9, of 150 diseased people 135 are positive, and of 850 healthy people 170 are false positives; PPV $= \frac{135}{305} = 44.3\%$. In question 13, changing the test after seeing the result (option d) is fishing for significance, and option a uses the wrong wording "H0 is

accepted". In question 14, RR $= \frac{\frac{18}{150}}{\frac{10}{250}} = \frac{0.12}{0.04} = 3.0$; 3.27 is the OR. In question 21, the intercept corresponds to a BMI of 0, an impossible value located far from the observed range (23.5-48.3), so it has no clinical interpretation. In question 24, the Youden threshold maximizes Se + Sp, not the PPV. In question 29, r = 0 rules out only a linear relationship; a curvilinear relationship (for example, U-shaped) can give an r close to 0.

Proposed Exercises: 1. Total 250: nonsmokers 120 (48.0% of the total; 52.2% valid), smokers 70 (28.0%; 30.4%), former smokers 40 (16.0%; 17.4%), missing 20 (8.0%); the valid percentages are calculated on 230. 2. OR $= \frac{56 \times 96}{64 \times 24} = 3.5$; 95% CI 1.97-6.21; $\chi^2 = 19.20$, $p < 0.001$. The odds of smoking are 3.5 times higher in patients with COPD. 3. Risk 30/200 = 15% in the exposed, 15/300 = 5% in the unexposed; RR = 3.0 (95% CI 1.66-5.43); RD = 10 percentage points; NNH = 10; $\chi^2 = 14.65$, $p < 0.001$. 4. Diseased 200: TP 180, FN 20; healthy 9,800: TN 9,310, FP 490. PPV $= \frac{180}{670} = 26.9\%$; NPV $= \frac{9310}{9330} = 99.8\%$. At a low prevalence, almost three quarters of the positive results are false positives, even though the test has an Sp of 95%; a positive result requires confirmation, while a negative result rules out the disease almost certainly. 5. SE $= \frac{18}{\sqrt{36}} =$ 3 mg/dL; CI $95\% = 220 \pm 2.03 \times 3 = 213.9 - 226.1$ mg/dL. 6. $R^2 = 0.36$; $t = \frac{0.6 \times \sqrt{25}}{\sqrt{1-0.36}} = \frac{3}{0.8} =$ $3.75 > 2.06$, so r is significant ($p \approx 0.001$). 7. SBP = 89.47 + 1.435 x 27 = 128.2 mmHg. A BMI of 15 lies outside the observed range (23.5-48.3 kg/m²), and extrapolating the line beyond the data is not justified. 8. Pre-test odds = 0.30/0.70 = 0.429; post-test odds = 0.429 x 6 = 2.57; post-test probability = 2.57/3.57 = 72%. 9. a) one-way ANOVA (with Tukey); b) the Wilcoxon signed-rank test for paired data (the pain score is ordinal); c) chi-square (Fisher when expected frequencies are small); d) Pearson (or Spearman) correlation and linear regression; e) the Mann-Whitney test; f) the Friedman test. 10. Model answer: "Mean HDL cholesterol was 53.9 mg/dL (SD 11.7) in women and 44.4 mg/dL (SD 9.7) in men. The difference of 9.5 mg/dL (95% CI 6.2-12.9) was statistically significant (Welch t-test, chosen because of unequal variances, Levene p = 0.034; t = 5.63, df 157.7, $p < 0.001$)." 11. Correct version: "The prevalence of diabetes was 33.3% in obese patients and 10.4% in non-obese patients ($\chi^2 = 12.70$, $p < 0.001$; prevalence ratio 3.21, 95% CI 1.64-6.31; OR 4.32, 95% CI 1.86-10.03). Because this is a cross-sectional study, the result shows an association, not a causal relationship." The original errors: causality, confusing the OR with risk, no CI. 12. The standard deviations are almost equal to the means, for a variable that cannot be negative: the distribution is strongly skewed. Recommendations: describe the data with the median and IQR, compare the groups with the Mann-Whitney test (or a t-test on log-transformed data), and report the difference in medians.

Appendix B. Glossary

The glossary contains the terms used in the course, in alphabetical order of the English terms. Each term is followed in parentheses by the abbreviation used in the book, where there is one, and by the Romanian term; then come a short definition and the chapter in which the concept is explained in detail. The definitions are worded for revision; they do not replace the explanations, examples and pitfalls discussed in the chapters.

- 2 x 2 table (two-by-two table; Rom. "tabel 2 x 2"): a contingency table with two rows and two columns (cells a, b, c, d), used both in risk analysis (exposure x disease) and in the evaluation of diagnostic tests (disease x test result). (Ch. 10)
- 68-95-99.7 rule (empirical rule; Rom. "regula 68-95-99,7"): in a normal distribution, approximately 68.3% of the values lie within one standard deviation of the mean, 95.4% within two and 99.7% within three. Exactly 95% lie within the mean ± 1.96 standard deviations. (Ch. 3)
- Absence of evidence (Rom. "absența dovezii"): a nonsignificant result does not prove that the effect is absent; it may simply reflect a study that was too small. The confidence interval shows which effects remain compatible with the data. (Ch. 5)
- Absolute effect (clinical effect size; Rom. "mărime a efectului clinic", "efect absolut"): expressing the result in units that are easy to understand for clinical decision-making (mmHg, percentage points, NNT), complementary to relative measures such as the RR. (Ch. 10)
- Absolute frequency (Rom. "frecvență absolută"): the number of individuals who have a given value or belong to a given category. The chi-square test is always calculated on absolute frequencies, never on percentages. (Ch. 2)
- Absolute risk reduction (ARR; Rom. "reducerea absolută a riscului"): the difference between the risk in the control group and the risk in the treated group, for a beneficial treatment. Its reciprocal is the NNT. (Ch. 10)
- Accuracy (Acc; Rom. "acuratețe"): the proportion of people correctly classified by the test, (a + d)/N. It depends on prevalence and can be misleading when the disease is rare. (Ch. 12)
- Addition rule (Rom. "regula adunării"): $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$. The subtracted term avoids counting twice the individuals who have both characteristics. (Ch. 1)
- Adjusted R-squared (adjusted R^2; Rom. "coeficient de determinare ajustat"): a version of R^2 that penalizes the addition of useless predictors. It is used to compare models with different numbers of variables. (Ch. 9)
- Adverse event (Rom. "eveniment advers"): any unfavorable medical event occurring in a participant during the study, whether or not it is related to the intervention. Serious events are reported immediately. (Ch. 13)
- Allocation concealment (Rom. "ascunderea alocării"): the measures that prevent the person enrolling the patient from knowing in advance the arm to which the patient will be allocated (sealed opaque envelopes, central randomization). It is different from blinding. (Ch. 13)
- Alternative hypothesis (H1, Ha; Rom. "ipoteza alternativă"): the hypothesis that contradicts H0, asserting that a difference or association exists. It may be two-sided ($\mu 1 \neq \mu 2$) or one-sided ($\mu 1 > \mu 2$). (Ch. 5)
- Analysis of variance (ANOVA; Rom. "analiza variației"): a test that compares the means of three or more independent groups by partitioning the total variability into between-group and within-group variability. A significant result shows only that at least one mean differs. (Ch. 6)
- Anonymization (Rom. "anonimizare"): processing data in such a way that the person can no longer be identified by any reasonable means. Truly anonymous data are no longer personal data. (Ch. 1)

- Area under the ROC curve (AUC; Rom. "aria de sub curba ROC"): an overall indicator of the discriminating ability of a test, between 0.5 (useless test) and 1 (perfect test). It is the probability that a randomly chosen diseased person has a more "abnormal" result than a randomly chosen non-diseased person. (Ch. 12)
- Arithmetic mean (Rom. "media aritmetică"): the sum of the values divided by their number. It is the preferred measure of central tendency for symmetric distributions, but it is sensitive to extreme values. (Ch. 2)
- Attributable fraction among the exposed (Rom. "fracție atribuibilă la expuși"): the proportion of the cases occurring in the exposed that would be due to the exposure, (RR - 1)/RR. (Ch. 10)
- Attributable risk (Rom. "risc atribuibil"): the excess risk in the exposed that is due to the exposure, equal to the risk difference, assuming a causal relationship. (Ch. 10)
- Bar chart (Rom. "diagramă cu bare"): a graph for categorical variables in which the length of each bar is proportional to the frequency of the category. Unlike in a histogram, the bars are separated. (Ch. 2)
- Bayes' theorem (Rom. "teorema lui Bayes"): the relationship that allows the probability of a cause (the disease) to be calculated from the observed effect (the positive test): $P(B \mid T) = \frac{P(T|B) \times P(B)}{P(T)}$. In medicine, it converts the pre-test probability into the post-test probability. (Ch. 11)
- Bayesian statistics (Rom. "statistică bayesiană"): the approach in which prior knowledge (the prior distribution) is updated with the data, producing a posterior distribution of the parameter. It allows direct statements such as "the probability that the effect is positive is 95%". (Ch. 11)
- Between-group variability (Rom. "variabilitate între grupuri"): the part of the total variability explained by the differences between the group means (the between-group sum of squares). (Ch. 6)
- Bias (systematic error; Rom. "eroare sistematică"): a constant deviation of the results from the true value, always in the same direction, produced by flaws in selection, measurement or analysis. Unlike random error, it does not decrease as the sample grows. (Ch. 1)
- Bimodal distribution (Rom. "distribuție bimodală"): a distribution with two peaks, which often suggests a mixture of two subpopulations (for example the heights of men and women taken together). (Ch. 2)
- Binary variable (dichotomous variable; Rom. "variabilă dihotomică", "variabilă binară"): a nominal variable with exactly two categories (diseased or healthy, 0 or 1). It is the type of variable found in 2 x 2 risk and diagnostic tables. (Ch. 2)
- Binomial distribution (Rom. "distribuție binomială"): the distribution of the number of "successes" in n independent trials, each with the same probability p (for example the number of patients with diabetes in a sample of 20). For large n it is approximated by the normal distribution. (Ch. 3)
- Biological variability (Rom. "variabilitate biologică"): the fact that the same characteristic differs between individuals and, in the same individual, from one moment to another. It is the reason why medicine needs statistics. (Ch. 1)
- Biostatistics (Rom. "biostatistică"): the branch of statistics concerned with collecting, describing, analyzing and interpreting data from biology and medicine. It starts from biological variability and offers methods by which uncertainty is measured instead of being ignored. (Ch. 1)
- Blinding (masking; Rom. "orbire", "mascare"): concealing the intervention received from the participant (single blind), from the participant and the investigator (double blind), or also from the person who assesses the outcomes and analyzes the data. It reduces performance and detection bias. (Ch. 13)
- Block randomization (Rom. "randomizare în blocuri"): randomization in which allocations are made in small blocks (for example of 4 or 6), each block containing an equal number of allocations to each arm. It keeps the groups numerically balanced throughout recruitment. (Ch. 13)

- Bonferroni correction (Rom. "corecția Bonferroni"): dividing the α level by the number of comparisons performed (α/k). It is simple and conservative. (Ch. 5)
- Box plot (box-and-whisker plot; Rom. "box-plot", "diagramă cutie cu mustăți"): a graph that shows the median, the quartiles, whiskers extending to the most distant non-outlying values and, separately, the outliers. It is ideal for comparing the distributions of several groups. (Ch. 2)
- Bradford Hill criteria (Rom. "criteriile Bradford Hill"): guideposts for judging whether an association is causal: strength of the association, consistency, specificity, temporal sequence, dose-response gradient, biological plausibility, coherence, experimental evidence and analogy. Of these, only the temporal sequence is mandatory. (Ch. 10)
- Case fatality rate (Rom. "letalitate"): the proportion of patients with a given disease who die of it within a given period. It measures the severity of the disease, not its frequency in the population. (Ch. 10)
- Case report form (CRF; Rom. "formular de raportare a cazului"): the form, on paper or electronic, in which the data required by the protocol are recorded for each participant. A well-designed form prevents a large share of data entry errors. (Ch. 13)
- Case series (Rom. "serie de cazuri"): the description of a group of patients with the same disease or the same treatment, without a comparison group. It can generate hypotheses, but it cannot test them. (Ch. 10)
- Case-control study (Rom. "studiu caz-control"): a study that starts from diseased people (cases) and people without the disease (controls) and compares the frequency of past exposure. It is efficient for rare diseases, but it allows only the OR to be calculated and is prone to recall bias. (Ch. 10)
- Categorical variable (qualitative variable; Rom. "variabilă calitativă", "variabilă categorială"): a variable whose values are categories, not measurements (sex, blood group, smoking status). It is described by frequencies and percentages. (Ch. 2)
- Causality (Rom. "cauzalitate"): the relationship in which a change in the exposure produces a change in the outcome. It is demonstrated most convincingly by randomized trials; in observational studies it is argued for using criteria such as those of Bradford Hill. (Ch. 10)
- Census (Rom. "recensământ"): the study of all individuals in a population. It is exact, but costly and rarely possible in medical research, where we almost always work with samples. (Ch. 1)
- Central limit theorem (Rom. "teorema limită centrală"): the mean of sufficiently large samples has an approximately normal distribution, whatever the shape of the distribution of the variable in the population. It justifies the use of tests based on the mean in large samples. (Ch. 4)
- Chi-square goodness-of-fit test (Rom. "testul chi-pătrat de concordanță"): a test that compares the observed distribution of a single categorical variable with a theoretical distribution (for example, equal proportions across three arms). (Ch. 7)
- Chi-square test of independence (chi-square test of association; Rom. "testul chi-pătrat de asociere", "testul χ^2 de independență"): a test that checks whether two categorical variables are associated by comparing the observed frequencies in a contingency table with those expected in the absence of an association. (Ch. 7)
- Classifier (Rom. "clasificator"): a rule or model that assigns each individual to one of the classes (diseased or healthy) on the basis of certain characteristics; its performance is described by sensitivity, specificity and the ROC curve. (Ch. 12)
- Clinical significance (Rom. "semnificație clinică"): the practical importance of an effect for the patient or for the medical decision. It is judged by the effect size and the confidence interval, compared with the minimal clinically important difference. (Ch. 5)
- Clinical trial (Rom. "studiu clinic"): an experimental study in humans that evaluates the effects of an intervention (drug, device, procedure, educational program) on health. It is conducted according to a protocol approved by the ethics committee. (Ch. 13)

- Clinical trial phases (Rom. "fazele studiilor clinice"): the stages in the development of a drug: phase I (safety, dose, few volunteers), phase II (preliminary efficacy), phase III (randomized comparison in large groups), phase IV (post-marketing surveillance). (Ch. 13)
- Cluster sampling (Rom. "eșantionare pe grupuri"): whole groups (practices, schools, villages) are drawn at random, and within them all individuals or a subsample are studied. It is practical in the field, but individuals in the same group resemble one another, which reduces precision. (Ch. 1)
- Coding (Rom. "codificare"): converting responses into standardized values (for example 0/1 for absent/present, 1-4 for level of education), described in the data dictionary. (Ch. 2)
- Coefficient of determination (R^2; Rom. "coeficient de determinare"): the proportion of the variability of the dependent variable explained by the model; in simple regression it equals r^2. It says nothing about the direction of the relationship. (Ch. 9)
- Coefficient of variation (CV; Rom. "coeficient de variație"): the standard deviation expressed as a percentage of the mean (100 x s/m). It allows the variability of measurements with different units or orders of magnitude to be compared. (Ch. 2)
- Cohen's d (Rom. "Cohen d"): the difference between two means expressed in standard deviations. Rough guide values: 0.2 small effect, 0.5 medium, 0.8 large. (Ch. 6)
- Cohen's kappa (κ; Rom. "coeficientul kappa al lui Cohen"): a measure of agreement between two raters for a categorical variable, corrected for the agreement expected by chance. Values above 0.60 indicate substantial agreement, and those above 0.80 almost perfect agreement. (Ch. 12)
- Cohort study (Rom. "studiu de cohortă"): a study in which a group of people without the disease, classified by exposure, is followed over time for the occurrence of the disease. It allows incidence and the RR to be calculated; it may be prospective or retrospective (historical). (Ch. 10)
- Complementary event (Rom. "eveniment contrar"): the event "not A", which occurs exactly when A does not occur; $P(\text{non} - A) = 1 - P(A)$. (Ch. 1)
- Conditional probability (Rom. "probabilitate condiționată"): the probability of an event given that another event has occurred, written $P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$. Sensitivity, specificity and the predictive values are conditional probabilities. (Ch. 1)
- Confidence interval (CI; Rom. "interval de încredere"): an interval calculated from the sample which, if the study were repeated many times, would contain the true value of the parameter in a given percentage of cases (usually 95%). Its width shows the precision of the estimate. (Ch. 4)
- Confidence interval for a difference (Rom. "interval de încredere pentru o diferență"): a CI calculated for the difference between two means or two proportions. If it does not contain 0, the difference is significant at the corresponding level; for ratios (RR, OR), the reference value is 1. (Ch. 4)
- Confidence level (Rom. "nivel de încredere"): the percentage (1 - α) x 100 associated with a confidence interval, usually 95% or 99%. A higher level produces a wider interval. (Ch. 4)
- Confounder (Rom. "factor de confuzie"): a variable associated with both the exposure and the outcome, without being an intermediate link between them, which can create or mask an association. It is controlled by randomization, restriction, matching, stratification or adjustment in multivariable models. (Ch. 10)
- CONSORT guideline (Consolidated Standards of Reporting Trials; Rom. "ghidul CONSORT"): a guideline for reporting randomized controlled trials, with a checklist and a flow diagram of participants (assessed, randomized, allocated, followed up, analyzed). (Ch. 13)
- Contingency table (Rom. "tabel de contingență"): a two-way table that cross-classifies the categories of two variables and gives the frequency of each combination. The 2 x 2 table is the simplest case. (Ch. 7)

- Continuous variable (Rom. "variabilă continuă"): a numerical variable that can, in theory, take any value within an interval (blood glucose, height); its precision is limited only by the measuring instrument. (Ch. 2)
- Control group (Rom. "grup de control"): the group that receives the comparison intervention (placebo, standard treatment or no intervention), against which the effect of the intervention under study is measured. (Ch. 13)
- Convenience sampling (Rom. "eșantionare de conveniență"): including patients who are easily available (those who come to the practice during a certain period, the respondents to an online questionnaire). It is common, but it does not guarantee representativeness. (Ch. 1)
- Correlation (Rom. "corelație"): the association between two variables, characterized by its direction (positive or negative) and its strength. Correlation does not prove a cause-and-effect relationship. (Ch. 9)
- Critical appraisal (Rom. "interpretarea critică a unui articol"): the systematic assessment of the validity, results and applicability of a published study: a design suited to the question, control of bias, appropriateness of the tests, reporting of CIs and effect sizes. (Ch. 14)
- Critical value (Rom. "valoare critică"): the value of the test statistic that separates the non-rejection region from the rejection region of H0, for a given α. It is read from statistical tables (Appendix E). (Ch. 5)
- Cross-sectional study (Rom. "studiu transversal"): a study in which exposure and disease are measured at the same point in time, like a snapshot of the population. It estimates prevalence, but it cannot establish the temporal sequence of exposure and disease. (Ch. 10)
- Crossover trial (Rom. "studiu încrucișat"): a trial in which each participant receives both interventions in succession, in random order, separated by a washout period. Each patient is his or her own control, which calls for tests for paired data. (Ch. 13)
- Cumulative frequency (Rom. "frecvență cumulată"): the sum of the frequencies of all categories up to and including the current category. It makes sense only for ordered variables (ordinal, or numerical grouped into classes). (Ch. 2)
- Cumulative incidence (risk; Rom. "incidență cumulativă"): the proportion of initially healthy people who develop the disease within a given period. It is always stated together with the period (for example 3% over 5 years). (Ch. 10)
- Cut-off value (Rom. "valoare prag"): the value of a numerical test (score, concentration) above which the result is considered positive. Moving the cut-off increases sensitivity at the expense of specificity, or vice versa. (Ch. 12)
- Data cleaning (Rom. "curățarea datelor"): the systematic checking of the file before analysis: inconsistently coded categories, impossible values, text in numerical columns, missing values, duplicates, inconsistencies between variables. Each correction is made against the source document and recorded in a log. (Ch. 2)
- Data dictionary (Rom. "dicționar de date"): a document that describes each variable in a file: its name, meaning, type, scale, unit of measurement or coding, valid range and the way missing values are recorded. (Ch. 13)
- Data monitoring committee (Rom. "comitet independent de monitorizare a datelor"): a group of experts, independent of the sponsor and the investigators, who follow the safety and efficacy data during the trial and recommend whether it should continue, be modified or be stopped. (Ch. 13)
- Data validation rules (Rom. "validarea datelor la introducere"): rules built into the electronic form (permitted ranges, mandatory fields, option lists) that prevent impossible or inconsistently coded values from being entered. (Ch. 13)
- Declaration of Helsinki (Rom. "Declarația de la Helsinki"): a set of ethical principles for medical research involving human subjects, adopted by the World Medical Association and revised periodically (most recently in 2024). (Ch. 13)

- Degrees of freedom (df; Rom. "grade de libertate"): the number of values that can vary freely once certain constraints (for example the mean) have been fixed. They determine the shape of the t, chi-square and F distributions; for the one-sample t-test, df = n - 1. (Ch. 4)
- Dependent variable (outcome variable; Rom. "variabilă dependentă", "variabilă rezultat"): the variable whose behavior we want to explain or predict (for example SBP or the presence of diabetes). (Ch. 8)
- Derived variable (Rom. "variabilă derivată"): a variable calculated from other variables (BMI from weight and height, weight loss from two measurements). It is recalculated, not entered by hand, and comparing it with the recorded value helps to uncover errors. (Ch. 2)
- Descriptive statistics (Rom. "statistică descriptivă"): the part of statistics that summarizes the data of a sample by frequency tables, summary measures (mean, median, standard deviation) and graphs, without making statements about the population. It is the first step of any analysis. (Ch. 2)
- Diagnostic odds ratio (DOR; Rom. "raportul cotelor de diagnostic"): the ratio between the odds of a positive test in diseased people and the odds of a positive test in non-diseased people, equal to ad/bc or LR+/LR-. It summarizes the performance of the test in a single number. (Ch. 12)
- Diagnostic test (Rom. "test diagnostic"): any procedure (laboratory test, imaging, clinical score, question) used to establish whether or not a patient has a given disease. (Ch. 12)
- Dichotomization (Rom. "dihotomizare"): converting a numerical variable into one with two categories by means of a cut-off (for example obese if BMI $\geq$ 30). It simplifies interpretation, but it loses information and reduces statistical power. (Ch. 2)
- Discrete variable (Rom. "variabilă discretă"): a numerical variable that takes only separate values, usually whole numbers resulting from counting (number of consultations, number of pregnancies). (Ch. 2)
- Dummy variable (indicator variable; Rom. "variabilă indicator", "variabilă dummy"): a 0/1 variable created in order to enter a category into a regression model; a variable with k categories requires k - 1 indicators, relative to a reference category. (Ch. 9)
- Dunn test (Rom. "testul Dunn"): a post hoc test for pairwise comparisons after a significant Kruskal-Wallis test, with p values adjusted for multiple comparisons. (Ch. 7)
- Ecological study (Rom. "studiu ecologic"): a study in which the unit of analysis is a group (country, county), not the individual. Associations observed at group level cannot automatically be transferred to individuals (the ecological fallacy). (Ch. 10)
- Effect modification (interaction; Rom. "modificare de efect", "interacțiune"): the situation in which the size of the association between exposure and disease differs according to the level of a third variable (for example, the effect of a treatment differs between men and women). (Ch. 10)
- Effect size (Rom. "mărimea efectului"): a measure of the magnitude of a difference or association, independent of the sample size (Cohen's d, eta squared, r, OR). It should be reported alongside the p value. (Ch. 5)
- Eligibility criteria (Rom. "criterii de includere și de excludere"): the conditions that a person must meet (or must not meet) in order to take part in the study. They define the population to which the results can be applied. (Ch. 13)
- Epi Info (Rom. "Epi Info"): a free program developed by the US Centers for Disease Control and Prevention, used in epidemiology for forms, 2 x 2 tables, RR, OR and the chi-square test. (Ch. 10)
- Epidemiology (Rom. "epidemiologie"): the science that studies the frequency, distribution and determinants of diseases in populations and applies this knowledge to disease prevention and control. (Ch. 10)
- Equivalence trial (Rom. "studiu de echivalență"): a trial that aims to show that two interventions do not differ, in either direction, by more than a prespecified margin. (Ch. 13)

- Eta squared (η^2; Rom. "eta pătrat"): the proportion of the total variability explained by the grouping factor in ANOVA (SSB/SST). Rough guide values: 0.01 small, 0.06 medium, 0.14 large. (Ch. 6)
- Event (Rom. "eveniment"): a possible result of an observation or experiment (the patient has diabetes, the test is positive). Events are combined by "and" (intersection) and "or" (union). (Ch. 1)
- Evidence-based medicine (EBM; Rom. "medicina bazată pe dovezi"): the conscientious and explicit use of the best available research evidence, integrated with clinical expertise and with patient values, in decisions about the care of individual patients. (Ch. 11)
- Expected frequency (Rom. "frecvență așteptată"): the number of observations that would fall in a cell if the variables were independent, calculated as row total x column total / N. The usual condition of the chi-square test is that the expected frequencies be at least 5. (Ch. 7)
- Experimental study (Rom. "studiu experimental"): a study in which the researcher assigns the intervention to the participants, ideally by randomization. The randomized controlled trial is the reference form. (Ch. 13)
- Explanatory trial (Rom. "studiu explicativ"): a clinical trial under ideal, strictly controlled conditions, which answers the question "can the intervention work?" (efficacy, not effectiveness). (Ch. 13)
- Exposure (Rom. "expunere"): the factor whose effect on the disease is being studied (a behavior, a treatment, an environment, a genetic characteristic). It may be a single event or chronic, in which case dose and duration matter. (Ch. 10)
- Extrapolation (Rom. "extrapolare"): using a model to predict values outside the range of predictor values observed in the data. It is risky, because the relationship may no longer be linear there. (Ch. 9)
- F distribution (Fisher distribution; Rom. "distribuția Fisher", "distribuția F"): the distribution of the ratio of two variances, defined by two numbers of degrees of freedom (numerator and denominator). It is used in ANOVA, in regression and in tests that compare variances. (Ch. 6)
- F statistic (Rom. "statistica F"): the ratio between the between-group mean square and the within-group mean square. Values well above 1 show that the means differ more than random variability would explain. (Ch. 6)
- Fagan nomogram (Rom. "nomograma Fagan"): a chart with three axes (pre-test probability, likelihood ratio, post-test probability) on which a straight line performs the Bayesian conversion without any calculation. (Ch. 11)
- False negative (FN; Rom. "fals negativ"): a diseased person with a negative test (cell b). The consequence is delayed diagnosis and treatment. (Ch. 12)
- False positive (FP; Rom. "fals pozitiv"): a person without the disease who has a positive test (cell c). The consequences are anxiety and unnecessary investigations. (Ch. 12)
- Fisher's exact test (Rom. "testul exact Fisher"): a test for 2 x 2 (and larger) tables that calculates exactly the probability of the observed configuration and of more extreme ones. It is used when the expected frequencies are small. (Ch. 7)
- Flow diagram (Rom. "diagramă de flux"): the chart that shows the number of participants at each stage of the study and the reasons for exclusions and losses. (Ch. 13)
- Forest plot (Rom. "grafic forest"): a graphical display of the estimates (RR, OR, differences in means) and their confidence intervals for several studies or subgroups, with a vertical line at the "no effect" value. (Ch. 10)
- Frequency table (Rom. "tabel de frecvențe"): a table that shows, for each value or category, the absolute frequency, the percentage, the valid percentage and, for ordered variables, the cumulative percentage. It is the simplest tool for detecting coding errors. (Ch. 2)

- Frequentist statistics (Rom. "statistică frecventistă"): the approach in which probability is the long-run frequency of an event in hypothetical repetitions of the study; the p values and confidence intervals in this course are based on it. (Ch. 11)
- Friedman test (Rom. "testul Friedman"): the nonparametric alternative to repeated measures ANOVA, for three or more measurements on the same individuals. Ranks are assigned within each individual. (Ch. 7)
- General Data Protection Regulation (GDPR; Rom. "Regulamentul general privind protecția datelor"): Regulation (EU) 2016/679, which sets the conditions under which personal data may be processed, including medical data used in research. (Ch. 1)
- Geometric mean (Rom. "media geometrică"): the n-th root of the product of the n values, equal to the exponential of the mean of the logarithms. It is suitable for positive variables skewed to the right, such as triglycerides or antibody titers. (Ch. 2)
- Good Clinical Practice (GCP; Rom. "buna practică în studiul clinic"): an international standard of ethics and scientific quality for the design, conduct, recording and reporting of clinical trials, described in the ICH E6 guideline. (Ch. 13)
- GRADE approach (Rom. "sistemul GRADE"): a method for rating the certainty of the evidence (high, moderate, low, very low) and the strength of guideline recommendations. (Ch. 11)
- Heterogeneity (Rom. "heterogenitate"): variation in the results of the studies included in a meta-analysis that is greater than chance would explain; it is usually quantified by the I^2 index. (Ch. 11)
- Hierarchy of evidence (Rom. "ierarhia dovezilor"): the ranking of study types according to how well they control systematic error, from expert opinion and case series, through observational studies, up to randomized trials and systematic reviews with meta-analysis. (Ch. 11)
- Histogram (Rom. "histogramă"): a graph made of adjacent rectangles that shows the distribution of a numerical variable grouped into classes. Its shape suggests symmetry, skewness or the presence of several peaks. (Ch. 2)
- Homogeneity of variance (homoscedasticity; Rom. "omogenitatea variațelor", "homoscedasticitate"): the condition that variability be approximately the same in the groups being compared (or, in regression, along the line). It is a condition of the classical t-test, of ANOVA and of linear regression. (Ch. 6)
- Incidence rate (incidence density; Rom. "densitate de incidență", "rata de incidență"): the number of new cases divided by the total follow-up time of the people at risk (person-years). It takes into account the fact that participants are followed for different periods. (Ch. 10)
- Independent events (Rom. "evenimente independente"): two events for which the occurrence of one does not change the probability of the other; then $P(A \text{ and } B) = P(A) \times P(B)$. (Ch. 1)
- Independent samples (Rom. "eșantioane independente"): groups made up of different individuals, in which the values in one group are not linked to the values in the other. (Ch. 6)
- Independent variable (predictor; Rom. "variabilă independentă"): the variable used to explain or predict the outcome variable (for example BMI, age, treatment arm). (Ch. 8)
- Independent-samples t-test (Rom. "testul t pentru eșantioane independente"): it compares the means of two groups made up of different individuals. The classical version (Student) assumes equal variances; the Welch version does not. (Ch. 6)
- Inferential statistics (Rom. "statistică inferențială"): the part of statistics that generalizes from the sample to the population, by estimating parameters with confidence intervals and by testing hypotheses. Its conclusions are probabilistic, not certain. (Ch. 1)
- Influential point (Rom. "valoare influentă"): an observation whose removal changes the correlation coefficient or the regression line substantially, usually a point that is extreme on the x axis. (Ch. 9)

- Information bias (Rom. "bias de informație"): a systematic error in measuring the exposure or the outcome, for example through uncalibrated instruments or questions worded differently for cases and controls. (Ch. 10)
- Informed consent (Rom. "consimțământ informat"): a person's free agreement, given in writing, to take part in a study after receiving clear information about its purpose, procedures, risks, benefits and the right to withdraw at any time. (Ch. 1)
- Integrated data analysis (Rom. "analiza integrată a datelor"): working through a dataset from start to finish, from cleaning and descriptive statistics to the choice of tests, interpretation and reporting, in a coherent report that answers the research question. (Ch. 13)
- Intention-to-treat analysis (ITT; Rom. "analiza după intenția de tratament"): the analysis in which each participant is included in the arm to which he or she was randomized, regardless of whether the intervention was actually received. It preserves the advantages of randomization and is the primary analysis in superiority trials. (Ch. 13)
- Intercept (a; Rom. "ordonata la origine"): the predicted value of y when x = 0. It often has no clinical interpretation, because x = 0 lies outside the data (a BMI of 0 does not exist). (Ch. 9)
- Interim analysis (Rom. "analiză interimară"): an analysis performed before the end of the study, according to rules set out in the protocol, which may lead to early stopping for benefit, lack of effect or harm. (Ch. 13)
- Interquartile range (IQR; Rom. "interval intercuartilic"): the difference Q3 - Q1, which contains the central 50% of the values. It accompanies the median in describing skewed distributions and is the basis of the rule for detecting outliers on a box plot. (Ch. 2)
- Interval estimation (Rom. "estimare prin interval"): giving a range of plausible values for the parameter, together with a confidence level. The usual form is the confidence interval. (Ch. 4)
- Interval scale (Rom. "scală de interval"): a numerical scale with equal distances between values but no absolute zero (temperature in degrees Celsius). Differences are meaningful, ratios are not. (Ch. 2)
- JASP (Rom. "JASP"): a free statistical program with a graphical interface, used in the practical sessions for descriptive statistics, t-tests, ANOVA, nonparametric tests, correlation and regression. (Ch. 2)
- Kendall's tau (Rom. "coeficientul tau Kendall"): a nonparametric correlation coefficient based on the number of concordant and discordant pairs. Some authors prefer it for small samples and when there are many tied values. (Ch. 9)
- Kolmogorov-Smirnov test (KS; Rom. "testul Kolmogorov-Smirnov"): a test that compares the observed distribution with a theoretical distribution, also used for normality. For checking normality it is less powerful than the Shapiro-Wilk test. (Ch. 3)
- Kruskal-Wallis test (Rom. "testul Kruskal-Wallis"): the nonparametric alternative to one-way ANOVA, for three or more independent groups. A significant result is followed by post hoc comparisons (for example the Dunn test). (Ch. 7)
- Kurtosis (excess kurtosis; Rom. "exces", "boltire"): a measure of the "heaviness" of the tails of a distribution compared with the normal distribution. Positive excess kurtosis indicates more extreme values than would be expected in a normal distribution. (Ch. 3)
- Law of large numbers (Rom. "legea numerelor mari"): as the number of observations increases, the relative frequency of an event approaches its probability, and the sample mean approaches the population mean. (Ch. 1)
- Least squares method (Rom. "metoda celor mai mici pătrate"): the method by which the coefficients of the regression line are chosen so that the sum of squared residuals is as small as possible. (Ch. 9)

- Levene test (Rom. "testul Levene"): a test of the equality of variances in two or more groups, with H0 "the variances are equal". A $p < 0.05$ points toward the Welch test or Welch ANOVA. (Ch. 6)
- Likelihood ratio (LR; Rom. "raport de verosimilitate"): how many times more likely a given test result is in a diseased person than in a non-diseased person. $LR^{+} = \frac{Se}{1-Sp}$, $LR^{-} = \frac{1-Se}{Sp}$; an LR+ above 10 or an LR- below 0.1 changes the probability of disease considerably. (Ch. 11)
- Likert scale (Rom. "scală Likert"): an ordered response scale, usually with 5 or 7 points (from "strongly disagree" to "strongly agree"), used in questionnaires. The data obtained are ordinal and are preferably analyzed with nonparametric tests. (Ch. 7)
- Linearity (Rom. "liniaritate"): the condition that the relationship between two variables can be described by a straight line. It is checked on the scatter plot and on the residual plot. (Ch. 9)
- Log transformation (Rom. "transformare logaritmică"): replacing each value with its logarithm, in order to reduce positive skewness and bring the distribution closer to the normal one. The results are brought back to the original scale by exponentiation, which gives the geometric mean. (Ch. 3)
- Logistic regression (Rom. "regresie logistică"): a model for a dichotomous dependent variable (diseased or healthy) whose coefficients, once exponentiated, give odds ratios adjusted for the other predictors. (Ch. 9)
- Lognormal distribution (Rom. "distribuție lognormală"): the distribution of a variable whose logarithm is normally distributed. It is skewed to the right and describes many biological concentrations well. (Ch. 3)
- Loss to follow-up (Rom. "pierdere din urmărire"): the situation in which a participant's outcome can no longer be measured (the person has moved, no longer responds). Large losses, or losses that are unequal between arms, can introduce bias. (Ch. 13)
- Mann-Whitney U test (Wilcoxon rank-sum test; Rom. "testul Mann-Whitney U", "testul Wilcoxon al sumei rangurilor"): the nonparametric alternative to the independent-samples t-test; it compares the distributions of two groups on the basis of ranks. The result is reported with medians and interquartile ranges. (Ch. 7)
- Margin of error (Rom. "marja de eroare"): half the width of the confidence interval, that is, the distance from the point estimate to each limit. It is the quantity that is fixed when the sample size is calculated. (Ch. 4)
- McNemar test (Rom. "testul McNemar"): a test for two dichotomous variables measured on the same individuals (for example, the results of two diagnostic tests in the same patients); it uses only the discordant pairs. (Ch. 7)
- Mean absolute deviation (Rom. "abatere medie absolută"): the mean of the absolute values of the deviations from the mean. It is intuitive, but it is used less often than the standard deviation. (Ch. 2)
- MedCalc (Rom. "MedCalc"): a statistical program geared toward medical research, used in the course mainly for ROC curves, 2 x 2 tables, relative risks and sample size calculation. (Ch. 12)
- Median (Rom. "mediana"): the value that divides the ordered series into two halves with equal numbers of observations. It is not influenced by extreme values and is preferred for skewed distributions and ordinal variables. (Ch. 2)
- Meta-analysis (Rom. "meta-analiză"): the statistical combination of the results of several independent studies that address the same question, in order to obtain a more precise overall estimate of the effect. (Ch. 11)
- Method agreement (Rom. "acord între metode"): the extent to which two measurement methods give the same values in the same patients. For numerical variables it is assessed with the Bland-Altman method, not with the correlation coefficient. (Ch. 12)

- Minimal clinically important difference (MCID; Rom. "diferența minimă importantă clinic"): the smallest difference in an outcome that patients or physicians consider relevant. It is set when the study is planned and is used to calculate the sample size. (Ch. 5)
- Misclassification rate (Rom. "rata erorii de clasificare"): the proportion of people incorrectly classified, $\frac{b+c}{N} = \frac{\text{FN+FP}}{N}$, that is, 1 - accuracy. (Ch. 12)
- Missing value (Rom. "valoare lipsă"): an observation that should have been recorded but is missing. Missing values are reported for each variable and are not filled in at will, only by an imputation method specified in the analysis plan. (Ch. 2)
- Mode (Rom. "moda"): the most frequent value or category. It is the only measure of central tendency applicable to nominal variables. (Ch. 2)
- Morbidity (Rom. "morbiditate"): the frequency of diseases in a population, expressed as prevalence or incidence. (Ch. 10)
- Mortality (Rom. "mortalitate"): the frequency of deaths in a population over a period, relative to the whole population (for example, deaths per 100,000 inhabitants per year). (Ch. 10)
- Multicollinearity (Rom. "multicoliniaritate"): a strong correlation between the predictors of a regression model, which makes the coefficients unstable and hard to interpret (for example weight and BMI in the same model). (Ch. 9)
- Multiple comparisons (Rom. "comparații multiple"): performing several tests on the same data, which increases the probability of obtaining at least one false positive result. This is controlled by corrections (Bonferroni) or by post hoc tests. (Ch. 5)
- Multiple linear regression (Rom. "regresie liniară multiplă"): a model with one numerical dependent variable and several predictors; each coefficient shows the effect of the corresponding predictor at constant values of the others (adjustment). (Ch. 9)
- Multiplication rule (Rom. "regula înmulțirii"): $P(A \text{ and } B) = P(B) \times P(A \mid B)$; for independent events it becomes P(A) x P(B). (Ch. 1)
- Mutually exclusive events (Rom. "evenimente incompatibile"): events that cannot occur at the same time (a patient cannot be both a smoker and a nonsmoker). For such events, $P(A \text{ or } B) = P(A) + P(B)$. (Ch. 1)
- Negative predictive value (NPV; Rom. "valoarea predictivă negativă"): the proportion of people without the disease among those with a negative test, d/(b + d). It increases as prevalence decreases. (Ch. 12)
- Nominal variable (Rom. "variabilă nominală"): a categorical variable with no natural order among its categories (urban or rural residence, treatment arm). The numerical codes assigned to the categories are mere labels. (Ch. 2)
- Non-inferiority trial (Rom. "studiu de non-inferioritate"): a trial that aims to show that a new intervention is not worse than the standard treatment by more than a prespecified margin, usually because it has other advantages (cost, safety, administration). (Ch. 13)
- Nonparametric test (Rom. "test neparametric"): a test that does not assume a particular distribution of the data (usually the normal one) and often works with the ranks of the values. It is used for ordinal variables, for skewed distributions with small samples and in the presence of outliers. (Ch. 7)
- Normal distribution (Gaussian distribution; Rom. "distribuție normală", "distribuția Gauss"): a continuous, symmetric, bell-shaped distribution, completely defined by its mean and standard deviation. The mean, median and mode coincide, and the tails theoretically extend to infinity. (Ch. 3)
- Normality (Rom. "normalitate"): the property of data of approximately following a normal distribution. It is a condition of parametric tests and is checked graphically (histogram, Q-Q plot) and by tests (Shapiro-Wilk). (Ch. 3)

- Null hypothesis (H0; Rom. "ipoteza nulă"): the statistical hypothesis of no difference or no association ($\mu1 = \mu2$, RR = 1, $\rho = 0$). The test calculates how compatible the data are with it; H0 is rejected or not rejected, but it is not "accepted". (Ch. 5)
- Number needed to harm (NNH; Rom. "numărul necesar pentru a produce un efect advers"): the number of patients who must be treated for one additional patient to experience an adverse effect, equal to the reciprocal of the absolute increase in the risk of the adverse effect. (Ch. 10)
- Number needed to treat (NNT; Rom. "numărul necesar de tratat"): the number of patients who must be treated for one additional patient to benefit, equal to 1/ARR and rounded up. A small NNT indicates an effective treatment. (Ch. 10)
- Numerical variable (quantitative variable; Rom. "variabilă numerică", "variabilă cantitativă"): a variable expressed as a number with a unit of measurement, for which arithmetic operations are meaningful (SBP in mmHg, weight in kg). (Ch. 2)
- Observational study (Rom. "studiu observațional"): a study in which the researcher only observes the exposure and the outcome without influencing them (cross-sectional, cohort, case-control). It is exposed to confounding factors. (Ch. 10)
- Odds (Rom. "cotă"): the ratio between the probability that an event occurs and the probability that it does not occur, p/(1 - p). A probability of 0.20 corresponds to odds of 0.25 (1 to 4). (Ch. 10)
- Odds ratio (OR; Rom. "raportul cotelor"): the ratio between the odds of disease in the exposed and the odds of disease in the unexposed, ad/bc in a 2 x 2 table. It is the only valid measure of association in case-control studies and approximates the RR when the disease is rare. (Ch. 10)
- One-sample t-test (Rom. "testul t pentru un eșantion"): it compares the mean of a sample with a fixed reference value (for example a published norm) when σ is unknown. It has n - 1 degrees of freedom. (Ch. 6)
- One-sided test (Rom. "test unilateral"): a test in which the alternative hypothesis allows only one direction ($\mu1 > \mu2$). It is justified only if a difference in the opposite direction is impossible or of no interest and if the direction was specified before the analysis. (Ch. 5)
- Ordinal variable (Rom. "variabilă ordinală"): a categorical variable whose categories have a natural order, although the distances between them are not necessarily equal (stage of a disease, level of education, satisfaction score 1-5). It is described by the median and quartiles. (Ch. 2)
- Outlier (Rom. "valoare aberantă"): an observation that deviates markedly from the rest of the data. It may be a data entry error (SBP 1540 mmHg) or a genuine extreme value; in the first case it is corrected against the source, in the second it is kept and its influence is analyzed. (Ch. 2)
- P value (Rom. "valoarea p"): the probability of obtaining a result at least as extreme as the one observed if H0 were true. It is not the probability that H0 is true, and it does not measure the size or the clinical importance of the effect. (Ch. 5)
- Paired data (Rom. "date pereche", "date legate"): observations that form natural pairs: the same individual measured twice, the two eyes of the same patient, twins or patients matched in pairs. They are analyzed with tests for dependent data. (Ch. 6)
- Paired t-test (Rom. "testul t pereche"): it compares the means of two linked series of measurements (the same patient before and after treatment) by analyzing the mean of the individual differences. The main condition is normality of the differences. (Ch. 6)
- Parallel testing (Rom. "teste în paralel"): applying two tests at the same time; the patient is considered positive if either test is positive. It increases sensitivity and decreases specificity. (Ch. 12)
- Parameter (Rom. "parametru"): a numerical value that describes the population (the mean μ, the standard deviation σ, the proportion π). As a rule it is unknown and is estimated from the sample. (Ch. 4)

- Parametric test (Rom. "test parametric"): a test that assumes a particular distribution of the data, usually the normal one, and compares parameters such as the mean (t-tests, ANOVA, Pearson correlation). When its conditions are met, it is more powerful than its nonparametric equivalent. (Ch. 8)
- Pearson correlation coefficient (r; Rom. "coeficientul de corelație Pearson"): a measure of the strength and direction of the linear relationship between two numerical variables, with values between -1 and 1. It assumes linearity, approximately normal distributions and no influential extreme values. (Ch. 9)
- Per-protocol analysis (PP; Rom. "analiza per protocol"): the analysis restricted to the participants who adhered to the protocol. It may overestimate the effect, because those who drop out often differ from those who stay. (Ch. 13)
- Percentile (Rom. "percentilă"): the value below which a given percentage of the observations lies; the 90th percentile has 90% of the values below it. Pediatric growth charts are built on percentiles. (Ch. 2)
- Person-time (Rom. "persoane-timp"): the sum of the follow-up times of all participants until disease, loss to follow-up or the end of the study. It is the denominator of the incidence rate. (Ch. 10)
- Personal data (Rom. "date cu caracter personal"): any information relating to an identified or identifiable person. Health data belong to the special categories, which enjoy enhanced protection under the General Data Protection Regulation. (Ch. 1)
- PICO question (Rom. "întrebarea PICO"): the structured formulation of a clinical question: the patient or problem (P), the intervention or exposure (I), the comparison (C) and the outcome (O). It makes searching the literature easier. (Ch. 11)
- Pie chart (Rom. "diagramă circulară"): a graph in which a circle is divided into sectors proportional to the frequencies of the categories. It is readable only for a few categories and is not used for numerical variables. (Ch. 2)
- Pilot study (Rom. "studiu pilot"): a small study that tests the feasibility of the procedures (recruitment, forms, adherence) before the main study; it is not designed to test efficacy. (Ch. 13)
- Placebo (Rom. "placebo"): an inactive intervention, identical in appearance to the active treatment, used in the control group to allow blinding and to separate the specific effect from the placebo effect. (Ch. 13)
- Point estimate (Rom. "estimare punctuală"): a single value calculated from the sample and used as the best approximation of the parameter (for example the mean SBP of 131.13 mmHg). (Ch. 4)
- Poisson distribution (Rom. "distribuție Poisson"): the distribution of the number of rare events occurring in an interval of time or space (the number of new cases per month in a hospital). Its mean and variance are equal. (Ch. 3)
- Population (Rom. "populație"): the set of all individuals who share at least one characteristic and about whom we want to draw conclusions, delimited in space and time. Example: all adults at cardiometabolic risk registered with family physicians in a region. (Ch. 1)
- Population attributable fraction (PAF; Rom. "fracție atribuibilă în populație"): the proportion of all cases in the population that would disappear if the exposure were eliminated, assuming a causal relationship. It depends both on the RR and on how common the exposure is. (Ch. 10)
- Positive predictive value (PPV; Rom. "valoarea predictivă pozitivă"): the proportion of diseased people among those with a positive test, a/(a + c). It depends strongly on prevalence: the same test has a lower PPV in a population in which the disease is rare. (Ch. 12)
- Post hoc tests (Rom. "teste post-hoc"): pairwise comparisons performed after a significant ANOVA, with control of the type I error (Tukey, Bonferroni, Games-Howell). They show which groups differ from one another. (Ch. 6)

- Post-test probability (posterior probability; Rom. "probabilitate post-test", "probabilitate a posteriori"): the probability of the disease once the test result is known. For a positive test it equals the PPV corresponding to the pre-test probability. (Ch. 11)
- Pragmatic trial (Rom. "studiu pragmatic"): a clinical trial conducted under conditions close to routine practice, with broad inclusion criteria, which answers the question "does the intervention work in real life?". (Ch. 13)
- Pre-test odds and post-test odds (Rom. "cota pre-test și cota post-test"): the pre-test and post-test probabilities expressed as odds; post-test odds = pre-test odds x likelihood ratio. (Ch. 11)
- Pre-test probability (prior probability; Rom. "probabilitate pre-test", "probabilitate a priori"): the probability that the patient has the disease before the test result is known, estimated from the prevalence and from the patient's clinical data. (Ch. 11)
- Precision (Rom. "precizie"): the degree of closeness between the results of repeated measurements or estimates; in estimation, it is reflected by the width of the confidence interval. An estimate can be precise but biased if there is bias. (Ch. 4)
- Prediction (Rom. "predicție"): estimating the value of the dependent variable for a new individual on the basis of the model. The prediction interval for an individual is much wider than the confidence interval for the mean of y. (Ch. 9)
- Predictive values and prevalence (Rom. "valoare predictivă și prevalență"): the relationship whereby, at constant sensitivity and specificity, the PPV increases and the NPV decreases as the prevalence of the disease increases. It explains why a test that performs well in hospital gives many false positive results in screening of the general population. (Ch. 12)
- Prevalence (Rom. "prevalență"): the proportion of people in a population who have the disease at a given point in time (point prevalence) or during a period (period prevalence). It includes both new and existing cases. (Ch. 10)
- Prevalence ratio (PR; Rom. "raportul prevalențelor"): the analogue of relative risk in cross-sectional studies, calculated as the ratio between the prevalence of the disease in the exposed and in the unexposed. (Ch. 10)
- Primary outcome, primary endpoint (Rom. "criteriul principal de evaluare"): the main outcome on which the efficacy of the intervention is judged and on which the sample size is calculated. It is specified in the protocol before data collection. (Ch. 13)
- Probability (Rom. "probabilitate"): a number between 0 and 1 that expresses how likely an event is; 0 means impossible, 1 certain. It can be estimated by the relative frequency of the event in a large number of observations. (Ch. 1)
- Probability distribution (Rom. "distribuție de probabilitate"): the description of all possible values of a random variable and of the probabilities with which they occur. For continuous variables it is described by a density curve, under which the total area is 1. (Ch. 3)
- Protective factor (Rom. "factor protector"): an exposure associated with a lower probability of disease, with an RR or OR below 1 (physical activity for diabetes). (Ch. 10)
- Pseudonymization (Rom. "pseudonimizare"): replacing direct identifiers (name, CNP, the Romanian personal numeric code) with a code, while keeping the linking key separately. Pseudonymized data remain personal data and must be protected as such. (Ch. 1)
- Q-Q plot (quantile-quantile plot; Rom. "grafic Q-Q"): a graph that plots the observed quantiles against the theoretical quantiles of the normal distribution. If the data are normal, the points lie approximately on a straight line. (Ch. 3)
- Quartiles (Rom. "cuartile"): the three values (Q1, Q2 = the median, Q3) that divide the ordered series into four equal parts. The middle half of the observations lies between Q1 and Q3. (Ch. 2)

- Random error (Rom. "eroare aleatoare"): the chance variation of results from one sample to another, with no preferred direction. It decreases as the sample size increases and is quantified by the standard error and the confidence interval. (Ch. 4)
- Random variable (Rom. "variabilă aleatoare"): a variable whose value depends on the outcome of a random process, for example the blood glucose value of a randomly selected patient. (Ch. 3)
- Randomization (Rom. "randomizare"): allocating participants to the study arms by a random procedure, so that each has a known (usually equal) chance of receiving each intervention. It is the main protection against confounding factors. (Ch. 13)
- Randomized controlled trial (RCT; Rom. "studiu clinic randomizat"): a study in which participants are allocated at random to different intervention groups, which makes the groups comparable for both known and unknown factors. (Ch. 13)
- Range (Rom. "amplitudine"): the difference between the maximum and the minimum value. It is easy to calculate, but it depends on only two values, so it is very sensitive to extreme values. (Ch. 2)
- Rank (Rom. "rang"): the position of a value in the series ordered in ascending order. Tied values receive the mean rank of the positions they occupy. (Ch. 7)
- Rank-biserial correlation (Rom. "corelația biserială a rangurilor"): an effect size for the Mann-Whitney test, between -1 and 1; it shows how often the values in one group exceed those in the other. (Ch. 7)
- Rate (Rom. "rata"): a ratio that includes time in the denominator (cases per person-years). In everyday language, the term is often used loosely for proportions. (Ch. 10)
- Rate standardization (Rom. "standardizarea ratelor"): a method for comparing rates between populations with different age structures by relating them to a common standard population. (Ch. 10)
- Ratio scale (Rom. "scală de raport"): a numerical scale with equal distances and an absolute zero (weight, the concentration of a substance), for which ratios are also meaningful: 100 kg means twice as much as 50 kg. (Ch. 2)
- Recall bias (Rom. "bias de memorie"): a form of information bias specific to case-control studies, in which patients remember past exposures better than (or differently from) healthy controls. (Ch. 10)
- Reference interval (normal values; Rom. "interval de referință"): the interval containing the central 95% of the values in a healthy population, calculated as the mean ± 1.96 standard deviations (normal data) or as the range between the 2.5th and 97.5th percentiles. It describes individuals, not the precision of the mean. (Ch. 3)
- Reference standard (gold standard; Rom. "standard de referință"): the best available method for establishing the presence of the disease, against which the test under study is evaluated (for example biopsy for liver fibrosis, the OGTT for diabetes). (Ch. 12)
- Reference value of the test (μ_0; null value; Rom. "valoare de referință a testului"): the value of the parameter specified by the null hypothesis, for example a mean of 120 mmHg in a one-sample t-test, 1 for the RR and OR, or 0 for a difference. (Ch. 5)
- Regulation (EU) No 536/2014 (EU Clinical Trials Regulation; Rom. "Regulamentul (UE) 536/2014"): the European regulation governing the authorization and conduct of clinical trials of medicinal products for human use in the European Union. (Ch. 13)
- Rejection region (Rom. "regiune de respingere"): the set of values of the test statistic for which H0 is rejected. In a two-sided test it is split between the two tails, each with area $\alpha/2$. (Ch. 5)
- Relative frequency (Rom. "frecvență relativă"): the ratio between the number of occurrences of a value and the total number of observations, expressed as a fraction or a percentage. The relative frequencies of all categories add up to 1 (100%). (Ch. 2)

- Relative risk (RR; risk ratio; Rom. "risc relativ"): the ratio between the risk (incidence) of disease in the exposed and the risk in the unexposed. RR = 1 indicates no association, RR > 1 a risk factor, RR < 1 a protective factor. It is calculated in cohort studies and in clinical trials. (Ch. 10)
- Relative risk reduction (RRR; Rom. "reducerea relativă a riscului"): the decrease in risk expressed as a proportion of the risk in the control group, 1 - RR. It looks impressive even when the absolute benefit is small. (Ch. 10)
- Repeated measures ANOVA (Rom. "ANOVA cu măsurători repetate"): the extension of the paired t-test to three or more measurements on the same individuals (for example fasting plasma glucose at 0, 3 and 6 months). Among other things, it assumes sphericity. (Ch. 6)
- Representative sample (Rom. "eșantion reprezentativ"): a sample that reproduces the structure of the population (by age, sex, area of residence, etc.) in proportions close to the real ones. It is obtained by random selection, in which every individual has the same chance of being chosen and the choices are independent. (Ch. 1)
- Reproducibility (Rom. "reproductibilitate"): the possibility for other researchers to obtain the same results starting from the same data and the same methods. It requires documented data, fully described scripts or analysis steps, and stated software versions. (Ch. 14)
- Research ethics committee (Rom. "comisie de etică a cercetării"): an independent body that reviews and approves the protocol of a study in humans before it begins, checking that the rights, safety and well-being of the participants are respected. (Ch. 13)
- Research hypothesis (Rom. "ipoteză de cercetare"): the statement that the researcher wants to test, formulated in clinical terms (for example, "teleconsultation lowers blood glucose more than standard counseling"). The statistical hypotheses are derived from it. (Ch. 5)
- Research integrity (Rom. "integritatea cercetării"): adherence to honesty at every stage of research. The serious forms of misconduct are fabrication, falsification and plagiarism; to these are added questionable practices, such as selective reporting of significant results. (Ch. 1)
- Residual (Rom. "reziduu"): the difference between the observed value and the value predicted by the model for an individual. Residual analysis checks linearity, homogeneity of variance and normality. (Ch. 9)
- Risk difference (RD; Rom. "diferența de risc"): the absolute difference between the risk in the exposed (or treated) and the risk in the unexposed (or controls). It expresses the effect in percentage points and is the basis of the NNT. (Ch. 10)
- Risk factor (Rom. "factor de risc"): a characteristic, behavior or exposure that increases the probability of developing a disease (smoking for COPD, obesity for diabetes). Association does not automatically mean causality. (Ch. 10)
- Robust statistic (Rom. "indicator robust"): a statistic that is little influenced by extreme values or by departures from normality, such as the median and the interquartile range. (Ch. 2)
- Robustness (Rom. "robustețe"): the property of a test of giving correct results even when its conditions are moderately violated. The t-test is fairly robust to slight departures from normality in large samples. (Ch. 8)
- ROC curve (receiver operating characteristic curve; Rom. "curba ROC"): the graph of sensitivity against 1 - specificity for all possible cut-off values of a numerical test. The closer the curve comes to the upper left corner, the better the test. (Ch. 12)
- Sample (Rom. "eșantion"): the subset of individuals from the population on whom the measurements are actually made. The quality of the conclusions depends more on how the sample was chosen than on its size. (Ch. 1)
- Sample size (Rom. "dimensiunea eșantionului"): the number of statistical units needed to estimate a parameter with the desired precision or to detect a given effect with a fixed power. It is calculated before the study and rounded up. (Ch. 4)

- Sampling distribution (Rom. "distribuție de eșantionare"): the distribution that a statistic (for example the mean) would have if we drew a very large number of samples of the same size from the same population. Its standard deviation is the standard error. (Ch. 4)
- Scatter plot (Rom. "diagramă de dispersie"): a graph in which each individual is a point whose coordinates are given by two numerical variables. It is mandatory before calculating a correlation coefficient. (Ch. 9)
- Screening test (Rom. "test de screening"): a test applied to people without symptoms in order to identify the disease at an early stage. It must be highly sensitive, inexpensive and acceptable; positive results are confirmed with a specific test. (Ch. 12)
- Secondary outcome (Rom. "criteriu secundar de evaluare"): an additional outcome, analyzed to complete the picture of the effects of the intervention; conclusions based only on secondary outcomes are exploratory. (Ch. 13)
- Selection bias (Rom. "bias de selecție"): a systematic error that arises when the individuals included differ systematically from the target population or when the groups being compared are formed differently (for example, volunteers are healthier than average). (Ch. 10)
- Selective reporting (Rom. "raportare selectivă"): presenting only the significant or favorable results out of all the analyses performed. It distorts the literature and is prevented by prospective registration of studies. (Ch. 5)
- Sensitivity (Se; Rom. "sensibilitate"): the proportion of diseased people in whom the test is positive, a/(a + b). A highly sensitive test misses few diseased people; if it is negative, it rules out the disease (SnNOut). (Ch. 12)
- Serial testing (Rom. "teste în serie"): applying the second test only to those who are positive on the first; the patient is considered diseased only if both are positive. It increases specificity and decreases sensitivity. (Ch. 12)
- Shapiro-Wilk test (Rom. "testul Shapiro-Wilk"): a normality test in which H0 states that the data come from a normal distribution. A $p < 0.05$ indicates a departure from normality; a $p \geq 0.05$ does not prove normality, only that no departure was detected. (Ch. 3)
- Significance level (α; Rom. "prag de semnificație"): the maximum accepted probability of a type I error, fixed before the analysis, usually 0.05. If $p < \alpha$, the result is declared statistically significant. (Ch. 5)
- Simple linear regression (Rom. "regresie liniară simplă"): a model that describes the dependent variable y as a linear function of a predictor x, $y = a + b \times x$, and allows values of y to be predicted. (Ch. 9)
- Simple random sampling (Rom. "eșantionare aleatoare simplă"): a selection method in which every individual in the population has the same probability of entering the sample, for example by drawing lots with a table of random digits or with a random number generator. (Ch. 1)
- Skewness (Rom. "asimetrie"): a measure of the lack of symmetry of a distribution. Positive skewness (a tail to the right) is common in biological variables that cannot fall below zero, such as triglycerides; in that case the mean exceeds the median. (Ch. 3)
- Slope of the regression line (regression coefficient, b; Rom. "panta dreptei de regresie", "coeficient de regresie"): the average change in y when x increases by one unit. Its sign is the same as the sign of the correlation coefficient. (Ch. 9)
- SnNOut (Rom. "SnNOut"): a mnemonic: a highly sensitive (Sn) test that is negative (N) rules out the disease (Out, rule out). (Ch. 12)
- Spearman rank correlation coefficient (rho; Rom. "coeficientul de corelație Spearman"): the Pearson coefficient calculated on ranks; it measures a monotonic (not necessarily linear) relationship. It is used for ordinal variables or for non-normal numerical variables. (Ch. 9)

- Specificity (Sp; Rom. "specificitate"): the proportion of non-diseased people in whom the test is negative, d/(c + d). A highly specific test gives few false positive results; if it is positive, it confirms the disease (SpPIn). (Ch. 12)
- Spectrum bias (Rom. "bias de spectru"): incorrect estimation of sensitivity and specificity when the patients studied do not reflect the spectrum of those seen in practice (for example only severe cases and perfectly healthy volunteers). (Ch. 12)
- Sphericity (Rom. "sfericitate"): the condition of repeated measures ANOVA that the variances of the differences between all pairs of time points be equal; it is checked with Mauchly's test and, if it is not met, a correction (Greenhouse-Geisser) is applied. (Ch. 6)
- SpPIn (Rom. "SpPIn"): a mnemonic: a highly specific (Sp) test that is positive (P) confirms the disease (In, rule in). (Ch. 12)
- Spreadsheet (Rom. "foaie de calcul"): a table-type program (Google Sheets, Excel) in which the data are organized with one row for each statistical unit and one column for each variable. (Ch. 2)
- Spurious correlation (Rom. "corelație falsă", "corelație spurioasă"): a statistical association between two variables that are not directly related, produced by a third, common variable or by chance. (Ch. 9)
- Stages of a statistical study (Rom. "etapele studiului statistic"): the sequence of formulating the question, planning (design, sample, variables), collecting the data, checking and cleaning, descriptive analysis, inferential analysis, interpretation and reporting. (Ch. 1)
- Standard deviation (SD; Rom. "abatere standard"): the square root of the variance; it shows how widely the individual values are scattered around the mean, in the unit of measurement of the variable. It does not decrease as the sample grows. (Ch. 2)
- Standard error of a proportion (Rom. "eroarea standard a proporției"): $\sqrt{\frac{p_i(1-p_i)}{n}}$; it measures the precision with which an observed proportion estimates the population proportion. (Ch. 4)
- Standard error of the estimate (SEE; Rom. "eroarea standard a estimării"): the standard deviation of the residuals, expressed in the unit of the dependent variable; it shows how far, on average, the points lie from the line. (Ch. 9)
- Standard error of the mean (SEM, SE; Rom. "eroarea standard a mediei"): the standard deviation of the distribution of sample means, estimated by $\frac{s}{\sqrt{n}}$. It measures the precision of the mean and decreases as the sample size increases. (Ch. 4)
- Standard normal distribution (Rom. "distribuție normală standard"): the normal distribution with mean 0 and standard deviation 1, obtained through the z transformation. Its areas are tabulated (Appendix E). (Ch. 3)
- STARD (Standards for Reporting of Diagnostic Accuracy Studies; Rom. "STARD"): a guideline listing the items that must be reported in a diagnostic accuracy study, including the patient flow diagram and the 2 x 2 table. (Ch. 12)
- Statistic (sample indicator; Rom. "statistică", "indicator de eșantion"): a value calculated from the sample data (the mean m, the standard deviation s, the proportion p), used as an estimate of the corresponding parameter. It varies from one sample to another. (Ch. 4)
- Statistical analysis plan (SAP; Rom. "plan de analiză statistică"): the document that specifies in advance the statistical methods for each outcome, the handling of missing values and the subgroup analyses, protecting the study from choices made after seeing the data. (Ch. 13)
- Statistical power (Rom. "puterea testului"): the probability of rejecting H0 when it is false, equal to 1 - β. It increases with the effect size, the sample size and the α level and decreases when variability is high; it is usually planned at 80% or 90%. (Ch. 5)

- Statistical significance (Rom. "semnificație statistică"): the situation in which $p < \alpha$, that is, the data are poorly compatible with H0. It does not guarantee that the effect is large, clinically important or causal. (Ch. 5)
- Statistical test (Rom. "test statistic"): a procedure by which we decide, with a known risk of error, whether the data are compatible with the null hypothesis. The choice of test depends on the question, the type of variables, the number of groups and whether they are independent. (Ch. 5)
- Statistical unit (Rom. "unitate statistică"): the individual element on which the observations are made (a patient, an eye, a blood sample). Defining the unit correctly matters: the two eyes of the same patient are not independent observations. (Ch. 1)
- Stratified randomization (Rom. "randomizare stratificată"): randomization carried out separately within strata defined by important prognostic factors (center, sex, severity), in order to balance these factors between arms. (Ch. 13)
- Stratified sampling (Rom. "eșantionare stratificată"): the population is divided into homogeneous strata (for example urban and rural), and a random sample is drawn from each stratum, usually in proportion to the size of the stratum. It increases precision and guarantees that all strata are represented. (Ch. 1)
- STROBE (Strengthening the Reporting of Observational Studies in Epidemiology; Rom. "STROBE"): a reporting guideline for observational studies (cohort, case-control, cross-sectional). (Ch. 10)
- Student t distribution (t distribution; Rom. "distribuția Student", "distribuția t"): a symmetric distribution, similar to the normal distribution but with heavier tails, used when the population standard deviation is unknown. Its shape depends on the degrees of freedom and approaches the normal distribution as they increase. (Ch. 4)
- Study arm (Rom. "braț al studiului"): each of the intervention groups in a clinical trial (for example arms A, B and C in the mg2026 dataset). (Ch. 13)
- Study protocol (Rom. "protocolul studiului"): the document that describes in detail the question, design, population, interventions, outcomes, sample size, randomization, data collection, statistical analysis plan and ethical aspects of a study. (Ch. 13)
- Superiority trial (Rom. "studiu de superioritate"): a trial that aims to show that an intervention is better than the comparator. (Ch. 13)
- Surrogate endpoint (Rom. "criteriu surogat"): an intermediate outcome (blood pressure, HbA1c, cholesterol) used instead of a clinically important outcome (myocardial infarction, death), on the assumption that a change in it predicts the clinical effect. (Ch. 13)
- Systematic review (Rom. "revizie sistematică"): a synthesis of all the studies relevant to a precise question, with an exhaustive search, explicit selection criteria and an assessment of the quality of each study. (Ch. 11)
- Systematic sampling (Rom. "eșantionare sistematică"): a starting point is chosen at random, then every k-th individual on an ordered list (for example every tenth record). It is simple, but it can introduce errors if the list has a hidden periodicity. (Ch. 1)
- Target population (Rom. "populație țintă"): the population for which the conclusions of the study are intended, often broader than the population from which the sample was actually drawn (the source population). The difference between the two limits external validity. (Ch. 1)
- Test selection algorithm (Rom. "algoritmul de alegere a testului"): the sequence of questions that leads to the appropriate test: what type the outcome variable is, how many groups or measurements are being compared, whether they are independent or paired, and whether the conditions of the parametric test are met. (Ch. 8)
- Test statistic (Rom. "statistica testului"): the value calculated from the data (t, z, F, χ^2, U) that measures how far the data are from what H0 would predict. The p value is obtained from it and from its distribution under H0. (Ch. 5)

- Trial registry (Rom. "registru de studii clinice"): a public database in which trials are registered before the first participant is enrolled, with prespecified outcomes. It prevents selective reporting and the publication of favorable results only. (Ch. 13)
- True negative (TN; Rom. "adevărat negativ"): a person without the disease who has a negative test (cell d). (Ch. 12)
- True positive (TP; Rom. "adevărat pozitiv"): a diseased person with a positive test (cell a of the diagnostic table). (Ch. 12)
- Tukey HSD test (Rom. "testul Tukey"): a post hoc test that compares all pairs of means while keeping the overall type I error at α; it gives the differences with adjusted CIs. (Ch. 6)
- Two-sided test (Rom. "test bilateral"): a test in which the alternative hypothesis allows differences in both directions ($\mu 1 \neq \mu 2$). It is the default choice in medical research. (Ch. 5)
- Type I error (Rom. "eroare de tip I"): rejecting the null hypothesis when it is true (a false positive result). Its probability is α. (Ch. 5)
- Type II error (Rom. "eroare de tip II"): failing to reject the null hypothesis when it is false (a false negative result, a real effect missed). Its probability is β. (Ch. 5)
- Unbiased estimator (Rom. "estimator nedeplasat"): an estimator whose mean value, over all possible samples, equals the parameter being estimated. The variance calculated with the denominator n - 1 is an unbiased estimator of the population variance. (Ch. 4)
- Valid percent (Rom. "frecvență validă"): the percentage calculated on the number of valid responses, excluding missing values. When data are missing, the valid percent differs from the percentage calculated on the total, and the report must state which of them was used. (Ch. 2)
- Variable (Rom. "variabilă"): a characteristic that can take different values from one individual to another (age, sex, blood glucose). The type of variable determines which summary measures, graphs and tests can be used. (Ch. 2)
- Variance (Rom. "varianță"): the mean of the squared deviations from the mean, calculated in a sample with the denominator n - 1. Its unit is the square of the unit of the variable, which is why the standard deviation is used in reporting. (Ch. 2)
- Verification bias (Rom. "bias de verificare"): a systematic error in diagnostic accuracy studies that arises when the reference standard is applied to only some of the patients, usually those with a positive test. (Ch. 12)
- Wald interval for a proportion (Rom. "intervalul Wald pentru proporție"): a CI calculated as $p_i \pm z \times \sqrt{\frac{p_i(1-p_i)}{n}}$. It is simple, but inaccurate for small samples or for proportions close to 0 or 1, where it can even give negative limits. (Ch. 4)
- Weighted mean (Rom. "media ponderată"): a mean in which each value contributes in proportion to a weight, for example the size of the subgroup. It is used to combine the means of groups of different sizes. (Ch. 2)
- Welch t-test (Rom. "testul t Welch"): a version of the independent-samples t-test that does not assume equal variances and adjusts the degrees of freedom. It is a safe choice in most situations. (Ch. 6)
- Wilcoxon signed-rank test (Rom. "testul Wilcoxon al rangurilor cu semn"): the nonparametric alternative to the paired t-test; it analyzes the ranks of the absolute values of the differences, taking their sign into account. (Ch. 7)
- Wilson score interval for a proportion (Rom. "intervalul Wilson pentru proporție"): a CI for a proportion with better properties than the Wald interval, recommended especially for small samples and extreme proportions. It is not symmetric around p. (Ch. 4)

- Within-group variability (Rom. "variabilitate în interiorul grupurilor"): the part of the total variability due to differences between individuals in the same group (the residual sum of squares). (Ch. 6)
- Yates continuity correction (Rom. "corecția Yates", "corecția de continuitate"): an adjustment of the chi-square test for 2 x 2 tables that slightly lowers the χ^2 value. It is conservative, and today Fisher's exact test is often preferred when frequencies are small. (Ch. 7)
- Youden index (J; Rom. "indicele Youden"): Se + Sp - 1, between 0 and 1. The cut-off that maximizes J is a common criterion for choosing the optimal cut-off when the costs of the two types of error are considered equal. (Ch. 12)
- Z-score (standardized score; Rom. "scor z", "scor standardizat"): the number of standard deviations by which a value lies from the mean, $z = \frac{x-m}{s}$. It allows values measured on different scales to be compared and areas to be read from the normal table. (Ch. 3)
- Z-test (Rom. "testul z"): a test based on the normal distribution, used for a mean when the population standard deviation is known or the sample is large, and for proportions. (Ch. 5)

Appendix C. The mg2026 Teaching Dataset: Scenario, Data Dictionary and the Errors in the Raw File

Almost all the numerical examples in this book, from the mean systolic blood pressure in Chapter 2 to the ROC curve of the FINDRISC score in Chapter 12, are calculated on the same dataset, called mg2026. We chose this solution for a simple reason: a student who recognizes the variables from one chapter to the next can concentrate on the method without having to relearn the clinical context each time. This appendix brings together in one place everything you need to know about the files: the scenario, the data dictionary, the errors deliberately left in the raw version and the way the dataset is cited in a report or an assignment.

C.1 The Scenario

The dataset describes a simulated pragmatic randomized trial conducted in 12 family practices in the west of the country. A total of 212 adults were assessed for eligibility, of whom 52 were excluded (31 did not meet the criteria, 21 declined to participate), and 160 were randomized (the CONSORT flow diagram is presented in Chapter 13). The main inclusion criterion was increased cardiometabolic risk: overweight or obesity, prediabetes or suspected undiagnosed diabetes.

The participants were randomly assigned to three arms:

- arm A (n = 53): standard lifestyle counseling given by the family physician;
- arm B (n = 53): the same counseling, plus a self-monitoring mobile app (weight, steps, diet);
- arm C (n = 54): counseling, the app and a monthly teleconsultation with the physician.

Measurements were taken at enrollment (time 0), at 3 months and at 6 months. At enrollment, all patients underwent an oral glucose tolerance test (OGTT), considered here the reference standard for the diagnosis of diabetes, which allows the FINDRISC score to be evaluated as a screening test (Chapter 12). The primary outcome of the trial is the decrease in fasting plasma glucose at 6 months; the secondary outcomes are weight loss (including the proportion of patients who lost at least 5%), systolic blood pressure at 6 months and satisfaction with the intervention.

The same dataset also allows observational, cross-sectional analyses of the baseline data: the association between obesity and diabetes, the correlation between BMI and blood pressure, the comparison between the sexes. This is a deliberate teaching choice. In a real study, secondary analyses would have to be declared in the protocol before data collection, and their results would be reported as exploratory.

C.2 The Files

There are two files, available on the discipline's Moodle platform [52]:

- **mg2026_curat.csv**: 160 rows (one row for each patient) and 28 columns, with no missing values and no data entry errors. This is the file on which all the analyses in the lectures and practical sessions are performed.
- **mg2026_brut.csv**: 161 rows and the same 28 columns, with 11 data entry errors left in deliberately and one duplicate row. It is the material for the practical session on data cleaning (Chapter 2 and the first practical sessions).

Both files use the comma as the field separator and the point as the decimal separator (the international CSV convention). When importing into Google Sheets with Romanian regional settings or into Excel in Romanian, the program may misinterpret the decimals; in JASP and

MedCalc the import works directly. The character encoding is UTF-8, and the variable names are written without diacritics, for compatibility with all programs.

C.3 The Data Dictionary

Table C.1 describes each variable. The "Valid range" column gives the validation rule we recommend when checking the raw file (clinically plausible values), followed, in parentheses, by the range actually observed in the clean file. There are no missing values in the clean file; the "Missing" column refers to the raw file.

Table C.1. Dictionary of the variables in the mg2026 dataset.

Variable	Description	Type	Scale	Unit / coding	Valid range (observed)	Missing in raw file
id	patient identifier	text	nominal (identifier)	P001-P160	unique on each row (P001-P160)	0; one id appears twice
brat	randomization arm	categorical	nominal	A, B, C (see C.1)	A, B, C (A 53, B 53, C 54)	0
sex	sex	categorical	nominal dichotomous	F, M	F, M (F 86, M 74)	0
varsta	age at enrollment	numerical discrete	ratio	completed years	18-100 (35-79)	1
mediu	area of residence	categorical	nominal dichotomous	urban, rural	urban, rural (95 / 65)	0
educatie	level of education	ordered categorical	ordinal	1 lower secondary school; 2 high school or vocational school; 3 university; 4 postgraduate	1-4 (1-4)	0
fumat	smoking status	categorical	nominal (3 categories)	nefumator, fost fumator, fumator	the three labels (74 / 33 / 53)	0
activitate	self-rated physical activity	ordered categorical	ordinal	1 very low... 5 very intense	1-5 (1-5)	0
inaltime	height	numerical continuous	ratio	cm, whole number	130-210 (145-192)	0
greutate0	weight at enrollment	numerical continuous	ratio	kg, one decimal place	35-250 (58,1-159,9)	0
imc0	body mass index at enrollment	numerical continuous (derived)	ratio	kg/m², one decimal place; greutate0 / (inaltime/100)²	15-70 (23,5-48,3)	0
talie	waist circumference	numerical continuous	ratio	cm, whole number	50-180 (65-140)	0
tas0	systolic blood pressure at enrollment	numerical	ratio	mmHg	70-260 (102-169)	0
tad0	diastolic blood pressure at enrollment	numerical	ratio	mmHg	40-150 (58-103)	0

Variable	Description	Type	Scale	Unit / coding	Valid range (observed)	Missing in raw file
fc0	heart rate at enrollment	numerical	ratio	beats/min	35-180 (51-102)	0
colesterol	total cholesterol	numerical	ratio	mg/dL	80-450 (119-306)	0
hdl	HDL cholesterol	numerical	ratio	mg/dL	10-120 (20-78)	0
trigliceride	serum triglycerides	numerical	ratio	mg/dL	20-1500 (44-442)	0
glicemie0	fasting plasma glucose at enrollment	numerical	ratio	mg/dL, whole number	40-400 (75-154)	0
glicemie3	fasting plasma glucose at 3 months	numerical	ratio	mg/dL	40-400 (49-144)	0
glicemie6	fasting plasma glucose at 6 months	numerical	ratio	mg/dL	40-400 (45-150)	1
hba1c0	glycated hemoglobin at enrollment	numerical continuous	ratio	%, one decimal place	3,5-15 (4,5-7,0)	0
findrisc	FINDRISC diabetes risk score	numerical discrete (score)	ordinal treated as numerical	points	0-26 (0-25)	0
diabet_ogtt	diabetes confirmed by OGTT at enrollment	categorical	nominal dichotomous	0 no, 1 yes	0, 1 (29 with diabetes)	0
greutate6	weight at 6 months	numerical continuous	ratio	kg, one decimal place	35-250 (50,7-155,6)	0
tas6	systolic blood pressure at 6 months	numerical	ratio	mmHg	70-260 (85-166)	0
scadere5	weight loss of at least 5% at 6 months	categorical (derived)	nominal dichotomous	0 no, 1 yes; 1 if (greutate0 - greutate6)/greutate0 ≥ 0.05	0, 1 (38 with 1)	0
satisfactie	satisfaction with the intervention at 6 months	ordered categorical	ordinal (Likert)	1 not at all satisfied... 5 very satisfied	1-5 (1-5)	0

A few remarks about scales, because this is exactly where mistakes appear in the exam. The variables educatie, activitate and satisfactie are coded with digits, but they are ordinal: their mean has limited meaning, and comparisons use the median, the interquartile range and nonparametric tests (Chapter 7). Strictly speaking, the FINDRISC score is also ordinal, but with 27 possible values it is treated as numerical in practice; we describe it with both the mean and the median. Blood pressure, blood glucose and weight are ratio variables (they have an absolute zero), although for the statistics in this course the distinction between the interval scale and the ratio scale matters little. Finally, imc0 and scadere5 are derived variables, calculated from other columns; they are not entered by hand but recalculated, and the difference between the recorded value and the recalculated one is an excellent tool for error control (see C.4).

C.4 The Errors in the Raw File and How to Correct Them

The file mg2026_brut.csv imitates what happens in any practice or hospital ward where data are copied by hand from patient records into a spreadsheet. The errors are of four kinds: inconsistent coding of categories, impossible values (usually an extra digit, a misplaced decimal separator or a wrong unit of measurement), text entered in a numerical column, and missing values. To these is added a row copied twice. Each error is a distortion of the true value, so once it has been detected it can be corrected; however, the correction is always confirmed in the source document (the patient record, the case report form, the laboratory report), whose role is played here by the clean file.

Table C.2. The errors in mg2026_brut.csv, how they are detected and the correct value.

No.	Patient (data row)	Variable	Value in the raw file	How it is detected	Nature of the error	Correct value (clean file)
1	P005 (5)	sex	f	frequency table for sex: a third category appears	lowercase letter instead of a capital	F
2	P021 (21)	sex	Feminin	the same frequency table: a fourth category	whole word instead of a code	F
3	P034 (34)	tas0	1540	minimum/maximum, box plot; the value breaks the 70-260 rule	an extra zero typed	154
4	P042 (42)	varsta	empty cell	number of valid values (160 instead of 161); COUNTBLANK	missing value	70 (recovered from the patient record)
5	P057 (57)	inaltime	1.72	minimum < 130; value in meters, not centimeters	wrong unit of measurement	172
6	P064 (64)	glicemie0	$\frac{87\text{mg}}{\text{dl}}$	the column is imported as text; ISNUMBER returns FALSE	unit written in the cell	87
7	P084 (84)	fumat	"Fumator " (capital letter and trailing space)	frequency table: a fourth category; LEN shows one extra character	capital letter and invisible space	fumator
8	P089 (89)	hba1c0	5,2	text with a comma in a column that uses the decimal point; the column	wrong decimal separator	5.2 (written 5.2, with a decimal point, in the CSV file)

No.	Patient (data row)	Variable	Value in the raw file	How it is detected	Nature of the error	Correct value (clean file)
				becomes text		
9	P102 (102)	greutate0	7.26	minimum < 35 kg; the recalculated BMI would be 2.7 kg/m²	misplaced decimal point	72,6
10	P121 (121)	mediu	Urban	frequency table: "Urban" separate from "urban"	capital letter	urban
11	P134 (134)	glicemie6	empty cell	number of valid values for glicemie6	missing value	135 (recovered from the patient record)
12	P091 (91 and 161)	all	identical row repeated at the end	COUNTIF on id > 1; in pandas, duplicated()	duplicate row	a single P091 row

A few comments on the table, in the order in which an attentive student would discover them.

The first step is always the frequency table for each categorical variable. In the raw file, sex has four categories (F 84, M 75, f 1, Feminin 1), smoking has four (nefumator 75, fumator 52, fost fumator 33, "Fumator " 1), and area of residence has three (urban 95, rural 65, Urban 1). The totals exceed the values in the clean file by 1 because of the duplicate row. Programs treat "f" and "F" as two different values, and likewise "Urban" and "urban", and an invisible space at the end of the word is enough for "Fumator " to be counted separately. The labels are made uniform automatically (lowercase or capitals according to the dictionary, removal of spaces with the TRIM function), and the result is then compared with the patient record.

The second step is descriptive statistics with the minimum and maximum of each numerical variable. Three errors stand out immediately: SBP 1540 mmHg, height 1.72 and weight 7.26 kg. The value of 1540 mmHg is not a harmless curiosity. If we leave it in the file, the mean SBP calculated on the 161 rows becomes 139.70 mmHg instead of 131.13, and the standard deviation rises from 13.88 to 111.89 mmHg; the median remains 131, which perfectly illustrates the robustness of the median to outliers (Chapter 2). For height and weight, the most elegant check uses the derived variable imc0, which remained correct in the raw file. For P057, the recorded BMI is 26.3 kg/m² and the weight 77.7 kg; the height compatible with these two values is $\sqrt{\frac{77.7}{26.3}} = 1.72m$, which confirms that the height was written in meters and that the correct value is 172 cm. For P102, a BMI of 27.0 and a height of 164 cm give a weight of 27.0 x 1.64² = 72.6 kg: the decimal point was shifted by one position. Cross-checking between variables turns a suspicion into a certainty and is safer than any guesswork.

The third step concerns the column types. When a single text value appears in a numerical column ("87 mg/dl" or "5,2"), the import program declares the whole column as text, and the mean can no longer be calculated at all. The symptom is easy to recognize: JASP displays the variable with the nominal variable icon, and in Google Sheets the ISNUMBER function returns FALSE. The correction (removing the unit or, respectively, replacing the comma with a point) is made in the cell concerned, after which the column is re-imported as numerical.

Missing values (age for P042 and fasting plasma glucose at 6 months for P134) are not filled in with the mean or with a "probable" value without an analysis plan that explicitly specifies the imputation method. In the practice of a study, the value is first looked up in the patient record; if it does not exist, the number of missing values is reported for each variable, and the analysis is performed on the available cases or, in clinical trials, according to the strategy specified in the protocol (Chapter 13). In our scenario, the source contains both values, so they were recovered.

The duplicate row (P091, which appears on row 91 and again on row 161) inflates the sample size to 161 and gives double weight to a single patient. It is detected by checking that the identifier is unique: in Google Sheets, =COUNTIF(A:A, A2) > 1; in JASP, by sorting by id; in Python, df.duplicated(). One of the copies is deleted after confirming that the two rows are identical in all columns.

> **The rule that prevents the error**
>
> Correct the data only against the source document, not by intuition, even when the correction seems obvious. Keep the raw file untouched, work on a copy and record every change in a cleaning log (row, variable, old value, new value, reason). A report that says "we corrected 11 errors and removed one duplicate, as documented in the log in the appendix" can be checked; one that says only "the data were cleaned" cannot.

C.5 How the Dataset Was Generated

The data are entirely simulated. The patients are fictitious, there is no link to real people, and the dataset does not come from any practice or hospital. Consequently, the question of personal data protection within the meaning of the General Data Protection Regulation [32] does not arise, and the files can be used freely for teaching purposes.

The simulation was done in Python, with the random number generator of the numpy library (default_rng) initialized with a fixed seed (2026). The fixed seed makes the dataset reproducible: running the script again gives exactly the same 160 rows. The variables were not generated independently but through clinically plausible relationships: systolic blood pressure increases with age, with BMI and in smokers; HbA1c depends on blood glucose; diabetes confirmed by OGTT and the FINDRISC score depend on the same latent risk variable (age, BMI, waist circumference, physical activity); the effect of the intervention on blood glucose and weight increases from arm A to arm C. The triglyceride distribution was generated as lognormal, to provide an example of a skewed variable that becomes normal after log transformation (Chapter 3). The effect sizes were chosen so that, for teaching purposes, the dataset contains clearly significant results (ANOVA across arms, the BMI-SBP correlation) as well as nonsignificant results (SBP by sex) and borderline results (the Wilcoxon test for weight in arm A, $p = 0.052$).

The practical consequence is important and must be stated honestly: the figures in mg2026 do not represent an estimate of any epidemiological reality. The 18.1% prevalence of diabetes, the effect of the mobile app or the area under the ROC curve of the FINDRISC score (0.852) are properties of the simulation, not results that can be cited as clinical evidence. For real evidence, the chapters of this book use the published studies of colleagues in the discipline and the specialist literature.

C.6 How to Cite the Dataset in Reports and Assignments

When you use mg2026 in a report, an assignment or the practical exam, state three things: that the data are simulated, which file you used (clean or raw) and what processing you did before the analysis. A model sentence for the methods section:

"The analysis used the mg2026 teaching dataset (simulated data, 160 fictitious patients, file mg2026_curat.csv), made available on the Moodle platform of the Discipline of Medical Informatics and Biostatistics, "Victor Babeș" University of Medicine and Pharmacy of Timișoara. Statistical processing was performed in JASP 0.98.1, with a significance level $\alpha = 0.05$."

Model entry for the reference list:

Muntean C. The mg2026 teaching dataset (simulated data for the Biostatistics course, Medicine, first year). In: Muntean C. Biostatistică aplicată în medicina generală. Orizont 2026 [Applied Biostatistics in General Medicine. Horizon 2026]. Timişoara: "Victor Babeş" Publishing House; 2026. Files mg2026_curat.csv and mg2026_brut.csv, available at moodle.umft.ro.

Software is cited with the version used: JASP 0.98.1 [47], MedCalc 24.0 [50], Epi Info 7.2.7 [51]. If you worked on the raw file, add the cleaning log to the appendix of your report (Table C.2 is a model) and report the final number of rows analyzed. Never present results obtained on mg2026 as if they came from real patients, not even in an exercise in writing a conference abstract: the habit of declaring the origin of the data correctly is formed from the first year.

Key Points

- mg2026 is a simulated, reproducible dataset (seed 2026), with 160 fictitious patients in three randomized arms (A 53, B 53, C 54) and measurements at 0, 3 and 6 months.
- The clean file (160 rows) is used for the analyses; the raw file (161 rows) contains 11 errors and one duplicate, for the data cleaning exercise.
- The order of checks: frequencies for categorical variables, minimum and maximum for numerical ones, column types, missing values, uniqueness of the identifier, consistency of derived variables (BMI).
- Each error is a distortion of the true value, but the correction is confirmed in the source document; consistency with derived variables (BMI) is the best check.
- In any piece of work, the data are declared as simulated and the processing is documented.

Appendix D. Formula Sheet

The sheet brings together the formulas used in the course, ordered by the chapter in which they appear. For each formula you will find the meaning of the symbols, the conditions under which it can be used and a one-line numerical example, most often on the mg2026 teaching dataset (Appendix C). The formulas are written in plain text: x means multiplication, √ the square root, Σ the sum of all terms, ln the natural logarithm, and e^x the exponential function. Intermediate results are rounded only at the end, so that you do not accumulate rounding errors; in the exam, keep at least three decimal places in intermediate calculations.

General notation, valid for the whole sheet: n is the sample size, N the total size (or the population size, depending on the context), xi an individual value, m the sample mean, μ the population mean, s the sample standard deviation, σ the population standard deviation, p an observed proportion, α the significance level, β the probability of a type II error.

For 2 x 2 tables two conventions are used, one for risk analysis and one for diagnostic tests. We repeat them here, because half of the mistakes in the exam come from confusing them.

Table D.1. The 2 x 2 table in risk analysis (Chapter 10).

	Disease present	**Disease absent**	**Total**
Exposed	a	b	a + b
Unexposed	c	d	c + d
Total	a + c	b + d	N

Table D.2. The 2 x 2 table for a diagnostic test (Chapter 12), with disease status in the rows, as in the discipline's practical sessions.

	Test positive	**Test negative**	**Total**
Diseased (B+)	a = true positives (TP)	b = false negatives (FN)	a + b
Non-diseased (B-)	c = false positives (FP)	d = true negatives (TN)	c + d
Total	a + c	b + d	N

D.1 Chapter 1. Probability

Classical probability. P(A) = number of favorable cases / number of possible cases. The value lies between 0 and 1. Condition: all cases are equally likely (for example, drawing a record at random). Example: the probability that a patient drawn at random from mg2026 is a smoker is 53/160 = 0.331.

Probability of the complementary event. $P(\text{non} - A) = 1 - P(A)$. Example: $P(\text{ non-smoker or former smoker }) = 1 - 0.331 = 0.669$.

Addition rule. $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$. If A and B are mutually exclusive, P(A and B) = 0 and the term disappears. Example: $P(\text{ woman or smoker }) = \frac{86}{160} + \frac{53}{160} - \frac{30}{160} = \frac{109}{160} = 0.681$.

Conditional probability. $P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$, read as "the probability of A given that B has occurred". Example: $P(\text{ diabetes } \mid \text{obese}) = \frac{\frac{18}{160}}{\frac{54}{160}} = \frac{18}{54} = 0.333$.

Multiplication rule. $P(A \text{ and } B) = P(B) \times P(A \mid B)$. If A and B are independent, P(A | B) = P(A) and $P(A \text{ and } B) = P(A) \times P(B)$. Example: under the hypothesis of independence,

P(woman and smoker) would be 0.5375 x 0.331 = 0.178; the observed value is 30/160 = 0.1875, very close, so sex and smoking appear to be independent in this dataset.

Relative frequency as an estimate of probability. $f = \frac{k}{n}$, where k is the number of occurrences of the event in n observations. Example: 29 patients with confirmed diabetes out of 160, f = 0.181.

D.2 Chapter 2. Describing Data

Absolute, relative and cumulative frequency. $f_i = \frac{n_i}{n}$; percentage frequency = 100 x ni / n; cumulative frequency at category k = the sum of the frequencies of categories 1...k (only for ordinal variables or grouped numerical variables). When there are missing values, the "valid" frequency is calculated on the number of valid responses, not on the total. Example: satisfaction 4 or 5: (55 + 28)/160 = 51.9%; cumulative frequency up to and including score 3: (4 + 19 + 54)/160 = 48.1%.

Arithmetic mean. $m = \frac{\sum x_i}{n}$. Conditions: numerical variable; sensitive to extreme values, so it is preferred for approximately symmetric distributions. Example: SBP in five patients 120, 135, 128, 142, 130 mmHg, $m = \frac{655}{5} = 131$ mmHg.

Weighted mean. mw $= \frac{\sum(w_i \times x_i)}{\sum w_i}$, where wi is the weight (usually the size of the subgroup). It is used when we combine the means of groups of different sizes. Example: the mean decrease in blood glucose across the three arms, $\frac{53\times3.0+53\times11.1+54\times16.3}{160} = 10.17$ mg/dL (the mean calculated directly on all patients is 10.19; the difference comes from rounding the group means).

Geometric mean. mg $= (x_1 \times x_2 \times ... \times \text{xn})^{\frac{1}{n}} = e$^(mean of ln x). Conditions: all values strictly positive; suitable for right-skewed, lognormal variables (triglycerides, antibody titers, viral loads). Example: for 100, 150 and 225, mg $= (100 \times 150 \times 225)^{\frac{1}{3}} = 150$; for the triglycerides in mg2026, the mean of the logarithms is 4.970 and mg $= e^{4.970} = 144.0$ mg/dL (the arithmetic mean is 161.18).

Position of the median. The values are sorted in ascending order; the median lies at position (n + 1)/2. If n is even, the median is the mean of the two central values (positions n/2 and n/2 + 1). Example: for n = 160, the position is 80.5, so the median SBP is the mean of the values in positions 80 and 81: (131 + 131)/2 = 131 mmHg.

Positions of the quartiles (simple method). Q1 at position (n + 1)/4, Q3 at position 3(n + 1)/4, with interpolation between neighboring values. Programs use slightly different interpolation methods, so their results may differ in the first decimal place. Example: in mg2026, Q1 of SBP = 122 and Q3 = 140 mmHg.

Range. A = xmax - xmin. Example: SBP 169 - 102 = 67 mmHg.

Interquartile range. IQR = Q3 - Q1; it contains the central 50% of the values. Example: SBP 140 - 122 = 18 mmHg; BMI $31.4 - 26.2 = \frac{5.2\ \text{kg}}{m^2}$.

Limits for outliers (Tukey's rule, used for the box plot). Lower limit = Q1 - 1.5 x IQR; upper limit = Q3 + 1.5 x IQR. Example: BMI, $31.4 + 1.5 \times 5.2 = \frac{39.2\ \text{kg}}{m^2}$; the maximum value of 48.3 appears as a separate point on the box plot.

Mean absolute deviation. MAD $= \frac{\sum|x_i - m|}{n}$. Example: for 120, 135, 128, 142, 130 (m = 131), the absolute deviations are 11, 4, 3, 11, 1, and MAD $= \frac{30}{5} = 6$ mmHg.

Sample variance. $s^2 = \frac{\sum(x_i - m)^2}{n-1}$. The denominator n - 1 (the degrees of freedom) corrects the tendency to underestimate the population variance. Example: for the same five values, $\sum(x_i - m)^2 = 121 + 16 + 9 + 121 + 1 = 268$, $s^2 = \frac{268}{4} = 67$ mmHg^2.

Standard deviation. $s = \sqrt{s^2}$. It has the same unit of measurement as the data. Example: $s = \sqrt{67} = 8.19$ mmHg; in mg2026, s(SBP) = 13.88 mmHg.

Coefficient of variation. CV = 100 x s / m (%). Conditions: ratio variable with a positive mean; it allows the variability of variables with different units to be compared. As a rough guide, a CV below 10% indicates low variability, 10-30% moderate and above 30% high variability. Example: SBP, $\frac{100 \times 13.88}{131.13} = 10.6\%$; triglycerides, 47.3%.

Standard error of the mean. SE $= \frac{s}{\sqrt{n}}$. It measures the precision of the mean as an estimate of μ, not the spread of the patients. Example: SBP, 13.88/√160 = 13.88/12.649 = 1.097 mmHg.

D.3 Chapter 3. The Normal Distribution, the z-Score, Transformations

The z-score (standardization). $z = \frac{x-m}{s}$ (in the sample) or $z = \frac{x-\mu}{\sigma}$ (in the population). It shows how many standard deviations a value lies from the mean. Condition for reading areas from the normal table: an approximately normal distribution. Example: SBP 160 mmHg, $z = \frac{160-131.13}{13.88} = 2.08$; from Table E.1, Φ(2.08) = 0.9812, so approximately 1.9% of the values would exceed 160 mmHg.

Inverse transformation. x = m + z x s. Example: the 97.5th percentile of SBP (z = 1.96) is 131.13 + 1.96 x 13.88 = 158.3 mmHg.

Intervals for individual values (the 68-95-99.7 rule). For a normal distribution, approximately 68.3% of the values lie within $m \pm s$, 95.4% within $m \pm 2s$, 99.7% within $m \pm 3s$, and exactly 95% within $m \pm 1.96s$. Example (SBP): $m \pm s = 117.3 - 145.0$; $m \pm 1.96s = 103.9 - 158.3$ (contains 93.8% of the actual observations); $m \pm 2s = 103.4 - 158.9$; $m \pm 3s = 89.5 - 172.8$ mmHg.

Intervals for the population mean (based on the standard error). μ lies, with the corresponding probabilities, within $m \pm \text{SE}(68.3\%)$, $m \pm 2\text{SE}(95.4\%)$, $m \pm 3\text{SE}(99.7\%)$ or $m \pm 1.96\text{SE}(95\%)$. Condition: n sufficiently large (central limit theorem) or normal data. Example (SBP, SE = 1.097): $m \pm \text{SE} = 130.03 - 132.23$; $m \pm 2\text{SE} = 128.94 - 133.32$; $m \pm 3\text{SE} = 127.84 - 134.42$ mmHg.

Log transformation. y = ln(x); after the analysis on the logarithmic scale, the results are brought back to the original scale with $x = e^y$. Condition: x > 0; useful for positive skewness. Example: the skewness of triglycerides falls from 0.95 to -0.28 after log transformation, and the 95% CI of the geometric mean is $e^{4.970 \pm \frac{1.975 \times 0.488}{\sqrt{160}}} = 133.44 - 155.42$ mg/dL.

Skewness and excess kurtosis (interpretation). A skewness close to 0 indicates symmetry; a positive skewness indicates a tail to the right. An excess kurtosis (kurtosis minus 3) close to 0 indicates a shape close to the normal one. Rough rule: absolute values above 1 suggest a substantial departure. Example: BMI, skewness 1.42 and excess kurtosis 3.68, hence a non-normal distribution, confirmed by Shapiro-Wilk ($p < 0.001$).

D.4 Chapter 4. Estimation, Confidence Intervals, Sample Size

CI for a mean using the normal distribution (σ known or n large). $\mathrm{CI}(1-\alpha) = m \pm \frac{z\left(1-\frac{\alpha}{2}\right)\times\sigma}{\sqrt{n}}$; z = 1.645 (90%), 1.960 (95%), 2.576 (99%). Example: SBP, $131.13 \pm 1.96 \times 1.097 = 128.98 - 133.28$ mmHg.

CI for a mean using the Student distribution (σ unknown, the usual case). $\mathrm{CI}(1-\alpha) = m \pm \frac{t(\alpha;\, n-1)\times s}{\sqrt{n}}$. Condition: approximately normal data or large n. Example: SBP, t(0.05; 159) = 1.975, $\mathrm{CI}\,95\% = 131.13 \pm 1.975 \times 1.097 = 128.96 - 133.30$ mmHg; 99% CI = 128.27-133.99 mmHg.

Width of the CI and the effect of n. Width = $\frac{2\times t\times s}{\sqrt{n}}$; if n is multiplied by 4, the width falls to approximately half. Example: the CI of the mean SBP has a width of 27.2 mmHg at n = 10, 10.6 at n = 40 and 4.3 at n = 160.

Standard error of a proportion. $\mathrm{SE}(p_i) = \sqrt{\frac{p_i\times(1-p_i)}{n}}$. Example: diabetes 29/160, $p_i = 0.181$, $\mathrm{SE} = \sqrt{\frac{0.181\times0.819}{160}} = 0.0305$.

CI for a proportion, Wald method. $\mathrm{CI}\,95\% = p_i \pm 1.96 \times \sqrt{\frac{p_i(1-p_i)}{n}}$. Conditions: n x $p_i \geq 5$ and n x (1 - p_i) ≥ 5 (some authors require 10); for small proportions or small samples it gives intervals that are too narrow or negative limits. Example: $0.181 \pm 1.96 \times 0.0305 = 0.122 - 0.241$, that is, 12.2-24.1%.

CI for a proportion, Wilson method (recommended). Center = (p_i + z²/(2n)) / (1 + z²/n); half-width = $\left[\frac{z}{1+\frac{z^2}{n}}\right] \times \sqrt{\frac{p_i(1-p_i)}{n} + \frac{z^2}{4\,n^2}}$; CI = center ± half-width. It also works well for proportions close to 0 or 1. Example: for 29/160 and z = 1.96, center = 0.1887, half-width = 0.0595, 95% CI = 12.9-24.8%.

CI of the geometric mean. $\mathrm{CI} = e^{\mathrm{ml}\pm\frac{t\times\mathrm{sl}}{\sqrt{n}}}$, where ml and sl are the mean and the standard deviation of ln x. Example: triglycerides, 133.44-155.42 mg/dL (see D.3).

Sample size for estimating a mean. $n = \left(\frac{z\left(1-\frac{\alpha}{2}\right)\times\sigma}{E}\right)^2$, where E is the desired margin of error (the half-width of the CI). It is always rounded up. Example: σ = 14 mmHg, $E = \pm 3$ mmHg, $n = \left(\frac{1.96\times14}{3}\right)^2 = 83.7$, so 84 patients.

Sample size for estimating a proportion. $n = \frac{z^2\times p_i(1-p_i)}{E^2}$. If p_i is unknown, $p_i = 0.5$ is used (the most conservative option). Example: $p_i = 0.20$, E = 0.05, $n = \frac{1.96^2\times0.2\times0.8}{0.0025} = 245.9$, so 246.

Sample size per group for comparing two means (independent-samples t-test). n per group = 2 x (z(1-α/2) + z(1-β))² / d², where $d = \frac{\mu_1-\mu_2}{\sigma}$ is the effect size; z(1-β) = 0.842 for 80% power and 1.282 for 90% power. Example: d = 0.5, α = 0.05, power 80%: $\frac{2\times(1.96+0.842)^2}{0.25} = 62.8$, that is, 63 per group; programs that use the t distribution (G*Power) give 64 per group, and for d = 0.3 they give 176.

Sample size per group for comparing two proportions. n per group = (z(1-α/2) + z(1-β))² x [p1(1 - p1) + p2(1 - p2)] / (p1 - p2)². Example: p1 = 0.30, p2 = 0.11, α = 0.05, power 80%: $\frac{7.849\times(0.21+0.0979)}{0.0361} = 66.9$, so 67 per group.

Correction for dropout. n adjusted $= \frac{n}{1-r}$, where r is the estimated proportion of losses to follow-up. Example: 64 per group and 15% dropout, 64/0.85 = 75.3, so 76 per group.

D.5 Chapter 5. Fundamentals of Hypothesis Testing

Errors and power. α = P(we reject H0 | H0 true), the type I error; β = P(we do not reject H0 | H0 false), the type II error; power = 1 - β; confidence level = 1 - α. Example: a study planned with α = 0.05 and 80% power accepts a 5% risk of a false positive result and a 20% risk of missing a real effect of the assumed size.

Decision rule using p. If p < α, H0 is rejected (statistically significant difference); if p ≥ α, H0 is not rejected (no difference has been demonstrated). Example: SBP in women vs men, p = 0.81 > 0.05, so H0 is not rejected.

Decision rule using the critical value. H0 is rejected if |calculated statistic| ≥ the critical value from the table (Appendix E) for the given α and degrees of freedom. Example: t = 11.14 with df = 159 > critical t 1.975, so p < 0.05.

Z-test for a mean (σ known). $z = \frac{m-\mu_0}{\frac{\sigma}{\sqrt{n}}}$. Conditions: σ known from the population or n large, approximately normal data. Example: mean SBP 131.13 compared with a reference value μ0 = 128 mmHg, with σ = 14: $z = \frac{3.13}{\frac{14}{\sqrt{160}}} = 2.83$, two-sided p = 0.005.

One-sample t-test. $t = \frac{m-\mu_0}{\frac{s}{\sqrt{n}}}$, with df = n - 1. Conditions: approximately normal data (or large n), independent observations. Example: the same comparison with s = 13.88: $t = \frac{3.13}{1.097} = 2.85$, df = 159, p = 0.005.

Z-test for a proportion. $z = \frac{p-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$. Conditions: n x p0 ≥ 5 and n x (1 - p0) ≥ 5. Example: diabetes 0.181 compared with an assumed prevalence p0 = 0.12: $z = \frac{0.061}{\sqrt{\frac{0.12\times0.88}{160}}} = \frac{0.061}{0.0257} = 2.38$, two-sided p = 0.017.

Bonferroni correction for multiple comparisons. α' = α / k, where k is the number of comparisons. Example: three pairwise comparisons between arms A, B and C, α' = 0.05/3 = 0.0167.

D.6 Chapter 6. Parametric Tests: t and ANOVA

Paired t-test. d = the difference for each patient; md = the mean of the differences, sd = their standard deviation; $t = \frac{\text{md}}{\frac{\text{sd}}{\sqrt{n}}}$, df = n - 1. Conditions: paired measurements (the same patient at two time points or with two methods), approximately normal differences. Example: fasting plasma glucose at 0 vs 6 months, md = 10.19, sd = 11.57, n = 160: $t = \frac{10.19}{\frac{11.57}{12.649}} = 11.14$, df = 159, p < 0.001.

CI for the mean of paired differences. $\text{CI}\,95\% = \text{md} \pm \frac{t(0.05;\, n-1)\times\text{sd}}{\sqrt{n}}$. Example: $10.19 \pm 1.975 \times 0.915 = 8.38 - 11.99$ mg/dL.

Pooled variance for two independent groups. $\text{sp}^2 = \frac{(n_1-1)\times {s_1}^2+(n_2-1)\times {s_2}^2}{n_1+n_2-2}$. Example: SBP, men (130.85; 13.25; 74) and women (131.37; 14.47; 86): $\text{sp}^2 = \frac{73\times 175.56+85\times 209.38}{158} = 193.76$, sp = 13.92 mmHg.

Classical independent-samples t-test (Student, equal variances). $t = \frac{m_1-m_2}{\text{sp}\times\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, df = n1 + n2 - 2. Conditions: independent groups, approximately normal data in each group, equal variances (nonsignificant Levene test). Example: $t = -\frac{0.52}{13.92\times\sqrt{\frac{1}{74}+\frac{1}{86}}} = -0.24$, df = 158, p = 0.81.

Welch t-test (unequal variances). $t = \frac{m_1-m_2}{\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}}$; $\text{df} = \frac{\left(\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}\right)^2}{\frac{\left(\frac{{s_1}^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{{s_2}^2}{n_2}\right)^2}{n_2-1}}$. It can be used at any time, not only when variances are unequal; JASP offers it as an option. Example: SBP by sex, $t = -\frac{0.52}{2.193} = -0.24$, df = 157.4, p = 0.81.

CI for the difference between two independent means. $\text{CI}\,95\% = (m_1 - m_2) \pm t \times \text{SE}(m_1 - m_2)$, with $\text{SE} = \text{sp} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ or the Welch SE. Example: the M - F difference in SBP, -0.5 mmHg, 95% CI from -4.9 to 3.8 mmHg (the interval contains 0).

Effect size Cohen's d (independent groups). $d = \frac{m_1-m_2}{\text{sp}}$. Rough benchmarks [76]: 0.2 small, 0.5 medium, 0.8 large. Example: SBP M vs F, $d = -\frac{0.52}{13.92} = -0.04$ (negligible effect).

Effect size for paired data. $\text{dz} = \frac{\text{md}}{\text{sd}}$. Example: fasting plasma glucose 0 vs 6 months, $\text{dz} = \frac{10.19}{11.57} = 0.88$.

One-way ANOVA: sums of squares. Grand mean mG; mj and nj the mean and size of group j; k the number of groups; N the total. SST (total) = $\sum(x_i - m_G)^2$SSB (between groups) = $\sum n_j \times \left(m_j - m_G\right)^2$SSW (within groups, residual) = $\sum\left(x_i - m_j\right)^2 = \sum\left(n_j - 1\right) \times {s_j}^2$SST = SSB + SSW. Example: decrease in fasting plasma glucose at 6 months by arm, SSB = 4806.9, SSW = 16459.5, SST = 21266.4.

One-way ANOVA: the table and the F statistic. Between-group df = k - 1; residual df = N - k; $\text{MSB} = \frac{\text{SSB}}{k-1}$; $\text{MSW} = \frac{\text{SSW}}{N-k}$; $F = \frac{\text{MSB}}{\text{MSW}}$. Conditions: independent groups, normality in each group (or of the residuals), homogeneous variances (Levene). Example: $\text{MSB} = \frac{4806.9}{2} = 2403.4$; $\text{MSW} = \frac{16459.5}{157} = 104.84$; F(2, 157) = 22.93, p < 0.001.

Table D.3. Structure of the ANOVA table, with the values from the example.

Source of variation	**SS**	**df**	$\text{MS} = \frac{\text{SS}}{\text{df}}$	**F**
Between groups	4806,9	2	2403,4	22,93
Within groups (residual)	16459,5	157	104,84	
Total	21266,4	159		

Eta squared. $\eta^2 = \frac{\text{SSB}}{\text{SST}}$; the proportion of the total variability explained by group membership. Rough benchmarks: 0.01 small, 0.06 medium, 0.14 large. Example: 4806.9/21266.4 = 0.226.

D.7 Chapter 7. Nonparametric Tests and the Chi-Square Test

Expected frequencies. E = (row total x column total) / N. Example: in the obesity x diabetes table, $E(\text{ obese with diabetes }) = \frac{54 \times 29}{160} = 9.79$; the others: 44.21; 19.21; 86.79.

The chi-square statistic (Pearson). $\chi^2 = \frac{\sum(O-E)^2}{E}$, summed over all cells, with df = (r - 1) x (c - 1). Conditions: independent observations, absolute frequencies (not percentages); at most 20% of cells with E < 5 and no cell with E < 1; otherwise, Fisher's exact test (for 2 x 2) or grouping of categories. Example: obesity x diabetes, χ^2 = 12.70, df = 1, p < 0.001.

Shortcut formula for the 2 x 2 table. $\chi^2 = \frac{N \times (\text{ad} - \text{bc})^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)}$. Example: $\frac{160 \times (18 \times 95 - 36 \times 11)^2}{54 \times 106 \times 29 \times 131} = 12.70$.

Yates continuity correction (2 x 2). $\chi^2 Y = \frac{N \times \left(|\text{ad} - \text{bc}| - \frac{N}{2}\right)^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)}$. Example: 11.20, p = 0.0008.

Degrees of freedom. df = (number of rows - 1) x (number of columns - 1). Example: arm (3) x decrease ≥5% (2), df = 2 x 1 = 2; χ^2 = 6.77, p = 0.034.

Mann-Whitney U test. All values are ordered together and ranks are assigned (for ties, the mean rank). R1 = the sum of the ranks in group 1. $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$; U2 = n1 x n2 - U1; U = min(U1, U2) for the table. Conditions: two independent groups, a variable that is at least ordinal. Example: group A {12, 15, 18}, group B {14, 20, 22, 25}; the ranks of A are 1, 3, 4, R1 = 8, U1 = 8 - 6 = 2, U2 = 12 - 2 = 10, U = 2.

Normal approximation for U (large samples). $z = \frac{U - \frac{\text{n1n2}}{2}}{\sqrt{\frac{n_1 \times n_2 \times (n_1 + n_2 + 1)}{12}}}$. Example: satisfaction in arm C vs A, U = 2265.5, 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 (without the correction for ties).

Rank-biserial correlation (effect size for Mann-Whitney). r = 2 x U1/(n1 x n2) - 1, where U1 corresponds to the group with the higher ranks. Example: 2 x 2265.5/2862 - 1 = 0.58.

Wilcoxon signed-rank test for paired data. The differences are calculated, zero differences are removed, the absolute values are ordered and ranks are assigned; W+ = the sum of the ranks of the positive differences, W- = the sum of the ranks of the negative ones, and $W^+ + W^- = \frac{n(n+1)}{2}$. The test statistic is min(W+, W-). Example: for 6 nonzero differences, $W^+ + W^- = \frac{6 \times 7}{2} = 21$.

Kruskal-Wallis test. $H = \left[\frac{12}{N(N+1)}\right] \times \sum\left(\frac{{R_j}^2}{n_j}\right) - 3(N+1)$, with df = k - 1 (chi-square approximation). Conditions: k ≥ 3 independent groups, a variable that is at least ordinal. Example: decrease in fasting plasma glucose by arm, H = 36.63, df = 2, p < 0.001.

Friedman test. $\chi^2 F = \left[\frac{12}{n \times k \times (k+1)}\right] \times \sum {R_j}^2 - 3n(k+1)$, where Rj is the sum of the ranks for time point j (ranks are assigned within each patient), df = k - 1. Example: fasting plasma glucose at 0, 3 and 6 months, χ^2F = 74.67, df = 2, p < 0.001.

Spearman correlation coefficient. $\text{rho} = 1 - \frac{6 \times \sum d^2}{n(n^2-1)}$, where d is the difference between the rank of x and the rank of y for the same individual (the exact formula when there are no ties).

Example: ranks x 1, 2, 3, 4, 5 and y 2, 1, 4, 3, 5 give d = -1, 1, -1, 1, 0, $\sum d^2 = 4$, rho $= 1 - \frac{24}{120} =$ 0.80.

D.8 Chapter 8. Choosing the Test (Numerical Benchmarks)

Checking normality. Shapiro-Wilk test: p ≥ 0.05 means that no departure from normality has been demonstrated (parametric tests can be used); p < 0.05 points toward a transformation or toward nonparametric tests. Example: SBP, p = 0.303 (parametric); BMI, p < 0.001 (nonparametric or transformation).

Checking homogeneity of variances. Levene test: p ≥ 0.05 allows the classical t-test or classical ANOVA; p < 0.05 calls for Welch. Example: SBP by sex, Levene p = 0.529.

The expected frequency rule. Chi-square if all E ≥ 5 (or at most 20% of cells below 5); Fisher otherwise. Example: arm x decrease ≥5%, the smallest E = 12.6, so chi-square.

D.9 Chapter 9. Correlation and Regression

Pearson correlation coefficient. $r = \frac{\sum(x_i - m_x)(y_i - m_y)}{\sqrt{\left[\sum(x_i - m_x)^2 \times \sum(y_i - m_y)^2\right]}}$, with values between -1 and 1. Conditions: both variables numerical, linear relationship, approximately normal distribution, no influential extreme values. Example: BMI and SBP in mg2026, r = 0.405.

The t-test for the correlation coefficient. $t = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}}$, df = n - 2; H0: ρ = 0. Example: $\frac{0.405 \times \sqrt{158}}{\sqrt{1-0.164}} = 5.57$, p < 0.001.

Coefficient of determination. $R^2 = r^2$ (in simple regression); the proportion of the variability of y explained by x. Example: 0.405² = 0.164, that is, BMI explains 16.4% of the variability of SBP.

Slope of the regression line (least squares method). $b = \frac{\sum(x_i - m_x)(y_i - m_y)}{\sum(x_i - m_x)^2} = \frac{r \times s_y}{s_x}$. Meaning: the average change in y when x increases by one unit. Example: $b = \frac{0.405 \times 13.88}{3.91} = 1.44$ (the software, using unrounded data, gives 1.435 mmHg per kg/m²).

Intercept. a = my - b x mx. Example: a = 131.13 - 1.435 x 29.04 = 89.46 (software: 89.47 mmHg).

Regression equation and prediction. ŷ = a + b x x. Condition: prediction makes sense only within the range of observed x values. Example: SBP = 89.47 + 1.435 x BMI; at a BMI of 30, ŷ = 132.5 mmHg; at a BMI of 35, ŷ = 139.7 mmHg.

The residual. ei = yi - ŷi. Example: a patient with a BMI of 30 and an SBP of 140 has a residual of 140 - 132.5 = 7.5 mmHg.

Standard error of the estimate (SEE). SEE $= \sqrt{\left[\frac{\sum e_i^2}{n-2}\right]} \approx s_y \times \sqrt{1-r^2} \times \sqrt{\frac{n-1}{n-2}}$. Expressed in the units of y, it shows how far, on average, the points lie from the line. Example: $13.88 \times \sqrt{0.836} \times \sqrt{\frac{159}{158}} = 12.73$ mmHg.

CI for the slope. $b \pm t(0.05;\ n-2) \times \mathrm{SE}(b)$. Example: 1.435, 95% CI 0.925-1.944; the interval does not contain 0, so the slope is significant.

Adjusted coefficient of determination (multiple regression). $R^2_{\text{adj}} = 1 - \frac{(1-R^2)\times(n-1)}{n-p-1}$, where p is the number of predictors (variables in the model, including indicator variables). Example: SBP ~ BMI + age + sex + smoking (5 terms), $R^2 = 0.308$, $R^2_{\text{adj}} = 1 - \frac{0.692\times159}{154} = 0.286$.

Logistic regression: from coefficient to OR. $\text{OR} = e^b$; $\text{CI}\,95\% = e^{b\pm1.96\times\text{SE}(b)}$. Example: for age, OR = 1.11 per year, that is, the odds of diabetes increase by approximately 11% for each additional year, at constant obesity status and sex.

D.10 Chapter 10. Epidemiology and Risk Analysis

Prevalence. P = number of existing cases at a given time / population at that time. Example: diabetes confirmed at enrollment, 29/160 = 18.1%.

Cumulative incidence (risk). CI = number of new cases in a period / population at risk at the start of the period. Example: if 12 of 400 people without diabetes develop the disease over 5 years, CI = 3.0% over 5 years.

Incidence rate (incidence density). IR = number of new cases / total person-time of follow-up. Example: 12 new cases in 1,850 person-years, IR = 6.5 per 1,000 person-years.

Approximate prevalence-incidence relationship (stable, rare disease). P ≈ IR x D, where D is the mean duration of the disease. Example: IR = 2 per 1,000 per year and a duration of 10 years, P ≈ 2%.

Relative risk (RR) or prevalence ratio (PR). $\text{RR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$. It is calculated from cohort studies and clinical trials (RR) or from cross-sectional studies (PR). RR = 1 means no association, RR > 1 a risk factor, RR < 1 a protective factor. Example: obesity and diabetes, (18/54)/(11/106) = 0.333/0.104 = 3.21.

CI for the RR on the logarithmic scale. $\text{SE}[\ln(\text{RR})] = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$; $\text{CI}\,95\% = e^{\ln\text{RR}\pm1.96\times\text{SE}}$. Example: ln 3.21 = 1.167, SE = 0.344, $\text{CI}\,95\% = e^{1.167\pm0.675} = 1.64 - 6.31$.

Odds ratio (OR). $\text{OR} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a\times d}{b\times c}$. The only valid measure of association in case-control studies; it approximates the RR when the disease is rare (below about 10%). Example: $\frac{18\times95}{36\times11} = \frac{1710}{396} = 4.32$.

CI for the OR on the logarithmic scale (Woolf method). $\text{SE}[\ln(\text{OR})] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$; $\text{CI}\,95\% = e^{\ln\text{OR}\pm1.96\times\text{SE}}$. Example: ln 4.32 = 1.463, SE = 0.430, 95% CI = 1.86-10.03. If the CI does not contain 1, the association is significant at the 5% level.

Risk difference (attributable risk in the exposed, absolute risk reduction). $\text{RD} = p_1 - p_0 = \frac{a}{a+b} - \frac{c}{c+d}$. $\text{CI}\,95\% = \text{RD} \pm 1.96 \times \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}}$. Example: decrease ≥5% in arm C vs A, 16/54 - 6/53 = 0.296 - 0.113 = 0.183, 95% CI 0.034-0.332, that is, 18.3 percentage points (3.4-33.2).

Number needed to treat (NNT). $\text{NNT} = \frac{1}{|\text{RD}|}$, rounded up; the CI limits are obtained by inverting the limits of the CI of the RD. For a harmful effect, the same formula gives the number

needed to harm (NNH). Example: 1/0.183 = 5.5, so 6 patients treated in arm C instead of A for one more patient to lose at least 5% of body weight; 95% CI 3.0-29.1.

Relative risk reduction. RRR $= \frac{p_0 - p_1}{p_0} = 1 -$ RR (for a treatment that lowers risk). Example: if mortality falls from 10% to 7%, RR = 0.7, RRR = 30%, RD = 3 percentage points, NNT = 34.

Attributable fraction among the exposed. AFe = (RR - 1) / RR. Example: (3.21 - 1)/3.21 = 0.689; 68.9% of the cases of diabetes in obese people would be attributable to obesity, if the relationship were causal and unconfounded.

Population attributable fraction (Levin's formula). PAF = pe x (RR - 1) / [1 + pe x (RR - 1)], where pe is the proportion of exposed people in the population; equivalently, PAF $= \frac{\text{Pt} - P_0}{\text{Pt}}$. Example: pe = 54/160 = 0.3375, PAF $= \frac{0.3375 \times 2.21}{1 + 0.3375 \times 2.21} = 0.427$; check: (0.1813 - 0.1038)/0.1813 = 0.427.

Hypothesis test for the RR and the OR. H0: RR = 1 (or OR = 1); the chi-square test on the 2 x 2 table (D.7) or Fisher's exact test is used. Example: $\chi^2 = 12.70$, $p < 0.001$, consistent with the CI of the OR, which excludes 1.

D.11 Chapter 11. Bayes and Evidence-Based Medicine

Bayes' theorem. $P(B \mid T) = \frac{P(T|B) \times P(B)}{P(T)}$, with $P(T) = P(T \mid B) \times P(B) + P(T \mid \text{non} - B) \times P(\text{non} - B)$. Example: see the PPV formula below.

Post-test probability after a positive test (PPV from Se, Sp and prevalence). PPV $= \frac{\text{Se} \times P}{\text{Se} \times P + (1 - \text{Sp}) \times (1 - P)}$. Example: prevalence 10%, Se 80%, Sp 90%: 0.08/(0.08 + 0.09) = 0.471 (47%).

NPV from Se, Sp and prevalence. NPV $= \frac{\text{Sp} \times (1 - P)}{\text{Sp} \times (1 - P) + (1 - \text{Se}) \times P}$. Example: the same values, 0.81/(0.81 + 0.02) = 0.976.

Converting probability to odds. odds = p / (1 - p); p = odds / (1 + odds). Example: p = 0.181, odds = 0.181/0.819 = 0.221; odds of 0.666 correspond to $p = \frac{0.666}{1.666} = 0.40$.

Positive likelihood ratio. $\text{LR}^+ = \frac{\text{Se}}{1 - \text{Sp}}$. Rough interpretation: above 10, a large increase in the probability of disease; 5-10, moderate; 2-5, small; 1, the test provides no information. Example: FINDRISC ≥ 12, 0.759/0.252 = 3.01.

Negative likelihood ratio. $\text{LR}^- = \frac{1 - \text{Se}}{\text{Sp}}$. Below 0.1, a large decrease in the probability of disease. Example: 0.241/0.748 = 0.32.

Post-test odds. Post-test odds = pre-test odds x LR; then post-test probability = odds/(1 + odds). This is the practical form of Bayes' theorem, on which the Fagan nomogram is based [122]. Example: pre-test probability 18.1% (odds 0.221), positive test with LR+ 3.01: post-test odds 0.666, post-test probability 40.0%, identical to the PPV from the table.

D.12 Chapter 12. Diagnostic Tests and the ROC Curve

All the formulas below use the notation of Table D.2. The common example is the FINDRISC score with the cut-off ≥ 12 against diabetes confirmed by OGTT: a = 22, b = 7, c = 33, d = 98, N = 160.

Prevalence (in the study sample). $P = \frac{a + b}{N}$. Example: 29/160 = 18.1%.

Sensitivity. Se $= \frac{a}{a+b} = P(T^+ \mid B^+)$. Example: 22/29 = 75.9%.

Specificity. Sp $= \frac{d}{c+d} = P(T^- \mid B^-)$. Example: 98/131 = 74.8%.

Positive predictive value. PPV $= \frac{a}{a+c} = P(B^+ \mid T^+)$. It depends on prevalence. Example: 22/55 = 40.0%.

Negative predictive value. NPV $= \frac{d}{b+d} = P(B^- \mid T^-)$. Example: 98/105 = 93.3%.

Accuracy (proportion of correct classifications). Acc $= \frac{a+d}{N}$. Example: (22 + 98)/160 = 75.0%.

Misclassification rate. ER $= \frac{b+c}{N} = \frac{\text{FN+FP}}{N} = 1 - \text{Acc}$. Caution: the incorrect formula (b + d)/N circulates in some materials; the correct one is (b + c)/N. Example: (7 + 33)/160 = 25.0%.

Proportion of false positives and false negatives. False positive rate $= \frac{c}{c+d} = 1 - \text{Sp}$; false negative rate $= \frac{b}{a+b} = 1 - \text{Se}$. Example: 33/131 = 25.2%; 7/29 = 24.1%.

Youden index. J = Se + Sp - 1, between 0 (useless test) and 1 (perfect test). The optimal cut-off according to this criterion is the one that maximizes J [134]. Example: at the FINDRISC cut-off ≥ 11, J = 0.931 + 0.664 - 1 = 0.595; at the cut-off ≥ 12, J = 0.759 + 0.748 - 1 = 0.507.

Diagnostic odds ratio (DOR). DOR $= \frac{\text{LR}^+}{\text{LR}^-} = \frac{a \times d}{b \times c}$. Example: $\frac{22 \times 98}{7 \times 33} = 9.33$.

Area under the ROC curve (AUC), interpretation. AUC = the probability that a randomly chosen diseased person has a higher score than a randomly chosen non-diseased person; 0.5 = useless test, 1 = perfect test. Benchmarks used in the course: 0.91-1 excellent, 0.81-0.90 very good, 0.71-0.80 good, 0.61-0.70 satisfactory, below 0.60 poor. Example: FINDRISC, AUC = 0.852 (95% CI 0.762-0.943).

Reconstructing the 2 x 2 table from the indicators. If N, the prevalence P, Se and Sp are known: 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. If the PPV and a are known: $a + c = \frac{a}{\text{PPV}}$. If the NPV and b are known: $b + d = \frac{b}{1-\text{NPV}}$. Example: N = 1,000, P = 10%, Se = 80%, Sp = 90% give a = 80, b = 20, c = 90, d = 810, and PPV $= \frac{80}{170} = 47.1\%$.

Cohen's kappa coefficient (agreement between two raters). κ = (po - pe) / (1 - pe), where po is the proportion of observed agreement and pe = Σ(marginal proportion of rater 1 x marginal proportion of rater 2), summed over categories. Interpretation according to Landis and Koch [103]: 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 substantial, above 0.80 almost perfect. Example: two physicians read 100 ECGs; both say "abnormal" for 20, both say "normal" for 60, and they disagree on 20 (10 + 10). po = 0.80; each physician classified 30% of the tracings as abnormal, so pe = 0.3 x 0.3 + 0.7 x 0.7 = 0.58; κ = 0.22/0.42 = 0.52 (moderate agreement).

D.13 Chapter 13. Clinical Trials (Planning Formulas)

The planning formulas (sample size for two means or two proportions and the correction for dropout) are in Section D.4. In reporting a clinical trial, RR, RD and NNT (D.10) are used for a binary primary outcome, and the difference in means with its 95% CI (D.6) for a numerical outcome.

Dropout rate and response rate. Dropout rate = lost to follow-up / randomized; participation rate = randomized / assessed for eligibility. Example: in the mg2026 scenario, 160 randomized out of 212 assessed, participation rate = 75.5%.

Block randomization (number of possible sequences). For a block of 4 with 2 allocations to A and 2 allocations to B there are $\frac{4!}{2!\times 2!} = 6$ possible sequences (AABB, ABAB, ABBA, BAAB, BABA, BBAA).

Key Points

- In the diagnostic table with disease status in the rows: $\mathrm{Se} = \frac{a}{a+b}$, $\mathrm{Sp} = \frac{d}{c+d}$, $\mathrm{PPV} = \frac{a}{a+c}$, $\mathrm{NPV} = \frac{d}{b+d}$, error rate = (b + c)/N.
- In the risk table with exposure in the rows: $\mathrm{RR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$, $\mathrm{OR} = \frac{\mathrm{ad}}{\mathrm{bc}}$; the CI is calculated on the logarithmic scale, so it is asymmetric around the estimate.
- $\mathrm{SE} = \frac{s}{\sqrt{n}}$ describes the precision of the mean; s describes the spread of the patients. The interval $m \pm 1.96s$ is for individual values, $m \pm 1.96\mathrm{SE}$ is for the mean.
- $\mathrm{NNT} = \frac{1}{\mathrm{RD}}$, rounded up; the RR and the OR say nothing about the absolute size of the benefit.
- Sample sizes are always rounded up and corrected for dropout.

Appendix E. Statistical Tables

The tables in this appendix were calculated with the scipy library (Python), using the inverse cumulative distribution functions of the normal, Student, chi-square and Fisher distributions, and were not copied from other textbooks. Statistical software (JASP, MedCalc, Epi Info) gives the p value directly, so in everyday practice the tables are rarely needed. In the exam and in seminars, however, they remain useful: they force you to understand what a critical value means, why a two-sided test requires a larger threshold than a one-sided test, and how the critical value of t decreases as the number of degrees of freedom increases. All decimal numbers are written with a decimal point.

The general reading rule is the same for Tables E.3-E.7: calculate the test statistic from the data, find in the table the critical value for the degrees of freedom and the chosen α level, then compare. If the calculated value (in absolute value) is greater than or equal to the critical value, $p \leq \alpha$ and H0 is rejected. If it is smaller, H0 is not rejected.

E.1 The Standard Normal Distribution: The Cumulative Distribution Function Φ(z)

The table gives Φ(z), that is, the area under the standard normal curve to the left of the value z (the probability $P(Z \leq z)$). The row gives the integer part and the first decimal place of z, and the column gives the second decimal place. From Φ(z) you obtain the other areas you need:

- the area between 0 and z: Φ(z) - 0.5;
- the area in the right tail ($P(Z > z)$): 1 - Φ(z);
- the area between -z and +z: 2 x Φ(z) - 1;
- the two-sided p value for an observed z-score: 2 x (1 - Φ(|z|));
- for negative z: Φ(-z) = 1 - Φ(z), because the curve is symmetric.

Example. In the mg2026 dataset, the z-score of a systolic blood pressure of 160 mmHg is 2.08. In row 2.0, column 0.08, we read Φ(2.08) = 0.9812. Consequently, approximately 1 - 0.9812 = 0.0188, that is, 1.9% of patients would be expected to have an SBP above 160 mmHg if the distribution were exactly normal, and the area between the mean and 160 mmHg is 0.9812 - 0.5 = 0.4812.

Table E.1. The cumulative distribution function Φ(z) of the standard normal distribution, 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

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
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
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 Critical Values of the Standard Normal Distribution

The critical value z is the point beyond which lies an area equal to α (one-sided test) or two areas of α/2 each (two-sided test). The same values are used to construct confidence intervals: for a 95% CI the two-sided z for α = 0.05 is used, that is, 1.960; for 99%, 2.576; for 90%, 1.645.

Example. A two-sided z-test gives z = 2.30. The critical value for two-sided α = 0.05 is 1.960, and for two-sided α = 0.01 it is 2.576. Since 1.960 < 2.30 < 2.576, it follows that 0.01 < p < 0.05 (the exact value, from Table E.1, is 2 x (1 - 0.9893) = 0.0214).

Table E.2. Critical z values for common significance levels.

α	One-sided critical z (z1-α)	Two-sided critical z (z1-α/2)	Two-sided CI level (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 Critical Values of the Student (t) Distribution

The table gives the value of t for which the area in the two tails taken together equals α (two-sided test). For a one-sided test at level α, read the 2α column (for example, a one-sided test at α = 0.05 uses the 0.10 column). The degrees of freedom (df) are n - 1 for the one-sample t-test and the paired t-test, n1 + n2 - 2 for the classical independent-samples t-test and n - 2 for the test of the correlation coefficient. The ∞ row coincides with the critical z values in Table E.2. If the number

of degrees of freedom is missing from the table, use the row for the next lower number of degrees of freedom (the result is conservative).

Example. For the 95% CI of the mean SBP in a subsample of n = 10 patients, df = 9 and t(0.05; 9) = 2.262. With a mean of 130.0 and standard deviation s, CI = $130.0 \pm \frac{2.262 \times s}{\sqrt{10}}$. For n = 160 (df = 159), row 120 is used, t = 1.980, very close to 1.96.

Table E.3. Two-sided critical t values.

df	$\alpha = 0.20$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.02$	$\alpha = 0.01$	$\alpha = 0.001$
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 Critical Values of the Chi-Square (χ^2) Distribution

The chi-square test of association is always one-sided, on the right tail: large values of χ^2 indicate a large deviation of the observed frequencies from the expected ones. The table gives the value of χ^2 that leaves an area equal to α to its right. For a contingency table with r rows and c columns, df = (r - 1) x (c - 1); a 2 x 2 table has df = 1.

Example. In mg2026, the table of arm x weight loss of at least 5% has 3 rows and 2 columns, so df = 2, and the calculated $\chi^2 = 6.77$. The critical value for $\alpha = 0.05$ and df = 2 is 5.991, and for $\alpha = 0.01$ it is 9.210. Consequently $0.01 < p < 0.05$ (the software gives $p = 0.034$) and H0 is rejected at the 5% level.

Table E.4. Critical χ^2 values (area to the right = α).

df	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
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 Critical Values of the Fisher (F) Distribution for α = 0.05

The F distribution has two numbers of degrees of freedom. In one-way ANOVA, df1 (numerator) = k - 1, where k is the number of groups, and df2 (denominator) = N - k, where N is the total number of observations. The columns of the table correspond to df1, the rows to df2. If the calculated F is greater than or equal to the value in the table, $p \leq 0.05$. The same table is used, with df1 = the number of predictors and df2 = n - p - 1, for the overall F test of a multiple regression.

Example. ANOVA for the decrease in fasting plasma glucose across the three arms of mg2026: k = 3, N = 160, so df1 = 2 and df2 = 157. Row 157 does not exist in the table; we use row 120 (conservative): critical F = 3.072. The calculated value, F = 22.93, far exceeds the threshold, so $p < 0.05$ (the software gives $p < 0.001$).

Table E.5. Critical F values for α = 0.05 (df1 in the columns, df2 in the rows).

df2 df1	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 Critical Values of the Pearson Correlation Coefficient (r)

To decide whether an observed correlation coefficient differs significantly from zero, compare |r| with the critical value for the sample size n. The values come from the relationship between r and the t-test, with df = n - 2:

$$\text{critical } r = \text{critical } t \,/\, \sqrt{(df + \text{critical } t^2)}$$

The table is two-sided (H1: $\rho \neq 0$). Notice how much the threshold falls as n increases: with 10 patients you need $|r| \geq 0.632$ for $p < 0.05$, whereas with 100 patients $|r| \geq 0.197$ is enough. Hence the warning in Chapter 9: in large samples, weak correlations of no clinical importance become "significant".

Example. The correlation between BMI and SBP in mg2026 is r = 0.405, with n = 160. Even the threshold for n = 100 at $\alpha = 0.01$ (0.256) is exceeded, so $p < 0.01$. At n = 160 the threshold is even lower.

Table E.6. Two-sided critical values of Pearson's r.

n	$\mathrm{df} = n - 2$	$\alpha = 0.05$	$\alpha = 0.01$
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 Approximate Critical Values of the Spearman Correlation Coefficient (rho)

For very small samples, the distribution of rho under H0 is not continuous, and the critical values are obtained by enumerating all permutations of the ranks. In the table below, the values for n = 5-

10 are exact (calculated by complete enumeration of the n! permutations), while for n = 11-30 they are approximated by the formula based on the t distribution with n - 2 degrees of freedom, the same as for Pearson's r. A dash means that, for that n, no value of rho can reach the required threshold. The t approximation slightly underestimates the exact threshold for n between 11 and 15 (by approximately 0.01-0.02), so a rho that lies right at the limit must be judged by the exact p value. The table is two-sided; decision rule: |rho| ≥ the critical value means p ≤ α. For reporting, use the exact p value given by the software.

Example. A family physician correlates, in 12 patients, the physical activity score (1-5) with the number of consultations per year and obtains rho = -0.61. The critical value for n = 12 and α = 0.05 is 0.576; since |-0.61| ≥ 0.576, the negative correlation is significant at the 5% level, but not at the 1% level (the threshold is 0.708).

Table E.7. Two-sided critical values of Spearman's rho (exact for n ≤ 10, approximate for n ≥ 11).

n	$\alpha = 0.05$	$\alpha = 0.01$	Method
5	1,000	-	exact
6	0,886	1,000	exact
7	0,786	0,929	exact
8	0,738	0,881	exact
9	0,700	0,833	exact
10	0,648	0,794	exact
11	0,602	0,735	t approximation
12	0,576	0,708	t approximation
13	0,553	0,684	t approximation
14	0,532	0,661	t approximation
15	0,514	0,641	t approximation
16	0,497	0,623	t approximation
17	0,482	0,606	t approximation
18	0,468	0,590	t approximation
19	0,456	0,575	t approximation
20	0,444	0,561	t approximation
21	0,433	0,549	t approximation
22	0,423	0,537	t approximation
23	0,413	0,526	t approximation
24	0,404	0,515	t approximation
25	0,396	0,505	t approximation
26	0,388	0,496	t approximation
27	0,381	0,487	t approximation
28	0,374	0,479	t approximation
29	0,367	0,471	t approximation
30	0,361	0,463	t approximation

E.8 Table of 1,000 Random Digits

The digits were generated with the numpy generator (default_rng) initialized with the seed 2026, so the table can be reproduced at any time. They are arranged in 25 rows of 40 digits each, grouped in fives for ease of reading. The table is used in the exercises on simple random sampling and randomization in Chapters 1, 4 and 13.

How to use it. (1) Number the units in the population, for example the patient records from 001 to N. (2) Choose a starting point without looking at the digits (for example, close your eyes and put your pencil on the table) and a reading direction (from left to right, then the next row). (3)

Read groups of as many digits as N has (three digits for N = 160). (4) Keep the numbers between 001 and N, ignore those outside the range and those that are repeated, until you obtain the desired sample size.

Example. We want to draw 5 patients from the 160 in the mg2026 dataset, starting at the beginning of row 01 and reading three digits at a time. We keep only the numbers 001-160, in the order in which they appear, without repetitions. The result for this starting point is given below the table, so that you can check yourself.

Table E.8. 1,000 random digits (seed 2026).

Row	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

Check for the example above: reading three digits at a time from the beginning of the table, the first five distinct numbers between 001 and 160 are 036, 021, 121, 042, 034. If you started from another point in the table, your sample will be different and just as valid.

Appendix F. Practice Multiple-Choice Tests for the Exam

The theoretical Biostatistics exam is a 50-question multiple-choice test graded in the format of the Romanian residency examination. This appendix offers you three complete tests, built on the same model, so that you can practice under conditions as close as possible to the real ones. Each test covers all 14 chapters of the course, roughly in proportion to their length, and mixes questions on theory, interpretation and calculation.

How the Questions Are Built and How They Are Scored

The questions are of two types, marked at the beginning of the stem:

- **SC (single choice)**: of the five options (a-e), exactly one is correct.
- **MC (multiple choice)**: of the five options, at least two and at most four are correct.

The exact scoring rules (how many points each type of question is worth and how a partial MC answer is treated) are set by the university's methodology for examinations graded in the residency format and are announced by the discipline before the exam session; please read them there, not in this book. The principle you need to remember is simple: an answer is considered correct when it matches the answer key exactly. For SC questions, this means the correct option and only that option. For MC questions, the safe strategy is to mark all the correct options and no incorrect option; an incorrect option ticked "just in case" or a correct option left out can cost you the points for the question.

A few practical tips, from experience in grading these exams:

- Read the stem to the end. The difference between "interval for individual values" and "confidence interval for the mean", or between "valid percent" and "percent of total", often lies in a single word.
- In calculations, write out the complete 2 x 2 table (a, b, c, d and the totals) on your scratch paper before calculating any indicator. Most mistakes come from reading the wrong cell.
- Check the order of magnitude of the result: a probability cannot exceed 1, a sensitivity of 90% cannot give a PPV of 150%, and a sample size is rounded up.
- For MC questions, judge each option separately, as if it were a true/false question.
- For the calculations you need only simple operations and, sometimes, the tables in Appendix E.

Work through each test in a single sitting, without interruption and without looking at the answers. Time yourself. Only then check the answer key in the "Answers and Explanations" section at the end of the appendix, and read the explanations for the questions you got wrong. Explanations are given for all calculation questions and for those that contain a trap; for simple definition questions, reread the corresponding section of the chapter or the glossary (Appendix B). The formulas are collected in Appendix D.

The numbers marked as coming from the mg2026 dataset are those calculated in the chapters of the course on the simulated teaching dataset (Appendix C); the rest of the data in the stems were constructed for the exercise.

Test 1

1. (SC) Which of the following operations is an example of statistical inference?

a) Calculating the mean SBP of the 160 patients in the mg2026 dataset

b) Constructing the histogram of fasting plasma glucose for the patients in the sample

c) Counting the smokers in the sample

d) Estimating, on the basis of the sample, the interval that contains the mean SBP of all adults at cardiometabolic risk in the region

e) Drawing up the frequency table for level of education

2. (MC) Regarding sampling methods, the following statements are correct:

a) In simple random sampling, every individual in the population has the same probability of being selected

b) A convenience sample becomes representative if it includes more than 1,000 people

c) Stratified sampling guarantees that all strata of the population are present in the sample

d) Systematic selection error decreases as the sample size increases

e) In cluster sampling, whole groups, for example practices, are drawn at random

3. (SC) In a family practice, 30% of patients are smokers, 20% have hypertension and 8% are hypertensive smokers. The probability that a randomly chosen patient is a smoker or hypertensive is:

a) 0,50

b) 0,42

c) 0,06

d) 0,58

e) 0,38

4. (SC) The NYHA functional class of heart failure (I, II, III, IV) is a variable that is:

a) nominal

b) ratio

c) continuous

d) interval

e) ordinal

5. (MC) The lengths of hospital stay (days) of five patients are 4, 7, 7, 9, 13. The following statements are correct:

a) The mean is 8 days

b) The median is 7 days

c) The mode is 7 days

d) The range is 13 days

e) If the value 13 were replaced by 43, the median would increase

6. (SC) A variable has a mean of 80 and a standard deviation of 12. The coefficient of variation is:

a) 6,7%

b) 12%

c) 15%

d) 66,7%

e) 0,15%

7. (MC) Regarding the box plot, the following statements are correct:

a) The line inside the box represents the arithmetic mean

b) The edges of the box correspond to the first and third quartiles

c) The height of the box is the interquartile range

d) The points drawn separately, beyond the whiskers, are outliers according to the 1.5 x IQR rule

e) It is the recommended graph for nominal variables

8. (SC) In a population, total cholesterol is normally distributed with a mean of 200 mg/dL and a standard deviation of 40 mg/dL. The approximate percentage of people with cholesterol above 260 mg/dL is:

a) 6,7%

b) 2,3%

c) 13,4%

d) 15,9%

e) 43,3%

9. (MC) Regarding the normal distribution, the following statements are correct:

a) It is symmetric about the mean

b) The mean, median and mode coincide

c) Approximately 95.4% of the values lie within the interval $m \pm 2s$

d) The total area under the curve depends on the standard deviation

e) A Shapiro-Wilk test with p = 0.30 proves that the data are normal

10. (SC) The triglycerides in the mg2026 dataset have a skewness of 0.95 and a Shapiro-Wilk test $p < 0.001$. The most appropriate way to describe central tendency and spread, without transformation, is:

a) mean ± standard deviation

b) median and interquartile range

c) mode and range

d) mean and standard error

e) the coefficient of variation alone

11. (SC) In the mg2026 dataset, SBP has a mean of 131.13 mmHg and a standard deviation of 13.88 mmHg. The z-score of a value of 160 mmHg is:

a) 0,21

b) 1,96

c) 2,08

d) 2,88

e) 28,87

12. (SC) In a sample of 100 patients, the mean of a variable is 150 and the standard deviation is 20. The 95% confidence interval for the population mean (using z = 1.96) is:

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. (MC) Regarding the standard deviation (SD) and the standard error of the mean (SE), the following statements are correct:

a) $\mathrm{SE} = \frac{\mathrm{SD}}{\sqrt{n}}$

b) The SD describes the spread of the patients, while the SE describes the precision of the mean

c) If the sample size is multiplied by 4, the SE falls to approximately half

d) If the sample size increases, the SD decreases in proportion to $\sqrt{n}$

e) The interval $m \pm 1.96$SE contains approximately 95% of the patients

14. (SC) In a sample of 200 pregnant women, 40 have anemia. The 95% confidence interval for the prevalence of anemia, calculated by the Wald method, is approximately:

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. (SC) What sample size is needed to estimate a completely unknown prevalence with a margin of error of ±5% and a confidence level of 95%?

a) 100

b) 196

c) 246

d) 385

e) 1537

16. (SC) The p value of a statistical test represents:

a) the probability of obtaining a result at least as extreme as the one observed, if the null hypothesis were true

b) the probability that the null hypothesis is true

c) the probability that the alternative hypothesis is true

d) the probability of a type II error

e) the size of the treatment effect

17. (MC) Regarding statistical errors and statistical power, the following statements are correct:

a) A type I error consists in rejecting the null hypothesis when it is true

b) β is the probability of a type II error

c) The power of the test is 1 - α

d) Power increases with the sample size

e) Lowering the α level from 0.05 to 0.01, at the same sample size, increases power

18. (SC) In mg2026, the difference in SBP between men and women is -0.5 mmHg (95% CI from -4.9 to 3.8 mmHg), p = 0.81. The correct conclusion is:

a) it has been demonstrated that mean SBP is equal in the two sexes

b) the null hypothesis is accepted

c) the null hypothesis is not rejected; the data are compatible both with no difference and with differences of a few mmHg in either direction

d) men have a significantly lower SBP

e) the power of the study is 81%

19. (MC) For a two-sided z-test at α = 0.05, the following statements are correct:

a) The critical value is 1.96

b) The critical value is 1.645

c) A calculated z of 2.30 corresponds to $p < 0.05$

d) A calculated z of 2.30 corresponds to $p < 0.01$

e) If the 95% CI of the difference does not contain 0, then $p < 0.05$

20. (SC) SBP is measured in 16 patients before and after a treatment. The mean of the differences is 4 mmHg, and the standard deviation of the differences is 8 mmHg. The t statistic of the paired t-test is:

a) 0,5

b) 1,0

c) 8,0

d) 16,0

e) 2,0

21. (MC) For the situation in question 20 (paired t, n = 16), the following statements are correct:

a) The number of degrees of freedom is 15

b) The two-sided critical value of t for $\alpha = 0.05$ is 2.131

c) The null hypothesis is rejected at the two-sided 0.05 level

d) The two-sided p value lies between 0.05 and 0.10

e) In a one-sided test, with the direction specified before the study and $\alpha = 0.05$, the result would be significant

22. (SC) To compare the mean decrease in blood glucose between three independent arms of a trial, with normal data and homogeneous variances, we use:

a) the paired t-test

b) three independent t-tests, without any correction

c) one-way ANOVA

d) the chi-square test

e) the Friedman test

23. (MC) In mg2026, the ANOVA for the decrease in fasting plasma glucose at 6 months by arm gives $F(2, 157) = 22.93$, $p < 0.001$, $\eta^2 = 0.226$, and the Tukey test gives $p = 0.024$ for the C-B difference. The following statements are correct:

a) ANOVA alone shows which arms differ from one another

b) The degrees of freedom correspond to 3 groups and 160 patients

c) Approximately 22.6% of the variability of the decrease in blood glucose is explained by arm

d) The difference between arms C and B is significant at the 0.05 level

e) The result makes it mandatory to repeat the analysis with the Kruskal-Wallis test

24. (SC) The nonparametric alternative to the paired t-test is:

a) the Mann-Whitney test

b) the Wilcoxon signed-rank test

c) the Kruskal-Wallis test

d) the Friedman test

e) the chi-square test

25. (SC) A contingency table has 3 rows and 4 columns. The number of degrees of freedom for the chi-square test is:

a) 6

b) 7

c) 12

d) 5

e) 2

26. (SC) In a contingency table with N = 200, the total of one row is 60 and the total of one column is 50. The expected frequency in the cell at their intersection is:

a) 10

b) 12

c) 15

d) 25

e) 30

27. (MC) Regarding the chi-square test of association, the following statements are correct:

a) It is calculated on absolute frequencies

b) It can be applied directly to the percentages in the table

c) For a 2 x 2 table with expected frequencies below 5, Fisher's exact test is preferred

d) A large value of χ^2 indicates a large deviation of the observed frequencies from the expected ones

e) The value of χ^2 directly shows the direction and size of the association

28. (SC) Satisfaction with the intervention (Likert scale 1-5) is compared between arm C and arm A. The appropriate test is:

a) the independent-samples t-test

b) the Wilcoxon signed-rank test

c) the paired t-test

d) the Mann-Whitney test

e) the Pearson correlation coefficient

29. (MC) Which of the following situations require a test for paired (dependent) data?

a) Blood glucose in the same patients at enrollment and at 6 months

b) SBP in men compared with women

c) Visual acuity of the right eye compared with the left eye, in the same patients

d) The result (positive or negative) of two diagnostic tests applied to the same patients

e) The BMI of patients from three different practices

30. (SC) A student compares a numerical variable between two independent groups of 25 patients each; the Shapiro-Wilk test gives p = 0.003 in one of the groups, and the histogram shows strong skewness. The most appropriate choice is:

a) the classical t-test, without reservations

b) the Mann-Whitney test

c) the chi-square test

d) the Pearson coefficient

e) repeated measures ANOVA

31. (SC) The Pearson correlation coefficient between two variables is r = 0.6. The coefficient of determination is:

a) 0,06

b) 0,36

c) 0,60

d) 0,77

e) 1,20

32. (MC) In mg2026, the regression line of SBP on BMI is SBP = 89.47 + 1.435 x BMI (r = 0.405). The following statements are correct:

a) At a BMI of 30 kg/m², the predicted SBP is approximately 132.5 mmHg

b) SBP increases on average by 1.435 mmHg for each additional kg/m²

c) 89.47 mmHg is the actual SBP of a patient with a BMI equal to 0

d) BMI explains approximately 16.4% of the variability of SBP

e) The value of r demonstrates that obesity causes SBP to rise

33. (SC) The Spearman correlation coefficient is preferred to the Pearson coefficient when:

a) both variables are normal and the relationship is linear

b) the sample size exceeds 1,000

c) both variables are nominal

d) a prediction equation is wanted

e) at least one of the variables is ordinal, or the relationship is monotonic but not linear

34. (MC) A study of 210 patients obtains r = 0.15, p = 0.03. The following statements are correct:

a) The correlation is statistically significant at the 0.05 level

b) The strength of the association is very small, although the result is significant

c) Variable x explains approximately 2.25% of the variability of y

d) The result demonstrates a causal relationship

e) A small coefficient can become significant in a large sample

35. (SC) In a cohort study, 30 of 100 exposed and 10 of 100 unexposed people develop the disease. The relative risk is:

a) 2,0

b) 3,0

c) 3,86

d) 0,33

e) 20

36. (MC) For the data in question 35, the following statements are correct:

a) The OR is approximately 3.86

b) The risk difference is 20 percentage points

c) The number needed to produce one additional case (NNH) is 5

d) The OR is smaller than the RR

e) The attributable fraction among the exposed is approximately 66.7%

37. (SC) The measure of association that can be calculated correctly in a case-control study is:

a) the odds ratio (OR)

b) the relative risk (RR)

c) cumulative incidence

d) the prevalence of the disease in the population

e) the number needed to treat

38. (SC) In a randomized trial, the adverse event occurs in 20% of the patients in the control group and in 15% of those treated. The number needed to treat (NNT) is:

a) 5

b) 15

c) 20

d) 25

e) 33

39. (MC) Regarding confounders, the following statements are correct:

a) It is associated with the exposure

b) It is associated with the outcome (the disease)

c) It is an intermediate link on the causal pathway between exposure and disease

d) Its influence can be controlled by randomization

e) It is eliminated automatically by increasing the sample size

40. (SC) A disease has a prevalence of 2%. A test has a sensitivity of 95% and a specificity of 90%. The probability that a person with a positive test has the disease is approximately:

a) 95%

b) 90%

c) 2%

d) 16%

e) 50%

41. (MC) The pre-test probability of a disease is 10%, and the test has a positive likelihood ratio LR+ = 9. The following statements are correct:

a) The pre-test odds are approximately 0.11

b) After a positive test, the post-test odds are 1.0

c) The post-test probability after a positive test is 50%

d) The post-test probability after a positive test is 90%

e) An LR+ of 9 corresponds, for example, to a test with Se 90% and Sp 90%

42. (SC) In formulating a clinical question according to the PICO model, the letter O stands for:

a) the observation

b) the outcome being studied

c) the objective of the study

d) the odds

e) blinding

43. (MC) A test is applied to 50 diseased and 150 non-diseased people: 45 of the diseased have a positive test, and 130 of the non-diseased have a negative test. The following statements are correct:

a) The sensitivity is 90%

b) The specificity is approximately 86.7%

c) The positive predictive value is approximately 69.2%

d) The misclassification rate is (5 + 130)/200

e) The accuracy of the test is 75%

44. (SC) For the test in question 43, the misclassification rate is:

a) 2,5%

b) 10%

c) 13,3%

d) 67,5%

e) 12,5%

45. (SC) In a population of 200 people, the prevalence of the disease is 30%. The test has a sensitivity of 80% and a specificity of 90%. The positive predictive value is approximately:

a) 30,0%

b) 77,4%

c) 80,0%

d) 90,0%

e) 91,3%

46. (MC) Regarding the ROC curve, the following statements are correct:

a) Sensitivity is plotted on the vertical axis

b) 1 - specificity is plotted on the horizontal axis

c) An area under the curve of 0.5 corresponds to a test with no discriminating value

d) An area of 0.852 means that 85.2% of patients are correctly classified at any cut-off

e) The cut-off that maximizes the Youden index is always the one with the maximum sensitivity

47. (SC) A test has a sensitivity of 85% and a specificity of 75%. The Youden index is:

a) 0,10

b) 0,40

c) 0,60

d) 0,64

e) 1,60

48. (SC) In a randomized superiority trial, the recommended primary analysis is:

a) the intention-to-treat analysis

b) the per-protocol analysis

c) the analysis of only the patients who completed the study

d) the analysis of a subgroup chosen after seeing the data

e) the interim analysis

49. (MC) Regarding randomized controlled trials, the following statements are correct:

a) Randomization tends to balance unknown confounders between the arms as well

b) Allocation concealment and blinding are the same thing

c) In a double-blind trial, neither the patient nor the investigator knows the treatment received

d) The primary outcome can be chosen after the data have been analyzed

e) The CONSORT guideline requires a participant flow diagram

50. (SC) For a cross-sectional study of 80 patients, an article reports "OR = 2.1; p = 0.04", with no other details. The missing information that is most important for interpretation is:

a) the name of the statistical software

b) the exact value of the α level

c) the mean age of the patients

d) the confidence interval of the OR

e) the number of authors

Test 2

1. (SC) The mean systolic blood pressure of all adults in Romania, an unknown value that we are trying to estimate, is:

a) a parameter

b) a sample statistic

c) a sample

d) a qualitative variable

e) a standard error

2. (MC) Regarding data protection and research ethics, the following statements are correct:

a) Pseudonymized data remain personal data

b) Informed consent can be withdrawn by the participant at any time

c) Anonymization means replacing the name with a code while keeping a correspondence list

d) Health data belong to the special categories of data protected by the GDPR

e) A simulated dataset with fictitious patients falls under the GDPR like any clinical dataset

3. (SC) Events A and B are independent, with P(A) = 0.3 and P(B) = 0.4. The probability that both occur is:

a) 0,70

b) 0,12

c) 0,10

d) 0,58

e) 0,75

4. (SC) The number of consultations per year in six patients is 3, 5, 8, 10, 12, 20. The median is:

a) 8

b) 9,67

c) 9

d) 10

e) 11,5

5. (MC) A questionnaire was completed by 200 people. To the question "Do you smoke?", 20 people did not respond and 54 answered "yes". The following statements are correct:

a) The valid percentage of smokers is 30%

b) The percentage of smokers relative to the total number of respondents is 27%

c) The valid percentage is calculated relative to 200

d) The report must state the number of missing values

e) Missing responses are replaced with "no", so as not to lose data

6. (SC) For a variable, Q1 = 20 and Q3 = 30. According to Tukey's rule used for the box plot, the upper limit above which a value is considered an outlier is:

a) 35

b) 40

c) 50

d) 45

e) 60

7. (SC) The appropriate graph for showing the shape of the distribution of fasting plasma glucose in 160 patients is:

a) the pie chart

b) the bar chart

c) the scatter plot

d) the forest plot

e) the histogram

8. (SC) In healthy women in a population, hemoglobin is normally distributed with a mean of 14 g/dL and a standard deviation of 1.2 g/dL. The approximate percentage of women with hemoglobin below 11.6 g/dL is:

a) 0,13%

b) 2,3%

c) 5%

d) 15,9%

e) 47,7%

9. (MC) A variable is normally distributed with a mean of 120 and a standard deviation of 10. The following statements are correct:

a) Approximately 95% of individuals have values between 100.4 and 139.6

b) Approximately 68.3% of individuals have values between 110 and 130

c) Approximately 99.7% of individuals have values between 90 and 150

d) The interval 100.4-139.6 is the 95% confidence interval for the population mean

e) Approximately 2.5% of individuals have values below 100.4

10. (SC) After a variable has been analyzed on the logarithmic scale, the mean of the logarithms brought back to the original scale by exponentiation is:

a) the arithmetic mean

b) the median

c) the geometric mean

d) the mode

e) the standard deviation

11. (SC) The central limit theorem states that:

a) the means of sufficiently large samples have an approximately normal distribution, whatever the distribution of the variable in the population

b) all biological variables are normally distributed

c) the median and the mean are always equal in large samples

d) the standard deviation decreases as the sample size increases

e) any test becomes significant if the sample exceeds 30 people

12. (SC) In 16 patients, the mean of a variable is 50 and the standard deviation is 8. Given that $t(0.05; 15) = 2.131$, the 95% confidence interval for the population mean is:

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. (SC) What sample size is needed to estimate the mean of a variable with $\sigma = 10$, with a margin of error of ±2 units and a confidence level of 95%?

a) 10

b) 49

c) 96

d) 97

e) 385

14. (MC) In mg2026, the 95% confidence interval for the mean SBP is 128.96-133.30 mmHg. The following statements are correct:

a) If the study were repeated many times, approximately 95% of the intervals constructed in the same way would contain the true mean

b) 95% of patients have an SBP between 128.96 and 133.30 mmHg

c) The 99% confidence interval, calculated on the same data, is wider

d) With a sample 4 times larger, the interval would be approximately half as wide

e) The interval is centered on the sample mean

15. (SC) In a pilot study, 2 of 25 patients have an adverse reaction. The Wald confidence interval for the proportion has a negative lower limit. The correct approach is:

a) report the negative limit as it came out

b) replace the negative limit with 0 and keep the Wald upper limit

c) abandon the interval and report only the proportion

d) artificially increase the number of cases

e) use the Wilson interval, which is suitable for small samples and extreme proportions

16. (MC) Regarding statistical hypotheses, the following statements are correct:

a) For comparing two means, H0 is $\mu 1 = \mu 2$

b) The two-sided alternative hypothesis is $\mu 1 \neq \mu 2$

c) When $p > 0.05$, the null hypothesis is accepted as demonstrated

d) For a relative risk, H0 is $RR = 1$

e) For a correlation coefficient, H0 is $\rho = 1$

17. (SC) A total of 5 comparisons are made on the same data and the Bonferroni correction is applied for an overall α of 0.05. The threshold used for each comparison is:

a) 0,05

b) 0,025

c) 0,01

d) 0,005

e) 0,25

18. (SC) In a study of 20,000 patients, a drug lowers SBP by 1.2 mmHg compared with placebo (95% CI 0.8-1.6 mmHg), $p < 0.001$. The most appropriate interpretation is:

a) the effect is statistically significant, but probably of no clinical importance

b) the effect is clinically important, because p is very small

c) the result is not statistically significant

d) the confidence interval contains 0

e) the study was underpowered

19. (MC) The power of a statistical test increases when:

a) the sample size increases

b) the size of the effect sought increases

c) the variability (standard deviation) of the variable increases

d) the α level increases from 0.01 to 0.05

e) a paired design is replaced by one with independent groups, with the same number of measurements

20. (SC) Two groups of 25 and 30 patients are compared with the classical independent-samples t-test. The number of degrees of freedom is:

a) 55

b) 54

c) 53

d) 29

e) 24

21. (MC) In mg2026, fasting plasma glucose at 0 and 6 months is compared with the paired t-test: mean of the differences 10.19 mg/dL, standard deviation of the differences 11.57, n = 160, t = 11.14, 95% CI 8.38-11.99. The following statements are correct:

a) The standard error of the mean difference is approximately 0.915

b) The number of degrees of freedom is 159

c) The confidence interval does not contain 0, in agreement with $p < 0.05$

d) The test assumes equal variances of blood glucose at the two time points

e) An effect size dz = 0.88 indicates a large effect

22. (SC) Two independent groups with normal data have different variances (Levene test $p = 0.002$). The recommended test for comparing the means is:

a) the paired t-test

b) the chi-square test

c) the Friedman test

d) the Welch t-test

e) the McNemar test

23. (SC) In a one-way ANOVA, the between-group sum of squares is 300 and the total sum of squares is 1200. Eta squared is:

a) 0,25

b) 0,33

c) 0,75

d) 4,00

e) 0,0025

24. (SC) Two groups have means of 136 and 130 mmHg, and the pooled standard deviation is 12 mmHg. The effect size Cohen's d and its interpretation are:

a) 2.0, large effect

b) 0.5, small effect

c) 72, very large effect

d) 0.05, negligible effect

e) 0.5, medium effect

25. (MC) In a 2 x 2 table, a = 20, b = 30, c = 10, d = 40 (N = 100). Given that the critical χ^2 values for df = 1 are 3.841 ($\alpha = 0.05$) and 6.635 ($\alpha = 0.01$), the following statements are correct:

a) χ^2 is approximately 4.76

b) The number of degrees of freedom is 1

c) $p < 0.05$

d) $p < 0.01$

e) The expected frequency for cell a is 15

26. (SC) Two screening tests (A and B) are applied to the same 100 patients, and the result of each is positive or negative. To compare the proportions of positive results we use:

a) the chi-square test of independence

b) the McNemar test

c) the Mann-Whitney test

d) the independent-samples t-test

e) the Kruskal-Wallis test

27. (SC) Group A has the values 12, 15, 18, and group B the values 14, 20, 22, 25. The U statistic of the Mann-Whitney test (the smaller of U1 and U2) is:

a) 2

b) 8

c) 10

d) 12

e) 20

28. (MC) Regarding the Kruskal-Wallis and Friedman tests, the following statements are correct:

a) The Kruskal-Wallis test is the nonparametric alternative to one-way ANOVA

b) The Friedman test is used for repeated measurements on the same individuals

c) A significant Kruskal-Wallis test shows exactly which groups differ from one another

d) The Kruskal-Wallis test is based on ranks

e) The Friedman test is applied to independent groups

29. (SC) To study the association between smoking (yes/no) and COPD (yes/no) in 300 patients, with all expected frequencies above 5, we use:

a) the Pearson coefficient

b) the independent-samples t-test

c) ANOVA

d) the chi-square test

e) the Wilcoxon test

30. (MC) The following pairings of situation and test are correct:

a) three measurement time points, the same patients, normal data: repeated measures ANOVA

b) two independent groups, ordinal variable: the Mann-Whitney test

c) two nominal variables: the chi-square test

d) two normal numerical variables, linear relationship: the Spearman coefficient is mandatory

e) three independent groups, normal data: the Friedman test

31. (SC) In 6 patients, the sum of the squared rank differences between two variables is $\sum d^2 = 14$. The Spearman coefficient is:

a) 0,4

b) 0,6

c) 0,8

d) 0,93

e) -0,6

32. (MC) In 27 patients, r = 0.5 is obtained between two variables. The critical t value for df = 25 and two-sided $\alpha = 0.05$ is 2.060. The following statements are correct:

a) The number of degrees of freedom is 25

b) The t statistic is approximately 2.89

c) The correlation is significant at the 0.05 level

d) The coefficient of determination is 0.25

e) The correlation is perfect

33. (SC) In mg2026, SBP = 108.3 + 0.406 x age. The predicted SBP for a 60-year-old patient is approximately:

a) 132.7 mmHg

b) 24.4 mmHg

c) 108.7 mmHg

d) 168.3 mmHg

e) 141.8 mmHg

34. (SC) In the multiple regression SBP ~ BMI + age + sex + smoking in mg2026, the BMI coefficient is 1.389 ($p < 0.001$). The correct interpretation is:

a) BMI explains 1.389% of the variability of SBP

b) the correlation between BMI and SBP is 1.389

c) SBP increases on average by 1.389 mmHg for each additional kg/m², at constant age, sex and smoking status

d) a patient with a BMI of 0 has an SBP of 1.389 mmHg

e) BMI increases by 1.389 kg/m² for each mmHg of SBP

35. (SC) In a case-control study, 40 of 100 cases and 20 of 100 controls were exposed. The odds ratio is approximately:

a) 2,0

b) 1,5

c) 0,375

d) 20

e) 2,67

36. (MC) A study reports RR = 1.8 (95% CI 0.9-3.6). The following statements are correct:

a) The association is not statistically significant at the 0.05 level

b) The data are compatible with no effect

c) The data are also compatible with a doubling of the risk

d) It has been demonstrated that the exposure has no effect

e) The p value is below 0.05

37. (SC) In a cohort of 2,000 people without diabetes followed for 2 years, 50 new cases of diabetes occur. The cumulative incidence is:

a) 2.5% over 2 years

b) 1.25% over 2 years

c) 25% over 2 years

d) 50 per 1,000 person-years

e) 2.5 per 1,000 person-years

38. (SC) A total of 120 patients with lung cancer and 240 people without cancer are selected, and the frequency of past smoking is compared. The type of study is:

a) prospective cohort study

b) randomized controlled trial

c) ecological study

d) case-control study

e) case series

39. (MC) In mg2026 (cross-sectional analysis at enrollment), 18 of 54 obese and 11 of 106 non-obese patients have diabetes. The following statements are correct:

a) The prevalence of diabetes in obese patients is 33.3%

b) The prevalence ratio is approximately 3.21

c) The OR is approximately 4.32

d) The study demonstrates that obesity preceded the onset of diabetes

e) The OR is larger than the prevalence ratio because the disease is not rare in this sample

40. (SC) A test has a positive likelihood ratio LR+ = 8. The pre-test probability of the disease is 20%. The post-test probability after a positive result is approximately:

a) 20%

b) 66,7%

c) 80%

d) 88,9%

e) 160%

41. (MC) Regarding evidence-based medicine, the following statements are correct:

a) Systematic reviews with meta-analysis of high-quality randomized trials are at the top of the hierarchy of evidence

b) A case series can generate hypotheses, but it cannot test them

c) Expert opinion is the strongest form of evidence

d) Evidence-based medicine integrates research evidence with clinical expertise and patient values

e) A large observational study is always superior to a randomized trial

42. (SC) The same diagnostic test (unchanged sensitivity and specificity) is used in a population in which the prevalence of the disease is lower. What happens to the predictive values?

a) PPV increases, NPV decreases

b) both remain unchanged

c) PPV decreases, NPV increases

d) both decrease

e) both increase

43. (MC) The following is known about a diagnostic test: 40 true positive results, PPV = 0.8, 10 false negative results and NPV = 0.9. The following statements are correct:

a) The number of false positive results is 10

b) The number of true negative results is 90

c) The sensitivity is 80%

d) The specificity is 90%

e) The total number of people tested is 200

44. (SC) Two radiologists classify the same radiographs; the observed agreement is 0.85, and the agreement expected by chance is 0.50. The kappa coefficient is:

a) 0,35

b) 0,50

c) 0,70

d) 0,85

e) 1,70

45. (SC) A test with very high sensitivity gives a negative result in a patient. The correct interpretation (the SnNOut rule) is:

a) the disease is practically ruled out

b) the disease is confirmed

c) the result provides no information

d) the specificity must be calculated in order to decide

e) the patient has the disease at an early stage

46. (SC) For the FINDRISC score in mg2026, the cut-off ≥ 12 has Se 75.9% and Sp 74.8%, and the cut-off ≥ 16 has Se 41.4% and Sp 94.7%. Raising the cut-off from 12 to 16 has the effect of:

a) increasing both sensitivity and specificity

b) decreasing both

c) increasing sensitivity and decreasing specificity

d) decreasing sensitivity and increasing specificity

e) changing the area under the ROC curve

47. (MC) Regarding screening tests and the combination of tests, the following statements are correct:

a) For a screening test, high sensitivity is preferred

b) Positive screening results must be confirmed with a specific test

c) In the general population, the PPV of a test is usually lower than in a specialist clinic

d) Serial testing increases sensitivity

e) Parallel testing increases sensitivity

48. (SC) The calculation shows that 64 patients per group are needed. A dropout rate of 20% is expected. How many patients must be recruited per group?

a) 51

b) 77

c) 80

d) 84

e) 128

49. (MC) Regarding the study protocol and data collection forms, the following statements are correct:

a) The protocol is written and approved before data collection

b) Validation rules for responses in an electronic form reduce data entry errors

c) A well-designed form asks for SBP as free text, with the unit of measurement written by the user

d) The statistical analysis plan is established before the data are analyzed

e) Clinical trials are registered in public registries after the results have been published

50. (SC) JASP displays "p <.001" for a test, and a student writes "p = 0.000" in a report. The correct wording in the report is:

a) $p = 0$

b) $p = 0.000$

c) p is zero, so H0 is false with certainty

d) $p = 0.001$

e) $p < 0.001$

Test 3

1. (SC) The first stage of a statistical study is:

a) the statistical analysis of the data

b) data collection

c) formulating the research question and planning the study

d) cleaning the data file

e) choosing the statistical software

2. (MC) Regarding errors in research, the following statements are correct:

a) Random error decreases as the sample size increases

b) Systematic error (bias) decreases as the sample size increases

c) Recall bias is characteristic of case-control studies

d) A sample made up of volunteers can introduce selection bias

e) The confidence interval quantifies the size of the bias

3. (MC) For two events, P(A) = 0.3, P(B) = 0.4 and P(A and B) = 0.12. The following statements are correct:

a) A and B are independent

b) $P(A \text{ or } B) = 0.58$

c) $P(A \mid B) = 0.3$

d) A and B are mutually exclusive

e) $P(\text{non} - A) = 0.7$

4. (SC) The antibody titers of three patients are 10, 100 and 1000. The geometric mean is:

a) 370

b) 100

c) 1000

d) 30

e) 10

5. (SC) Variable A has a mean of 50 and a standard deviation of 5; variable B has a mean of 200 and a standard deviation of 30. Which statement is correct?

a) A has the greater relative variability, because its mean is smaller

b) The two variables have the same relative variability

c) Relative variability cannot be compared

d) B has the smaller absolute variability

e) B has the greater relative variability (CV 15% vs 10%)

6. (MC) The following pairings of variable and variable type are correct:

a) Blood glucose (mg/dL): continuous numerical

b) Number of pregnancies: discrete numerical

c) Blood group: ordinal

d) Satisfaction score 1-5: ordinal

e) Temperature in degrees Celsius: ratio scale

7. (SC) The measure of central tendency least influenced by extreme values is:

a) the median

b) the arithmetic mean

c) the weighted mean

d) the range

e) the standard deviation

8. (SC) On the Q-Q plot of a variable, the points lie approximately on the line. This suggests that:

a) the variable is ordinal

b) there are many outliers

c) the distribution is strongly skewed

d) the distribution is close to normal

e) the variables are correlated

9. (MC) In mg2026, BMI has a mean of 29.04, a median of 28.25, a skewness of 1.42 and a Shapiro-Wilk test $p < 0.001$. The following statements are correct:

a) The distribution is skewed to the right

b) The mean exceeds the median, in agreement with the positive skewness

c) The Shapiro-Wilk test indicates a departure from normality

d) The interval $m \pm 1.96s$ contains exactly 95% of the patients

e) Nonparametric tests may be preferred for comparing BMI between groups

10. (SC) The z-score value that leaves 2.5% of the area under the standard normal curve to its right is:

a) 1,645

b) 1,96

c) 2,33

d) 2,576

e) 3,00

11. (SC) A child's weight is at the 90th percentile for age and sex. This means that:

a) the child weighs 90% more than the mean weight

b) the child is obese

c) 10% of children weigh less than this child

d) the child's weight is 90 times the mean

e) approximately 90% of children of the same age and sex weigh less

12. (SC) In a sample of 60 patients, 18 have uncontrolled hypertension. The 95% confidence interval for the proportion (Wald method) is approximately:

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. (MC) The sample size needed to estimate a parameter increases when:

a) the desired margin of error decreases

b) the confidence level increases from 95% to 99%

c) the variability of the variable (σ) is greater

d) the estimated proportion approaches 0.5

e) the population from which the sample is drawn increases from 1 million to 2 million people

14. (SC) The sample size increases from 40 to 160 patients, and the standard deviation remains approximately the same. The width of the confidence interval for the mean becomes approximately:

a) a quarter as wide

b) unchanged

c) half as wide

d) twice as wide

e) four times as wide

15. (SC) How many patients are needed to estimate an assumed prevalence of 10%, with a margin of error of ±3% and a confidence level of 95%?

a) 100

b) 196

c) 400

d) 385

e) 1000

16. (SC) In 64 patients, the mean of a variable is 128, the standard deviation is 16, and the reference value is 120. The t statistic of the one-sample t-test is:

a) 0,5

b) 2

c) 4

d) 8

e) 32

17. (MC) Regarding the p value, the following statements are correct:

a) $p = 0.03$ means that the probability that H0 is true is 3%

b) A very small p can also occur for a clinically negligible effect if the sample is large

c) A $p > 0.05$ does not prove the absence of an effect

d) For the same effect, the p value depends on the sample size

e) The values $p = 0.049$ and $p = 0.051$ represent radically different evidence

18. (SC) A one-sided test is justified when:

a) the direction of the effect was specified before the study, and an effect in the opposite direction is impossible or of no interest

b) the two-sided test gave $p = 0.08$

c) the researcher wants a significant result

d) the sample is small

e) the data are not normal

19. (SC) A total of 20 independent tests are performed, and H0 is true for all of them. At $\alpha = 0.05$, the probability of obtaining at least one falsely significant result is approximately:

a) 0,05

b) 0,20

c) 0,36

d) 0,64

e) 1,00

20. (SC) In an ANOVA with three groups and 60 patients, the between-group mean square is 150 and the residual mean square is 30. The F statistic is:

a) 0,2

b) 5

c) 4,5

d) 120

e) 180

21. (MC) For the ANOVA in question 20, given that the critical F for df (2, 57) and α = 0.05 is approximately 3.16, the following statements are correct:

a) The number of between-group degrees of freedom is 2

b) The number of residual degrees of freedom is 57

c) $p < 0.05$

d) It has been demonstrated that all three means differ from one another

e) The next step is a post hoc test, for example Tukey

22. (SC) The role of the Tukey test, applied after a significant ANOVA, is to:

a) check the normality of the data

b) check the equality of variances

c) replace ANOVA when the data are ordinal

d) calculate the overall effect size

e) identify the pairs of groups that differ, while controlling the type I error

23. (SC) The weight of 30 patients is measured at the beginning and at the end of a weight-loss program; the differences are normally distributed. The appropriate test is:

a) the paired t-test

b) the independent-samples t-test

c) the Mann-Whitney test

d) the chi-square test

e) one-way ANOVA

24. (MC) The conditions of the classical (Student) independent-samples t-test are:

a) independent observations

b) an approximately normal distribution in each group (or large samples)

c) equal variances in the two groups

d) equal sizes of the two groups

e) an ordinal outcome variable

25. (SC) For a contingency table with 2 rows and 3 columns, the number of degrees of freedom of the chi-square test is:

a) 1

b) 2

c) 3

d) 5

e) 6

26. (SC) In a Mann-Whitney test, group 1 has 4 patients, group 2 has 5 patients, and the sum of the ranks in group 1 is 12. The value of $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$ is:

a) 2

b) 12

c) 18

d) 20

e) 10

27. (MC) In mg2026, weight loss of at least 5% occurred in 6/53 patients in arm A, 16/53 in arm B and 16/54 in arm C; $\chi^2 = 6.77$, df = 2, p = 0.034, and the smallest expected frequency is 12.6. The following statements are correct:

a) The condition on expected frequencies is met

b) The proportion of patients with a loss of at least 5% differs significantly between arms

c) The test shows that arm C differs significantly from arm B

d) The degrees of freedom are calculated as (3 - 1) x (2 - 1)

e) Fisher's exact test was mandatory

28. (SC) Blood glucose was measured in the same patients at 0, 3 and 6 months, and the distribution is strongly skewed in a small sample. The appropriate test is:

a) one-way ANOVA

b) the Kruskal-Wallis test

c) the chi-square test

d) the Friedman test

e) the Mann-Whitney test

29. (SC) HbA1c (normally distributed in both groups) is compared between patients with and without confirmed diabetes; the Levene test gives p = 0.40. The appropriate test is:

a) the paired t-test

b) the Wilcoxon test

c) the independent-samples t-test (classical or Welch version)

d) the chi-square test

e) the Friedman test

30. (MC) The stage of liver fibrosis (F0-F4) is an ordinal variable. Appropriate methods are:

a) the Mann-Whitney test for comparing two groups

b) the Kruskal-Wallis test for comparing three groups

c) the Spearman coefficient for the association with a numerical variable

d) the independent-samples t-test applied to the codes 0-4, as if it were a continuous normal variable

e) the chi-square test for the stage x group table, accepting that the information about order is lost

31. (SC) In a simple linear regression, $R^2 = 0.25$, and the slope of the line is negative. The Pearson correlation coefficient is:

a) 0,25

b) -0,25

c) 0,5

d) -0,5

e) 0,0625

32. (MC) Regarding linear regression, the following statements are correct:

a) The residual is the difference between the observed value and the predicted value

b) The least squares method minimizes the sum of squared residuals

c) The standard error of the estimate is expressed in the units of the dependent variable

d) Prediction for values of x far outside the observed range is just as reliable

e) A large R^2 proves a causal relationship

33. (SC) In a logistic regression, the coefficient for smoking is b = 0.693. The corresponding odds ratio is approximately:

a) 0,693

b) 2,0

c) 1,693

d) 0,5

e) 6,93

34. (SC) On a scatter plot, a single patient has very high values on both axes, far from the others. This point:

a) cannot influence the correlation coefficient

b) always lowers the correlation coefficient

c) automatically turns the correlation into a nonlinear one

d) must be deleted without checking

e) can change the correlation coefficient substantially and must be checked

35. (SC) In a randomized trial, the event occurs in 8 of 200 treated patients and in 16 of 200 patients on placebo. The relative risk is:

a) 0,5

b) 2,0

c) 0,04

d) 0,08

e) 0,96

36. (MC) For the data in question 35, the following statements are correct:

a) The absolute risk reduction is 4 percentage points

b) The relative risk reduction is 50%

c) The NNT is 25

d) The OR is exactly equal to the RR

e) Reporting only the relative risk reduction can exaggerate the impression of benefit

37. (SC) An exposure has RR = 3, and 25% of the population is exposed. The population attributable fraction (Levin's formula) is approximately:

a) 0,25

b) 0,333

c) 0,50

d) 0,667

e) 0,75

38. (SC) A chronic disease has a low incidence, but patients survive for many years. In this situation:

a) prevalence is always equal to incidence

b) prevalence is lower than annual incidence

c) prevalence is relatively high compared with annual incidence

d) incidence cannot be calculated

e) the case fatality rate is high

39. (MC) Regarding the types of epidemiological studies, the following statements are correct:

a) The cohort study allows incidence to be calculated

b) The case-control study is efficient for rare diseases

c) The case-control study allows the relative risk to be calculated directly

d) A prospective cohort study can be costly and lengthy

e) The cross-sectional study establishes the temporal sequence of exposure and disease

40. (SC) A test has a sensitivity of 90% and a specificity of 95%. In a population with a prevalence of 1%, the PPV is approximately:

a) 15,4%

b) 90%

c) 95%

d) 1%

e) 50%

41. (MC) Regarding likelihood ratios, the following statements are correct:

a) $\mathrm{LR}^{-} = \frac{1-\mathrm{Se}}{\mathrm{Sp}}$

b) For Se = 80% and Sp = 90%, the LR- is approximately 0.22

c) An LR equal to 1 does not change the probability of disease

d) The Fagan nomogram links the pre-test probability, the likelihood ratio and the post-test probability

e) The LR+ depends on the prevalence of the disease in the tested population

42. (SC) The odds of an event are 3. The probability of the event is:

a) 0,25

b) 0,33

c) 0,75

d) 3,00

e) 0,67

43. (SC) A diagnostic test gives: a = 80 true positives, b = 20 false negatives, c = 10 false positives, d = 90 true negatives. The diagnostic odds ratio (DOR) is:

a) 4

b) 8

c) 9

d) 36

e) 72

44. (MC) For the test in question 43, the following statements are correct:

a) The sensitivity is 80%

b) The specificity is 90%

c) The misclassification rate is (20 + 10)/200 = 15%

d) The misclassification rate is (20 + 90)/200

e) The accuracy is 85%

45. (SC) In the scenario of the mg2026 dataset, the reference standard for the diagnosis of diabetes, against which the FINDRISC score is evaluated, is:

a) fasting plasma glucose

b) HbA1c

c) the FINDRISC score itself

d) the patient's own statement

e) the oral glucose tolerance test

46. (MC) In mg2026, the area under the ROC curve is 0.852 (95% CI 0.762-0.943) for the FINDRISC score and 0.844 for fasting plasma glucose. The following statements are correct:

a) FINDRISC has a very good discriminating ability

b) The CI of the area does not contain 0.5, so the score discriminates better than chance

c) The difference between 0.852 and 0.844 demonstrates the superiority of FINDRISC

d) The area under the curve does not depend on the chosen cut-off

e) The area is the probability that a randomly chosen diseased person has a higher score than a randomly chosen non-diseased person

47. (SC) In an accuracy study, biopsy (the reference standard) is performed only in patients with a positive test. The error introduced is called:

a) recall bias

b) verification bias

c) type II error

d) confounding factor

e) publication bias

48. (SC) The phase of a clinical trial in which the drug is compared, in a randomized fashion and in large groups, with the standard treatment or with placebo, before marketing authorization, is:

a) phase I

b) phase II

c) phase IV

d) phase III

e) the preclinical phase

49. (MC) Regarding randomization and blinding, the following statements are correct:

a) Block randomization keeps the arms numerically balanced throughout recruitment

b) Allocation concealment prevents the arm from being known before the patient is enrolled

c) An open-label (unblinded) trial cannot be randomized

d) Using a placebo allows the participants to be blinded

e) Per-protocol analysis always preserves the advantages of randomization

50. (SC) A researcher wants to find out whether the proportion of patients with a weight loss of at least 5% differs between the three arms and then to quantify the effect of arm C compared with arm A. The correct sequence of analyses is:

a) ANOVA, then the Tukey test

b) the paired t-test, then the Pearson correlation

c) the chi-square test on the 3 x 2 table, then the RR, risk difference and NNT with 95% CIs for C vs A

d) simple linear regression, then R^2

e) the Kruskal-Wallis test, then the Friedman test

Answers and Explanations

First check your answers against the key, without reading the explanations, and write down your score. Then read the explanations for all the questions you got wrong and for those where you guessed. All the calculations below were checked with software (Python, scipy library).

Test 1: Answer Key

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: Explanations

Question 1. All the other options describe the sample we have (descriptive statistics). Inference means generalizing to the population, and option d does exactly that, by means of a confidence interval.

Question 2. The trap is option d: selection bias is a systematic error and does not disappear in large samples; a convenience sample of 10,000 people remains just as unrepresentative. Option b is wrong for the same reason.

Question 3. Addition rule: $P(F \text{ or HTN}) = 0.30 + 0.20 - 0.08 = 0.42$. Anyone who simply adds 0.30 + 0.20 = 0.50 counts the hypertensive smokers twice.

Question 4. The NYHA classes have a natural order (I milder than IV), but the distances between the classes are not equal, so the variable is ordinal.

Question 5. Mean = (4 + 7 + 7 + 9 + 13)/5 = 40/5 = 8. The ordered values have the value 7 in the middle (position 3), so the median is 7; the mode is also 7. The range is 13 - 4 = 9, not 13. If 13 becomes 43, the mean rises to 14, but the median remains 7: the median is robust to extreme values.

Question 6. $\text{CV} = \frac{100 \times 12}{80} = 15\%$. Option e (0.15%) is the typical mistake of forgetting to multiply by 100, and 6.7% is the inverted ratio (80/12 = 6.67).

Question 7. The line in the box is the median, not the mean (trap a). The box plot describes numerical or ordinal variables; for nominal variables the bar chart or the pie chart is used.

Question 8. $z = \frac{260-200}{40} = 1.5$. From Table E.1, Φ(1.5) = 0.9332, so the area to the right is 1 - 0.9332 = 0.0668, that is, 6.7%. Option b (2.3%) corresponds to z = 2.

Question 9. The total area under any density curve is 1, regardless of the standard deviation (d is false). A Shapiro-Wilk p = 0.30 does not "prove" normality; it only shows that no departure from normality was detected (e is false).

Question 10. In a skewed distribution, the mean and the standard deviation are pulled by the long tail. The median and the interquartile range (151.5 mg/dL; 107-203.25) describe the data correctly. A correct alternative, which does not appear among the options, would be the geometric mean after log transformation.

Question 11. $z = \frac{160-131.13}{13.88} = \frac{28.87}{13.88} = 2.08$. Option e (28.87) is only the numerator.

Question 12. $\text{SE} = \frac{20}{\sqrt{100}} = 2$; $\text{CI}\,95\% = 150 \pm 1.96 \times 2 = 150 \pm 3.92$, that is, 146.1-153.9. Option a (110.8-189.2) is the interval for individual values ($m \pm 1.96s$), not for the mean, the most frequent confusion in the exam.

Question 13. $\text{SE} = \frac{\text{SD}}{\sqrt{n}}$; with an n 4 times larger, $\sqrt{n}$ doubles and the SE is halved. The SD does not decrease systematically with n: it describes biological variability, which does not depend on how many patients we measure. The interval $m \pm 1.96\text{SE}$ contains the population mean with 95% confidence, not 95% of the patients.

Question 14. $p = \frac{40}{200} = 0.20$; $\text{SE} = \sqrt{\frac{0.2 \times 0.8}{200}} = 0.0283$; $\text{CI} = 0.20 \pm 1.96 \times 0.0283 = 0.20 \pm 0.055$, that is, 14.5-25.5%.

Question 15. When the prevalence is unknown, p = 0.5 is used (the most conservative option): $n = \frac{1.96^2 \times 0.5 \times 0.5}{0.05^2} = 384.16$, rounded up to 385. The value 246 corresponds to p = 0.20.

Question 16. The definition in option a is the only correct one. Options b and c are the most widespread misinterpretations: the p value is calculated assuming that H0 is true, so it cannot be the probability that H0 is true.

Question 17. Power is 1 - β, not 1 - α (c is false). A stricter α (0.01) makes H0 harder to reject and, at the same n, lowers power (e is false).

Question 18. This is the classic example of a nonsignificant result from mg2026. H0 is not "accepted" and equality is not "demonstrated": the interval from -4.9 to 3.8 mmHg shows that the data are also compatible with differences of a few mmHg. The value 0.81 is the p value, not the power.

Question 19. The value 1.645 is the critical value for a one-sided test at α = 0.05. For z = 2.30, two-sided p = 2 x (1 - 0.9893) = 0.021: smaller than 0.05 but larger than 0.01, so d is false.

Question 20. SE of the differences = 8/√16 = 8/4 = 2; $t = \frac{4}{2} = 2.0$. Option a (0.5) appears if the mean is divided by the standard deviation (4/8), that is, if dz is calculated instead of t.

Question 21. df = n - 1 = 15; two-sided critical t (α = 0.05) = 2.131 (Table E.3). Since 2.0 < 2.131, H0 is not rejected (c is false). The critical value for two-sided α = 0.10 is 1.753 < 2.0, so 0.05 < p < 0.10 (exactly, p = 0.064). In a prespecified one-sided test, the threshold is 1.753 and the one-sided p = 0.032, so the result would be significant (e is correct). This option does not mean that we are allowed to switch to a one-sided test after seeing the data.

Question 22. Three t-tests without correction increase the overall risk of a type I error. For three independent groups with the conditions met, one-way ANOVA is used.

Question 23. df1 = k - 1 = 2, so k = 3; df2 = N - k = 157, so N = 160. $\eta^2 = 0.226$ means that 22.6% of the variability is explained by arm. ANOVA alone says only that at least one mean differs; the pairs are identified by the post hoc test (Tukey). A Levene test p = 0.163 does not call for the nonparametric alternative.

Question 25. df = (3 - 1) x (4 - 1) = 2 x 3 = 6. Option c (12) is the number of cells.

Question 26. $E = \frac{60 \times 50}{200} = \frac{3000}{200} = 15$.

Question 27. The test is applied to absolute frequencies; on percentages, the χ^2 value would depend on the arbitrary base of 100. χ^2 shows how far the data are from independence, but neither the direction nor the size of the association; for these, the RR, the OR or the difference in proportions are used.

Question 28. Satisfaction on the Likert scale is ordinal, and the arms are independent groups, so the Mann-Whitney test is used (in mg2026: U = 2265.5, p < 0.001).

Question 29. The eyes of the same patient and the results of two tests applied to the same patients are paired observations (for paired dichotomous variables the McNemar test is used). Option e describes three independent groups.

Question 30. With 25 patients per group, strong skewness and a Shapiro-Wilk p = 0.003, the safe alternative for two independent groups is the Mann-Whitney test (or, alternatively, a log transformation followed by the t-test).

Question 31. $R^2 = 0.6^2 = 0.36$.

Question 32. Predicted SBP = 89.47 + 1.435 x 30 = 89.47 + 43.05 = 132.5 mmHg. $R^2 = 0.405^2 = 0.164$. The intercept corresponds to a BMI of 0, an impossible value that lies far from the data, so it is not the "actual" SBP of any patient. Correlation does not demonstrate causality.

Question 33. Spearman works on ranks and measures a monotonic relationship; it is suitable for ordinal variables or for nonlinear monotonic relationships. For nominal variables, contingency tables are used, and for a prediction equation, regression.

Question 34. $R^2 = 0.15^2 = 0.0225$, that is, 2.25%. With n = 210, even an r of 0.15 becomes significant (Table E.6: for n = 100 the threshold at α = 0.05 is 0.197, and at n = 210 the threshold is lower). According to the rules of thumb presented in the course, an |r| below 0.25 means a practically negligible association. Statistical significance means neither relevance nor causality.

Question 35. Risk in the exposed = 30/100 = 0.30; in the unexposed = 10/100 = 0.10; RR = $\frac{0.30}{0.10} = 3.0$.

Question 36. OR = $\frac{30\times90}{70\times10} = \frac{2700}{700} = 3.86$, larger than the RR (d is false: when the disease is not rare, the OR exceeds the RR for a risk factor). Risk difference = 0.30 - 0.10 = 0.20; NNH = $\frac{1}{0.20} = 5$. Attributable fraction among the exposed = (3 - 1)/3 = 0.667.

Question 37. In a case-control study the proportion of diseased people is fixed by the researcher, so risks and incidences cannot be calculated; the OR is the only valid measure.

Question 38. ARR = 20% - 15% = 5% = 0.05; NNT = $\frac{1}{0.05} = 20$. The option 33 would be 1/0.03 (a subtraction error), and 5 is 1/0.20.

Question 39. A confounder is related both to the exposure and to the outcome, but it is not an intermediate link (that is called a mediator). Randomization controls it; increasing the sample size does not, because it is a systematic error.

Question 40. Out of 1,000 people: 20 diseased, of whom 19 are positive (95%); 980 non-diseased, of whom 98 are false positives (10%). PPV = $\frac{19}{19+98} = \frac{19}{117} = 0.162$, that is, approximately 16%. This is the typical example: with a rare disease, even a good test gives more false positives than true positives.

Question 41. Pre-test odds = 0.10/0.90 = 0.111; post-test odds = 0.111 x 9 = 1.0; post-test probability = 1/(1 + 1) = 0.50. A test with Se 90% and Sp 90% has $LR^+ = \frac{0.90}{0.10} = 9$.

Question 43. The complete table: a = 45, b = 5, c = 20, d = 130, N = 200. Se = $\frac{45}{50} = 90\%$; Sp = $\frac{130}{150} = 86.7\%$; PPV = $\frac{45}{65} = 69.2\%$; NPV = $\frac{130}{135} = 96.3\%$; accuracy = (45 + 130)/200 = 87.5%. The error rate is $\frac{b+c}{N} = \frac{5+20}{200}$, not (5 + 130)/200.

Question 44. Misclassification rate = $\frac{\text{FN+FP}}{N} = \frac{5+20}{200} = \frac{25}{200} = 12.5\% = 1 - 0.875$.

Question 45. Diseased = 0.30 x 200 = 60; a = 0.80 x 60 = 48, b = 12. Non-diseased = 140; d = 0.90 x 140 = 126, c = 14. PPV $= \frac{48}{48+14} = \frac{48}{62} = 77.4\%$. Option e (91.3%) is NPV $= \frac{126}{138}$.

Question 46. The AUC is not the percentage of patients correctly classified: it is the probability that a randomly chosen diseased person has a higher score than a randomly chosen non-diseased person. The optimal cut-off according to Youden maximizes Se + Sp - 1, not sensitivity alone.

Question 47. J = 0.85 + 0.75 - 1 = 0.60.

Question 48. The intention-to-treat analysis keeps the groups as randomization formed them and is the primary analysis in superiority trials.

Question 49. Allocation concealment (the next arm is not known in advance) and blinding (the treatment received is not known during the study) are different things. The primary outcome is fixed in the protocol, before data collection.

Question 50. Without a CI we cannot judge the precision of the estimate: an OR of 2.1 in 80 patients can have a very wide interval, compatible with both a small and a large effect.

Test 2: Answer Key

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: Explanations

Question 1. A value that describes the population and that we estimate is a parameter (μ); the mean calculated in the sample would be the statistic (m).

Question 2. Anonymization means that the person can no longer be re-identified; replacing the name with a code while keeping the linking key is pseudonymization (c is false). A simulated dataset with fictitious patients contains no data on real people, so these are not personal data (e is false).

Question 3. For independent events, $P(A \text{ and } B) = 0.3 \times 0.4 = 0.12$. The value 0.70 would be the sum, and 0.58 would be P(A or B).

Question 4. n = 6 (even), so the median is the mean of the values in positions 3 and 4: (8 + 10)/2 = 9. The mean is 58/6 = 9.67, the trap in option b.

Question 5. Valid responses = 200 - 20 = 180; valid percentage = 54/180 = 30%; percentage of total $= \frac{54}{200} = 27\%$. The difference shows why the report must state the base of the calculation and the number of missing values. Missing values are not replaced arbitrarily.

Question 6. IQR = 30 - 20 = 10; upper limit = Q3 + 1.5 x IQR = 30 + 15 = 45.

Question 7. The histogram shows the shape of the distribution of a numerical variable. The bar chart and the pie chart are for categorical variables, the scatter plot for two numerical variables, and the forest plot for estimates with CIs from several studies.

Question 8. $z = \frac{11.6-14}{1.2} = -2.0$; Φ(-2) = 1 - Φ(2) = 1 - 0.9772 = 0.0228, that is, 2.3%.

Question 9. m ± 1.96s = 120 ± 19.6 = 100.4-139.6 (95% of individuals); $m \pm s = 110 - 130(68.3\%)$; $m \pm 3s = 90 - 150(99.7\%)$. Below 100.4 lie 2.5% (half of the 5% that remain outside the interval). The interval 100.4-139.6 describes individuals, not the population mean: a CI for the mean would use SE $= \frac{s}{\sqrt{n}}$ and would be much narrower.

Question 10. The exponential of the mean of the logarithms is the geometric mean; in mg2026, for triglycerides, it is 144.0 mg/dL, compared with an arithmetic mean of 161.18 mg/dL.

Question 11. The theorem refers to the distribution of sample means, not to the distribution of the variable. Option d is false: the standard deviation does not decrease with n; the standard error does.

Question 12. SE $= \frac{8}{\sqrt{16}} = 2$; CI $95\% = 50 \pm 2.131 \times 2 = 50 \pm 4.26$, that is, 45.7-54.3. Option a (46.1-53.9) uses z = 1.96, which is inappropriate for n = 16 with σ unknown.

Question 13. $n = \left(\frac{1.96\times10}{2}\right)^2 = 9.8^2 = 96.04$. The sample size is always rounded up, so 97; with 96 patients the margin would be slightly above 2 units.

Question 14. The confidence interval refers to the population mean, not to the patients' values (b is false). A higher confidence level requires a wider interval; with an n 4 times larger, the SE is halved.

Question 15. $p = \frac{2}{25} = 0.08$; SE $= \sqrt{\frac{0.08\times0.92}{25}} = 0.054$; Wald lower limit = 0.08 - 1.96 x 0.054 = -0.026, which is impossible for a proportion. The Wilson interval solves the problem and stays between 0 and 1.

Question 16. H0 is neither "accepted" nor "demonstrated" (c is false). For correlation, H0 is ρ = 0 (no correlation), not ρ = 1.

Question 17. α' = 0.05/5 = 0.01.

Question 18. With 20,000 patients, even a decrease of 1.2 mmHg becomes extremely significant statistically. The upper limit of the CI (1.6 mmHg), however, shows that the effect is at most very small from a clinical point of view.

Question 19. Power increases with n, with the effect size and with α, and decreases with variability. The paired design removes between-individual variability; giving it up, with the same number of measurements, usually lowers power.

Question 20. df = n1 + n2 - 2 = 25 + 30 - 2 = 53.

Question 21. SE $= \frac{11.57}{\sqrt{160}} = \frac{11.57}{12.649} = 0.915$; $t = \frac{10.19}{0.915} = 11.14$; df = 159; dz $= \frac{10.19}{11.57} = 0.88$ (a large effect according to Cohen's benchmarks). The paired t-test does not require equal variances at the two time points: it works on a single series, that of the differences, whose normality must be checked (in mg2026, Shapiro-Wilk p = 0.725).

Question 22. With unequal variances, the Welch t-test is used, which adjusts the degrees of freedom.

Question 23. $\eta^2 = \frac{\text{SSB}}{\text{SST}} = \frac{300}{1200} = 0.25$.

Question 24. $d = \frac{136-130}{12} = 0.5$, a medium effect according to Cohen's benchmarks. Option b has the correct figure but the wrong interpretation.

Question 25. $\chi^2 = \frac{N(\text{ad}-\text{bc})^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{100\times(800-300)^2}{50\times50\times30\times70} = 25{,}000{,}000/5{,}250{,}000 = 4.76$, with df = 1. Since 3.841 < 4.76 < 6.635, it follows that 0.01 < p < 0.05 (exactly, p = 0.029). The expected frequency for $a = \frac{(a+b)(a+c)}{N} = \frac{50\times30}{100} = 15$.

Question 26. The results of the two tests come from the same patients (paired dichotomous data), so the McNemar test is used. The chi-square test of independence would assume independent groups.

Question 27. We order all 7 values: 12 (A, rank 1), 14 (B, 2), 15 (A, 3), 18 (A, 4), 20 (B, 5), 22 (B, 6), 25 (B, 7). $R(A) = 1 + 3 + 4 = 8$; $U(A) = 8 - \frac{3\times4}{2} = 2$; $U(B) = 3 \times 4 - 2 = 10$. The smaller value is U = 2 (exact p = 0.23, not significant with such small samples).

Question 28. A significant Kruskal-Wallis test says only that at least one group differs; the pairs are identified by post hoc tests (Dunn). The Friedman test is for repeated measurements, not for independent groups.

Question 29. Both variables are dichotomous and the expected frequencies are above 5: the chi-square test on the 2 x 2 table (followed by calculation of the OR or RR with its CI).

Question 30. For two normal numerical variables with a linear relationship, the Pearson coefficient is the appropriate choice; Spearman is not "mandatory". For three independent groups with normal data, one-way ANOVA is used, not Friedman.

Question 31. rho $= 1 - \frac{6\times14}{6\times(36-1)} = 1 - \frac{84}{210} = 1 - 0.4 = 0.6$.

Question 32. df = 27 - 2 = 25; $t = \frac{0.5\times\sqrt{25}}{\sqrt{1-0.25}} = \frac{2.5}{0.866} = 2.89 > 2.060$, so p < 0.05 (exactly, p = 0.008). $R^2 = 0.25$. Table E.6 confirms this: for n = 27, the critical r at 0.05 is below 0.5.

Question 33. 108.3 + 0.406 x 60 = 108.3 + 24.36 = 132.7 mmHg. Option b (24.4) is only the product.

Question 34. In a multiple model, the coefficient of a predictor is its effect adjusted for the other predictors: +1.389 mmHg for each kg/m², at equal age, sex and smoking status.

Question 35. Odds of exposure in cases = 40/60; in controls = 20/80; OR $= \frac{40\times80}{60\times20} = \frac{3200}{1200} = 2.67$. Option a (2.0) is the ratio of the proportions exposed (40%/20%), which has no meaning as a risk in a case-control study.

Question 36. The CI contains 1, so p > 0.05 and the association is not significant. But the CI extends up to 3.6: the data do not exclude an important effect. A nonsignificant result does not demonstrate the absence of an effect (d is false).

Question 37. Cumulative incidence = 50/2,000 = 0.025 = 2.5% over 2 years. The incidence rate would require person-time (approximately 50/4,000 = 12.5 per 1,000 person-years, if everyone was followed for 2 years); options d and e do not correspond to any correct calculation.

Question 38. Selection starts from the disease (cases and controls), and exposure is sought retrospectively: a case-control study.

Question 39. Prevalences: 18/54 = 33.3% and 11/106 = 10.4%; PR = 3.21; OR $= \frac{18\times95}{36\times11} = \frac{1710}{396} = 4.32$. A cross-sectional study measures exposure and disease simultaneously, so it cannot establish their sequence (d is false). Diabetes is not rare in the sample (18.1%), which is why the OR moves away from the PR.

Question 40. Pre-test odds = 0.2/0.8 = 0.25; post-test odds = 0.25 x 8 = 2.0; post-test probability = 2/3 = 66.7%. Option e (160%) appears when the probability is multiplied directly by the LR, an unforgivable error: probabilities cannot exceed 100%.

Question 41. Expert opinion lies at the bottom of the hierarchy of evidence, and an observational study, however large, does not eliminate confounding factors the way randomization does.

Question 42. At constant Se and Sp, the PPV decreases and the NPV increases as prevalence decreases. In mg2026, for FINDRISC ≥ 12, the PPV would be 5.8% at a prevalence of 2% and 75.1% at a prevalence of 50%.

Question 43. $a + c = \frac{a}{\text{PPV}} = \frac{40}{0.8} = 50$, so c = 10. $b + d = \frac{b}{1-\text{NPV}} = \frac{10}{0.1} = 100$, so d = 90. Se $= \frac{40}{40+10} = 80\%$; Sp $= \frac{90}{10+90} = 90\%$. N = 40 + 10 + 10 + 90 = 150, not 200.

Question 44. κ = (0.85 - 0.50)/(1 - 0.50) = 0.35/0.50 = 0.70 (substantial agreement). Option a (0.35) is only the numerator.

Question 45. SnNOut: a highly sensitive test misses few diseased people, so a negative result practically rules out the disease.

Question 46. A higher cut-off makes the test more "severe": fewer positives, hence lower sensitivity and higher specificity. The area under the ROC curve characterizes the test across all cut-offs and does not change when the cut-off is moved.

Question 47. Serial testing (positive only if both tests are positive) increases specificity and decreases sensitivity (d is false); parallel testing does the opposite.

Question 48. n adjusted $= \frac{64}{1-0.20} = \frac{64}{0.8} = 80$. Option b (77) comes from the incorrect calculation 64 x 1.2 = 76.8: adding 20% to 64 does not compensate for a loss of 20% of the number recruited.

Question 49. Numerical variables are collected as numbers, with the unit fixed in the form and with a validation range; free text with a unit ("87 mg/dl") is exactly the error in the mg2026 raw file. Clinical trials are registered before the first patient is enrolled.

Question 50. A p value is never exactly 0; the software displays "<.001" because the value is very small. Report p < 0.001.

Test 3: Answer Key

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: Explanations

Question 2. Bias does not decrease with the sample size (b is false), and the confidence interval quantifies only random error, not bias (e is false). A narrow CI from a biased study is a precise estimate of a wrong value.

Question 3. $P(A) \times P(B) = 0.3 \times 0.4 = 0.12 = P(A \text{ and } B)$, so A and B are independent; $P(A \mid B) = \frac{0.12}{0.4} = 0.3 = P(A)$, the same conclusion. $P(A \text{ or } B) = 0.3 + 0.4 - 0.12 = 0.58$. The events are not mutually exclusive, because P(A and B) $\neq 0$. $P(\text{non} - A) = 1 - 0.3 = 0.7$.

Question 4. Geometric mean = (10 x 100 x 1000)^(1/3) = (1,000,000)^(1/3) = 100. The arithmetic mean (370) is dominated by the largest value.

Question 5. $\text{CV}(A) = \frac{5}{50} = 10\%$; $\text{CV}(B) = \frac{30}{200} = 15\%$. B has both the greater absolute variability (s) and the greater relative variability.

Question 6. Blood group is nominal (the categories have no order). Temperature in degrees Celsius is on an interval scale: 0 °C does not mean the absence of temperature, and 20 °C is not "twice as warm" as 10 °C.

Question 9. The positive skewness (1.42) pulls the mean (29.04) above the median (28.25). In a non-normal distribution, the interval $m \pm 1.96s$ does not contain exactly 95% of the patients.

Question 10. The area to the right of 1.96 is 2.5% (Φ(1.96) = 0.9750); 1.645 leaves 5% to its right, and 2.576 leaves 0.5%.

Question 11. The 90th percentile means that 90% of the reference children have lower values; it is not a diagnosis of obesity and does not express a percentage of the mean.

Question 12. $p = \frac{18}{60} = 0.30$; $\text{SE} = \sqrt{\frac{0.3 \times 0.7}{60}} = 0.0592$; $\text{CI} = 0.30 \pm 1.96 \times 0.0592 = 0.30 \pm 0.116$, that is, 18.4-41.6%. Option e confuses the proportion with the percentage.

Question 13. $n = \frac{z^2 \times p_i(1-p_i)}{E^2}$ increases when E decreases and when z increases (99% vs 95%), and for means $n = \left(\frac{z\sigma}{E}\right)^2$ increases with σ. The product $p_i(1 - p_i)$ is largest at $p_i = 0.5$. The size of the population matters only when the sample represents a large fraction of it; going from 1 to 2 million changes practically nothing.

Question 14. The width of the CI is proportional to $\frac{1}{\sqrt{n}}$: √(40/160) = √0.25 = 0.5, so the interval becomes approximately half as wide.

Question 15. $n = \frac{1.96^2 \times 0.10 \times 0.90}{0.03^2} = \frac{3.8416 \times 0.09}{0.0009} = 384.16$, so 385.

Question 16. $\text{SE} = \frac{16}{\sqrt{64}} = 2$; $t = \frac{128-120}{2} = 4$, with df = 63 ($p < 0.001$).

Question 17. The p value is not the probability that H0 is true (a is false). The values 0.049 and 0.051 provide practically the same evidence; the 0.05 threshold is a convention, not a natural boundary (e is false).

Question 18. Choosing a one-sided test after seeing the data (for example, to "rescue" a two-sided p of 0.08) is an improper practice.

Question 19. The probability that no test is falsely significant = 0.95^20 = 0.358; the probability of at least one false positive result = 1 - 0.358 = 0.64. Hence the need for corrections for multiple comparisons.

Question 20. $F = \frac{\text{MSB}}{\text{MSW}} = \frac{150}{30} = 5$.

Question 21. Between-group df = k - 1 = 2; residual df = N - k = 60 - 3 = 57. Since 5 > 3.16, $p < 0.05$ (exactly, p = 0.010). A significant ANOVA shows only that at least one mean differs, not that they all differ from one another.

Question 24. The t-test does not require equal group sizes; it requires independent observations, normality (or large samples) and, in the classical version, equal variances.

Question 25. df = (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. The minimum expected frequency (12.6) exceeds 5, so the chi-square test is appropriate and Fisher's test is not necessary. The overall test on the 3 x 2 table shows only that the proportions are not all equal; moreover, arms B and C have almost identical proportions (16/53 and 16/54), so option c is false.

Question 28. Repeated measurements (the same patients, three time points) and non-normal data in a small sample: the Friedman test. Kruskal-Wallis would be appropriate for three independent groups.

Question 29. Two independent groups, normal data, homogeneous variances (Levene p = 0.40): the independent-samples t-test; the Welch version is also acceptable.

Question 30. For an ordinal variable, rank-based tests are used or, at the cost of losing the information about order, the chi-square test. Treating the codes 0-4 as continuous normal values assumes equal distances between stages, which is not justified.

Question 31. |r| = √0.25 = 0.5; the sign of r is the sign of the slope, so r = -0.5.

Question 32. Extrapolation outside the observed range is risky, and a large R^2 shows a strong association, not causality.

Question 33. $\text{OR} = e^{0.693} = 2.0$. The coefficient b is on the log-odds scale and must be exponentiated.

Question 34. A value that is extreme on both axes can create or inflate a correlation. The point is checked in the source document and the analysis is run with and without it, but it is not deleted just because it "spoils" the result.

Question 35. Risk in the treated = 8/200 = 0.04; risk on placebo = 16/200 = 0.08; $\text{RR} = \frac{0.04}{0.08} = 0.5$.

Question 36. ARR = 0.08 - 0.04 = 0.04 (4 percentage points); $\text{RRR} = \frac{0.04}{0.08} = 50\%$; $\text{NNT} = \frac{1}{0.04} = 25$. $\text{OR} = \frac{8\times184}{192\times16} = \frac{1472}{3072} = 0.48$, close to the RR (the event is rare) but not equal to it. A "risk cut in half" sounds spectacular, although it means only 4 events avoided per 100 patients treated.

Question 37. $\text{PAF} = \frac{0.25\times(3-1)}{1+0.25\times(3-1)} = \frac{0.5}{1.5} = 0.333$.

Question 38. Prevalence is approximately incidence multiplied by the duration of the disease; a long duration makes prevalence high even when incidence is low.

Question 39. Risks cannot be calculated in a case-control study, so neither can the RR directly (c is false). The cross-sectional study does not establish the temporal sequence (e is false).

Question 40. Out of 10,000 people: 100 diseased, 90 positive; 9,900 non-diseased, 495 false positives (5%). $\text{PPV} = \frac{90}{90+495} = \frac{90}{585} = 15.4\%$. At a prevalence of 20%, the same test would have a PPV of approximately 82%.

Question 41. $\text{LR}^- = \frac{1-0.8}{0.9} = \frac{0.2}{0.9} = 0.22$. Likelihood ratios are calculated only from Se and Sp, so they do not depend on prevalence (e is false); this is why they are easier to transfer from one study to another than predictive values.

Question 42. p = odds/(1 + odds) = 3/4 = 0.75. Option b (0.33) is the reciprocal of the odds.

Question 43. $\mathrm{DOR} = \frac{a \times d}{b \times c} = \frac{80 \times 90}{20 \times 10} = \frac{7200}{200} = 36$. Check: $\mathrm{LR}^{+} = \frac{0.8}{0.1} = 8$, $\mathrm{LR}^{-} = \frac{0.2}{0.9} = 0.222$, and 8/0.222 = 36.

Question 44. $\mathrm{Se} = \frac{80}{100} = 80\%$; $\mathrm{Sp} = \frac{90}{100} = 90\%$; accuracy = (80 + 90)/200 = 85%; error rate $= \frac{b+c}{N} = \frac{20+10}{200} = 15\% = 1 - 0.85$. The formula (b + d)/N, found in some materials, is wrong.

Question 45. In the mg2026 scenario, diabetes was confirmed by the oral glucose tolerance test (OGTT), the reference standard against which the Se and Sp of the FINDRISC score are calculated.

Question 46. An area between 0.81 and 0.90 indicates very good discrimination. The CI 0.762-0.943 excludes 0.5. The difference of 0.008 between the two areas is much smaller than the uncertainty of each estimate; superiority could be claimed only after a formal test comparing the areas (c is false).

Question 47. When the reference standard is applied selectively, depending on the result of the test under study, verification bias arises, which usually overestimates sensitivity.

Question 48. Phase III compares the new treatment, in a randomized fashion and in large groups, with the standard treatment or with placebo, before marketing authorization. Phase IV takes place after authorization.

Question 49. Many randomized trials are open-label (for example, those with educational interventions, like the arms in mg2026); randomization does not depend on blinding. Per-protocol analysis excludes patients who did not adhere to the protocol and can destroy the balance created by randomization.

Question 50. The outcome is dichotomous and there are three groups, so the chi-square test on the 3 x 2 table is used ($\chi^2 = 6.77$, p = 0.034 in mg2026). For the clinical comparison of C vs A, report RR = 2.62 (95% CI 1.11-6.17), a risk difference of 18.3 percentage points (95% CI 3.4-33.2) and NNT = 5.5, that is, 6.

References

The references are numbered in the order of their first appearance in the text (Vancouver style). Each article was checked in PubMed or CrossRef before being cited; official web pages were accessed on the date indicated.

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, Şirli 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, Secoșan 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 Hațieganu"; 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 Hațieganu"; 2025. ISBN 978-606-075-342-1 (PDF).

151. Țarcă V. Noțiuni fundamentale de biostatistică. Iași: Junimea; 2025. ISBN 978-973-37-2931-0.