AN MD COMMUNITY MEDICINE COMPANION

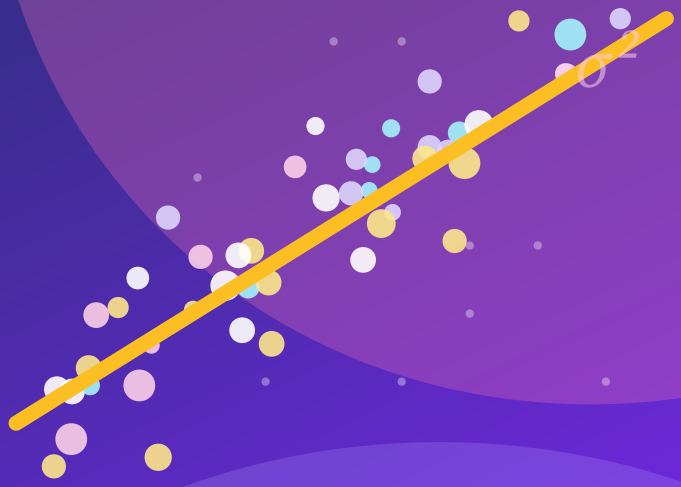
VERSION 3.1

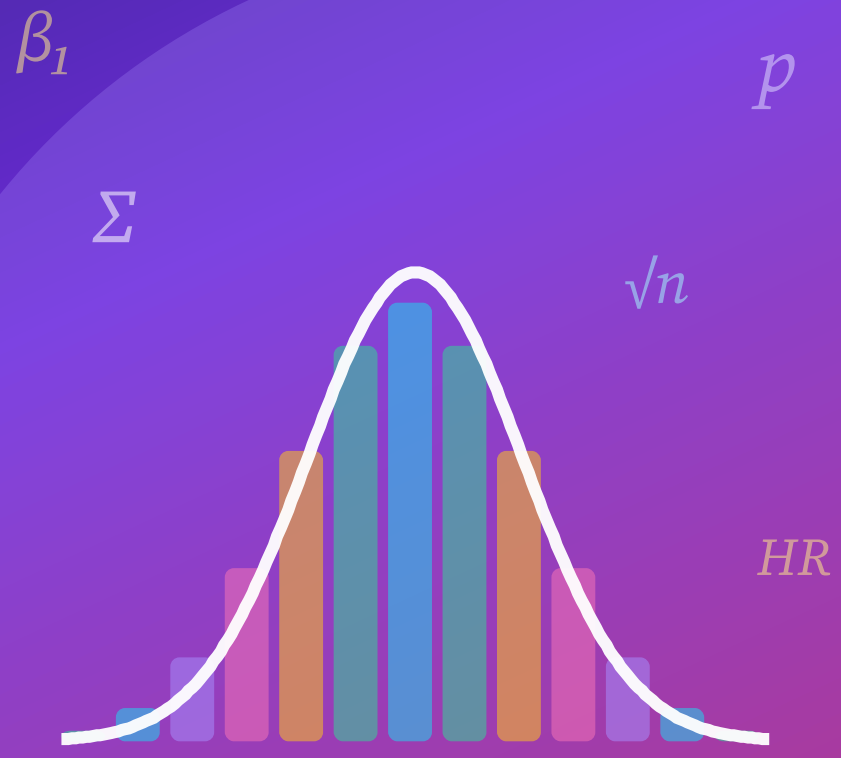
Statistics

From First Principles to Regression

An introductory book, written as an unbroken chain of logic

- Derived, not memorised
- Worked Indian examples
- Optional R walk-throughs
- 94 checkpoints with answers





Dr. Harsh Maheshwari

MD Community Medicine Resident, AIIMS Raipur

AN MD COMMUNITY MEDICINE COMPANION

Statistics

From First Principles to Regression

An introductory book, written as an unbroken chain of logic

Dr. Harsh Maheshwari
MD Community Medicine Resident, AIIMS Raipur

contact@drharshmaheshwari.com · drharshmaheshwari.com

Version 3.1 · 28 September 2026

Title	Statistics: From First Principles to Regression
Version	3.1
Last updated	28 September 2026
Author	Dr. Harsh Maheshwari, MD Community Medicine Resident, AIIMS Raipur
Contact	contact@drharshmaheshwari.com
Newest version	drharshmaheshwari.com

© 2026 Harsh Maheshwari.

This book is shared under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International licence (CC BY-NC-SA 4.0). You may copy, share and adapt it for non-commercial purposes, if you credit the author and share what you make under the same licence.

It quotes short passages from papers, guidelines and other works, each credited where it is used; those passages remain under their owners' terms. The datasets used in the worked examples and the “In R” boxes are synthetic teaching data, not records of real people. Fonts: Source Serif 4, Inter and JetBrains Mono, SIL Open Font License 1.1.

HOW THIS BOOK WAS MADE

Written for the author's own study, then shared. Revised and checked with AI assistance (Claude, by Anthropic) under the author's direction. Every worked number was recomputed, every R example re-run, and every cited source opened. Mistakes can remain; the last pages say how to report one.

NOT ADVICE

This book is for learning. It is not medical, legal or financial advice, and it does not replace a qualified professional who knows your situation.

ABOUT THE AUTHOR



Dr. Harsh Maheshwari

MD Community Medicine Resident, AIIMS Raipur

Harsh Maheshwari completed his MBBS at King George's Medical University (KGMU), Lucknow, and is a resident in MD Community Medicine at the All India Institute of Medical Sciences (AIIMS), Raipur. His MD thesis, on chrononutrition and obesity, was selected by the Indian Council of Medical Research (ICMR) for funding in 2026.

He learns best from first principles: rebuilding an idea from what came before it until the formula feels like the only reasonable answer. This book grew out of that habit.

His long-term aims are to reduce obesity in India and to help make India healthy. He is also the author of *Obesity Expertise*, a first-principles curriculum written towards the first of those aims.

Email **contact@drharshmaheshwari.com**

Website **drharshmaheshwari.com**

Contents

1 **Statistics – Foundations of Data** 14

1.1 What Is a Variable 14
1.2 Population vs Sample 20
1.3 Parameter vs Statistic 21
1.4 Organising Raw Data 22
1.5 Study Designs – A First Map 29
1.6 Data in Software – Codebooks, Entry, Cleaning and an R Primer 32

2 **Descriptive Statistics – Central Tendency** 36

2.1 Why We Need a Typical Value 36
2.2 Mean 37
2.3 Median 41
2.4 Mode 43
2.5 Choosing the Right Measure of Central Tendency 45

3 **Descriptive Statistics – Dispersion** 47

3.1 Why Average Alone Is Not Enough 47
3.2 Range 48
3.3 Quartiles Deciles and Percentiles 48
3.4 Interquartile Range 49
3.5 Mean Deviation 50
3.6 Variance 51
3.7 Standard Deviation 55
3.8 Coefficient of Variation 57
3.9 Skewness 57
3.10 Kurtosis 59

4 **Data Visualisation** 61

4.1 Tables for Data Presentation 61
4.2 Diagrams for Qualitative Data 62
4.3 Diagrams for Quantitative Data 66
4.4 Scatter Diagram 74
4.5 Diagrams for Time, Place and Population 75

5 **Foundations of Probability** 87

5.1 From Counting to Probability 87
5.2 Basic Probability Rules 91
5.3 Conditional Probability 94
5.4 Independence of Events 95
5.5 Bayes' Theorem 96
5.6 Random Variables 102
5.7 Probability Distributions 104

6 **Sampling** 124

6.1 Why We Sample 124
6.2 Probability Sampling Methods 125
6.3 Non-Probability Sampling Methods 134
6.4 Sampling Error vs Bias 136

6.5 Sampling Distribution of a Statistic 138
6.6 Central Limit Theorem 140
6.7 Standard Error 141

7 **Estimation** **143**

7.1 Point Estimation 143
7.2 Confidence Intervals 145
7.3 Margin of Error and Sample Size Estimation 154
7.4 Rates, Ratios and Standardisation 156
7.5 The Life Table 165

8 **Hypothesis Testing – The Framework** **169**

8.1 Logic of Hypothesis Testing 169
8.2 Null and Alternative Hypothesis 170
8.3 Type I and Type II Error 170
8.4 Significance Level and Power 172
8.5 p-value – Correct Interpretation 174
8.6 One-Tailed vs Two-Tailed Tests 176
8.7 General Steps in Hypothesis Testing 178
8.8 Choosing the Right Statistical Test 179

9 **Comparing Groups – Parametric Tests** **183**

9.1 z-test 183
9.2 t-test 188
9.3 Tests of Normality 194
9.4 Data Transformation Techniques 197
9.5 ANOVA 198

10 **Comparing Groups – Non-Parametric Tests** **208**

10.1 When to Use Non-Parametric Tests 208
10.2 Mann-Whitney U Test 211
10.3 Wilcoxon Signed-Rank Test 213
10.4 Kruskal-Wallis Test 214
10.5 Friedman Test 216
10.6 Chi-Square Test 217
10.7 Fisher's Exact Test 228
10.8 McNemar's Test 230

11 **Correlation** **234**

11.1 The Idea of Association 234
11.2 Pearson Correlation Coefficient 235
11.3 Spearman Rank Correlation 239
11.4 Point-Biserial and Phi Coefficient 240
11.5 Correlation vs Causation 241
11.6 Confounding – A Working Introduction 242
11.7 Stratified Analysis – Mantel–Haenszel, Effect Modification and Simpson's Paradox 247

12 **Regression – Foundations** **253**

12.1 From Correlation to Prediction 253
12.2 Simple Linear Regression 254
12.3 Assumptions of Linear Regression 263
12.4 Multiple Linear Regression 265
12.5 Regression Diagnostics 275

12.6 Analysis Journey – Multiple Linear Regression 278

13 **Generalised Linear Models 282**

13.1 Why We Need a Generalised Framework 282
13.2 Logistic Regression 284
13.3 Analysis Journey – Logistic Regression 296
13.4 Counts, Person-Time, Rate, and the Offset 298
13.5 Poisson Regression 299
13.6 Negative Binomial Regression 301
13.7 Estimating Prevalence and Risk Ratios Directly 303

14 **Survival Analysis 308**

14.1 Why Ordinary Regression Fails for Time-to-Event Data 308
14.2 Censoring 309
14.3 Survival Function, Hazard, and the At-Risk Set 310
14.4 Kaplan-Meier Survival Curve 311
14.5 Log-Rank Test 315
14.6 Cox Proportional Hazards Regression 317
14.7 Analysis Journey – Cox Regression 323
14.8 Competing Risks 326

15 **Advanced Topics for the Community Medicine Researcher 331**

15.1 Model Selection 331
15.2 Clustered and Hierarchical Data 340
15.3 Confounding Handling Techniques 355
15.4 Missing Data 362
15.5 Diagnostic Test Statistics 367
15.6 Sample Size and Power Calculation 375
15.7 Meta-Analysis Statistics 378
15.8 Introduction to Bayesian Statistics 387
15.9 Randomised Trials – Randomisation, ITT, Baseline Adjustment, Cluster and Non-Inferiority Designs 390

16 **Measurement Theory – Constructs, Validity, and Reliability 395**

16.1 What Is a Construct 395
16.2 Validity of a Measurement Instrument 396
16.3 Reliability of a Measurement Instrument 401
16.4 Relationship Between Validity and Reliability 415
16.5 Principal Component Analysis 416
16.6 Building Composite Variables 420
16.7 Questionnaire Validation Methodology 421

17 **Threats to Study Validity 423**

17.1 Validity of a Study 423
17.2 Bias in Research 426
17.3 Confounding 437
17.4 Reporting Guidelines and Writing the Statistical Methods 441
17.5 Qualitative and Mixed Methods – What a Quantitative Researcher Must Know 443

- Checkpoint Answers.....446
- R Quick Reference 496
- References.....500

Introduction

Why this book exists

Most statistics teaching — and most statistics textbooks — present formulas as things to be memorised: here is the formula for variance, here is the formula for a t-statistic, learn them, apply them. This book takes a different, deliberately slower path, in the spirit of Richard Feynman's teaching: start from what you already know and build, without a single unexplained leap. Every idea here is derived, not asserted. Before a formula is given, the specific problem it solves is explained first, so that by the time the formula appears, it feels less like something to memorise and more like the only reasonable answer to a question you already understand.

This matters more, not less, for a working clinician. An MD Community Medicine resident who has memorised "divide by n−1 for sample variance" without knowing *why* will apply the rule correctly on an exam and then completely fail to recognise when it doesn't apply — a clustered dataset, an unusual missingness pattern, a test built on an assumption quietly violated. Understanding the chain of reasoning behind a technique is what lets you recognise its edges. That is the entire premise of this book: no mugging up, an unbroken chain from first principles to the techniques used in real research.

How to use this book

Read it in order, front to back. Every concept builds only on concepts that came before it — nothing is used before it is taught. The sign **§** means "section": §7.3 is section 7.3. When an earlier idea is needed again, the text says "recall §n.n" and points back to where it was built; when a later section is mentioned early, it is named with a one-line explanation.

The book has seventeen chapters. Chapters 1–4 describe data: variables, study designs, summaries and diagrams. Chapters 5–8 build probability, sampling, estimation and the logic of the hypothesis test. Chapters 9–14 compare groups and model outcomes: parametric and non-parametric tests, correlation and confounding, linear regression, generalised linear models and survival analysis. Chapters 15–17 turn to research practice: clustered and survey data, missing data, prediction, sample size, meta-analysis and trials, then measurement and the threats to a study's validity.

Each concept follows the same four-part structure:

- **Definition** — the precise, technical statement.
- **Simplified Explanation** — the same idea rebuilt from first principles, in plain language, usually with an analogy or a callback to something already covered.
- **Illustration / Example** — one concrete worked example, drawn mostly from a running Community Medicine context (an under-five clinic, a CBHI household survey, a district outbreak).
- **Must-Know Notes (MD-level)** — the compressed, exam- and practice-relevant takeaways.

Four further blocks appear where they earn their place:

- **Derivation** — the few lines that show *why* a formula is true.
- **Worked Calculation** — numbered steps that carry an example through to a stated result.
- **In R (optional)** — the same analysis run in the free statistical software R, with its real output and a reading of what each number means.
- **Common Misreading** — the tempting wrong interpretation of a result, and why it is wrong.

After each group of related concepts comes a **Checkpoint**, numbered within its chapter (Checkpoint 5.3 is the third in Chapter 5): a few questions to answer from memory, in your own words, before looking anything up. Trying to retrieve and explain an idea is one of the most effective ways to find out whether you understood it, and to make it stick. Answers, with the working shown, are in **Appendix A** at the back. Try each question before you turn there.

Conventions, cut-offs and named rules are cited author–year in the text, and the full sources are listed in the **References** at the end.

If you are new to statistics

- **You need only school mathematics:** fractions, percentages, powers, square roots and logarithms. Anything beyond that is explained where it first appears, or marked as beyond this book.
- **On a first reading,** read the Definition, Simplified Explanation, Illustration and Must-Know Notes of each section, and do the Checkpoints. The Derivation boxes can be skimmed and revisited once the idea feels familiar.
- **Keep a calculator beside you.** Every worked calculation can be followed with pen, paper and a calculator.
- **Use the list of symbols below** whenever a sign or letter looks unfamiliar. Each symbol is also explained in the text where it is first used.

R and software are optional

You do not need R, or any other software, to read this book. Every idea, calculation and Checkpoint can be followed without a computer, and no section depends on an In R box. The boxes are there for readers who want to run the analyses themselves. They use small synthetic teaching datasets that behave like real community-medicine data; no row describes a real person, household or village. If you want to try them, §1.6 includes a short R primer, and **Appendix B** lists every R function used, with the sections where it appears. On a first reading you can skip every In R box and Appendix B.

Symbols and notation

Statistics borrows symbols from mathematics and letters from the Greek alphabet. The tables below list the ones used in this book, how to read them aloud, what they mean here, and where they are first explained. You do not need to learn them in advance; come back to this page whenever a symbol stops you.

Signs and operations

Symbol	Read as	Meaning in this book	First explained
§	"section"	A section of this book: §7.3 is section 7.3	Here
±	"plus or minus"	A range either side of a value: 11.8 ± 0.6 means 11.2 to 12.4	§3.7
× , · , ÷	"times", "divided by"	Multiplication (× or a raised dot: $n \cdot \bar{x} = n \times \bar{x}$) and division	—
≈	"approximately equal to"	The two sides are close but not exactly equal	§1.4.4
≠	"not equal to"	The two sides differ	§3.6.2
< , >	"less than", "greater than"	Strict inequalities	—

Symbol	Read as	Meaning in this book	First explained
$\leq$, $\geq$	“less than or equal to”, “greater than or equal to”	Inequalities that include the boundary	§1.4.2
$\sqrt{}$	“square root of”	$\sqrt{9} = 3$; $\sqrt[3]{}$ is the cube root	§2.2.3
x^2, x^3, a^b	“x squared”, “x cubed”, “a to the power b”	Powers: $x^2 = x \times x$; a^b is used when the power is long, as in e^(−λ)	—
10^{-5}	“ten to the minus five”	0.00001 (one divided by 10^5)	—
∞	“infinity”	Without limit	§5.7.2.3
$\propto$	“is proportional to”	Changes in fixed ratio with	§7.2
$\rightarrow$, $\Longrightarrow$, $\Rightarrow$	“gives”, “implies”	Leads to, or follows from; in a causal diagram, an arrow ($\rightarrow$ or $\leftarrow$) points from cause to effect (§11.6.1); $\leftrightarrow$ means “corresponds to”	§1.1.2
n!, C(n, k)	“n factorial”, “n choose k”	$n \times (n-1) \times \ldots \times 2 \times 1$, e.g. 4! = 24; C(n, k) is the number of ways of choosing k items from n	§5.7.1.2
Σ	“sigma”, “the sum of”	Add up everything that follows: $\Sigma x = x_1 + x_2 + \ldots + x_n$	§2.2.1
Π	“pi (capital)”, “the product of”	Multiply everything that follows	§13.2.3
$\lvert x \rvert$	“absolute value of x”	x without its sign: $\lvert -3 \rvert = 3$	§2.3
P(A)	“P of A”	The probability of event A	§5.1
$\mid$	“given”	P(A $\mid$ B): the probability of A given that B has happened	§5.3
~	“is distributed as”	X ~ Binomial(n, p): X follows a binomial distribution	§5.7.1.2
e, exp	“e”, “exponential”	The constant 2.718..., the base of natural logarithms; exp(x) = e raised to the power x, which undoes ln	§5.7.1.3 (exp: §2.2.3)
ln, log	“natural log”, “log”	The logarithm to base e (ln), or to base 10 ($\log_{10}$) where stated	§2.2.3

Letters with marks and subscripts

Symbol	Read as	Meaning in this book	First explained
x_1, x_2, x_i	“x one”, “x i”	The first, second, i-th value; the small letter below is a label, not a power	§2.2.1
$\bar{x}$, $\bar{y}$	“x bar”, “y bar”	The mean of a sample of x (or y) values	§1.3
$\hat{p}$, $\hat{y}$	“p hat”, “y hat”	An estimate or a predicted value (the “hat” marks something calculated from data)	§1.3, §12.2
p	“p”	A proportion or probability; also the **p-value** of a test (the context makes clear which)	§1.3, §8.5
n, N	“n”, “N”	The number of observations in a sample (n) or a population (N)	§2.2.1
H_0, H_1	“H nought”, “H one”	The null and alternative hypotheses	§8.2
E(X)	“expected value of X”	The long-run average of a random quantity	§3.6.2, §5.6.1

Symbol	Read as	Meaning in this book	First explained
Var(X), SD, SE	"variance", "standard deviation", "standard error"	Measures of spread of values (SD) and of an estimate (SE)	§3.6, §3.7, §6.4
s, s^2	"s", "s squared"	The standard deviation and variance of a sample	§3.6.2, §3.7
df	"degrees of freedom"	The number of values free to vary once constraints are fixed	§3.6.3
z, t, F	"z", "t", "F"	Test statistics and the distributions they follow	§5.7.2.2.1, §5.7.2.3, §5.7.2.5
r, R^2	"r", "R squared"	Pearson's correlation in a sample; the share of variation a regression explains	§11.2, §12.2.5
b_0, b_1	"b nought", "b one"	Intercept and slope of a regression line fitted to a sample	§12.2
n'	"n prime"	A modified version of a quantity (here, an adjusted number at risk)	§14.4

Greek letters

Greek letters usually stand for **population** values, which are fixed but unknown; Roman letters stand for what you calculate from a sample (recall §1.3). A few letters have more than one use; the text always says which one it means.

Letter	Name	Meaning in this book	First explained
μ	mu	Population mean	§1.3
σ, σ^2	sigma	Population standard deviation and variance	§1.3, §3.6.1
α	alpha	Significance level: the Type I error rate a test accepts (also the negative binomial dispersion, §13.6, and Cronbach's alpha, §16.3.1.1)	§8.3, §8.4
β	beta	Probability of a Type II error (§8.3); a regression coefficient in the population, as in β_0, β_1 (§12.2.1)	§8.3, §12.2.1
χ^2	chi-squared	The chi-square distribution and test statistic	§5.7.2.4
ν	nu	Degrees of freedom of a t or χ^2 distribution	§5.7.2.3
λ	lambda	Mean number of events in a Poisson distribution (also an eigenvalue in §16.5)	§5.7.1.3
ρ	rho	Population correlation (§11.2); the intraclass correlation in the design effect (§6.2.6); ρ_s is Spearman's rank correlation (§11.3)	§6.2.6, §11.2
τ, τ^2	tau	Kendall's rank correlation (§11.3); the between-study variance in meta-analysis (§15.7.3)	§11.3
κ	kappa	Cohen's kappa, agreement beyond chance	§16.3.3.1
ϕ	phi	The phi coefficient for two binary variables	§11.4
Φ	phi (capital)	The probability that a standard Normal value falls below a given point	§8.4

Letter	Name	Meaning in this book	First explained
Δ	delta (capital)	A true difference between groups, used in power and sample size	§8.4
δ	delta	A small step or change	§3.5
ε	epsilon	The error term of a regression model (§12.2.1); a relative margin of error (§7.3)	§7.3, §12.2.1
θ	theta	The dispersion parameter some software reports for a negative binomial model, θ = 1/α (§13.6); the true effect all studies share in a meta-analysis (§15.7.4)	§13.6, §15.7.4
π	pi	The constant 3.14159… (this book writes a population proportion as p, not π)	§1.3
ψ	psi	The baseline odds within a matched set	§13.2.9

Statistics – Foundations of Data

Statistics does not begin with a formula. It begins with a recorded observation: a **datum** (plural *data*) is any recorded value describing a defined unit under study. This chapter covers what must be settled before any arithmetic: what a variable is and how it is measured (§1.1), whom the data describe (§1.2–1.3), how a raw list becomes a table (§1.4), where Indian health data come from (§1.4.4), which study design produced them (§1.5), and how they are coded, entered and cleaned in software (§1.6, which ends with an optional R primer).

1.1 What Is a Variable

DEFINITION

The **unit of observation** is the thing each record describes: one child, one household, one village. A **variable** is a characteristic of that unit that can take *different values* across the units being studied. A characteristic that is the same for every unit is a **constant**, not a variable, and it carries no statistical information.

SIMPLIFIED EXPLANATION

Think of a health worker's register at a sub-centre. Each **row** is one person who walked in, the unit of observation. Each **column** (age, sex, vaccination status, haemoglobin) is a variable. One cell, one unit's value on one variable, is one datum.

Not everything you could record is useful. If every person in your study lives in Chhattisgarh, "state of residence" distinguishes no one from anyone else. It is a constant. Age, sex and blood pressure differ from person to person, and that difference is exactly what statistics analyses. No variation, no statistics: this idea returns as the logic behind variance (roughly, the average squared distance of values from their mean, §3.6).

ILLUSTRATION / EXAMPLE

At an under-five clinic you record, for every child: age in months, sex, nutritional status (normal / moderate wasting / severe wasting) and weight in kilograms. The unit is one child; the four characteristics are four variables. Before you can honestly say "the mean weight (the ordinary average, §2.2.1) of the children seen today was 11.8 kg", you must have agreed, for every child, that weight was measured the same way, on the same scale, in the same units. Skip that agreement and the mean is meaningless however correctly you compute it.

Now suppose a study enrols only pregnant women attending an antenatal clinic. "Sex: female" is true for every unit, so within *this* dataset sex is a constant, even though it is a perfectly good variable in a general

outpatient (OPD) register. Whether a characteristic is a variable depends on whether it varies *within the units you are studying*.

MUST-KNOW NOTES (MD-LEVEL)

- Variables are the columns of a dataset; units are the rows; one cell is one datum.
- A characteristic identical across all units in your dataset is a constant and adds nothing to an analysis.
- Analysis can rarely repair a variable that was badly defined or measured ("garbage in, garbage out").
- The type of a variable limits which summaries and tests are legitimate. Variables are classified two ways, by kind of value (§1.1.1) and by scale of measurement (§1.1.2). The two classifications overlap but are not identical.

1.1.1 Qualitative vs Quantitative Data

DEFINITION

A variable can vary in **kind** (a category) or in **amount** (a quantity). The first is **qualitative** (categorical) data. The second is **quantitative** (numerical) data, which splits into **discrete** data (separate, countable values, often whole numbers) and **continuous** data (any value within a range).

SIMPLIFIED EXPLANATION

Sort a basket of fruit two ways. By *type* (apple, banana, mango) you group things by kind, and "apple" is neither more nor less than "banana". By *weight in grams*, 150 g really is more than 100 g, and adding or averaging the numbers makes sense.

Ask of any variable: does it describe a *category* the unit falls into, or an *amount* the unit has? If a category, arithmetic is meaningless: averaging blood groups, (A + O) ÷ 2, produces nonsense even though a calculator will accept the codes. If an amount, arithmetic is legitimate.

Within amounts there is a second split. Some quantities move in separate steps: number of children, number of visits. You cannot have 2.5 children, so the variable is **discrete**. The steps need not be whole numbers (a score recorded in half-points is still discrete); what matters is that the possible values can be listed one by one. Others can fall anywhere on a scale: 62.3 kg is as valid as 62 kg. That is **continuous**.

ILLUSTRATION / EXAMPLE

In the under-five register, nutritional status is **qualitative** (three categories). Weight in kilograms is **quantitative and continuous**: 11.2 kg sits meaningfully between 11 and 12. Number of diarrhoeal episodes in the past month is **quantitative and discrete**: 0, 1, 2, 3 or 4 in this register, never 1.5. All three describe the same children; type is a property of each variable, not of the dataset.

MUST-KNOW NOTES (MD-LEVEL)

- **Qualitative:** attributes or categories; sums and means are not meaningful.

- **Quantitative:** measurable amounts; arithmetic is meaningful. **Discrete:** separate, countable values (pregnancies, teeth). **Continuous:** any value within an interval (height, haemoglobin, blood pressure).
- A continuous quantity is always *recorded* to finite precision (weight to 0.1 kg), but it stays continuous in nature. Classify by what the quantity is, not by how many decimals the scale shows.
- This classification differs from the scales of measurement (§1.1.2), though qualitative data usually sit on nominal or ordinal scales and quantitative data on interval or ratio scales. Exams test both framings.
- Misclassifying a variable leads straight to wrong analyses: a "mean" of a categorical variable, or a t-test (a test comparing two means, §9.2) run on category codes.

CHECKPOINT 1.1

1. Explain, without using the word "variable", why "sex" might be a variable in one dataset but a constant in another.
2. A field worker records "type of latrine used" and later reports its "mean" across households. What is wrong with that statement, and what should be reported instead?
3. The clinic register holds 100 children and 11 columns. How many data points (cells) does it contain? Classify "age in completed months" and "number of diarrhoeal episodes in the past month" as discrete or continuous, with a reason.

(Answers in the Appendix at the end of the book.)

1.1.2 Scales of Measurement

DEFINITION

Scales of measurement classify a variable by which operations are meaningful on its values: equality, ordering, subtraction and division. There are four scales, each permitting everything the one below it permits plus one more operation: nominal → ordinal → interval → ratio.

SIMPLIFIED EXPLANATION

Think of a staircase of permissions:

- **Nominal**: you may only check whether two values are the *same* or *different*.
- **Ordinal**: you may also say one value is *higher or lower* than another.
- **Interval**: you may also *subtract*, because the gaps between values are equal.
- **Ratio**: you may also *divide*, because there is a true zero, so "twice as much" is a real statement.

Skip a step and you are doing invalid arithmetic. The classic slip is a ratio on an interval scale: 20°C is not "twice as hot" as 10°C, because 0°C is an arbitrary reference point, not "no heat".

ILLUSTRATION / EXAMPLE

- Blood group (A, B, AB, O): **nominal**. You can group, not order.

- Pain severity (mild / moderate / severe): **ordinal**. Severe is worse than mild, but not by a fixed amount.
- Temperature in Celsius: **interval**. The 10° gap between 10°C and 20°C equals the gap between 20°C and 30°C, but 0°C is not the absence of temperature.
- Body weight in kg: **ratio**. 0 kg means no mass, so 20 kg really is twice 10 kg.

MUST-KNOW NOTES (MD-LEVEL)

- Nominal < ordinal < interval < ratio; each includes every operation of the ones before it.
- The scale guides the choice of method, but it does not settle it alone. Nominal data are analysed with categorical-data methods: counts and percentages, the chi-square test (§10.6), exact tests (§10.7) and logistic regression (§13.2). Ordinal data suit rank-based methods (§10.1–10.5) or models for ordered categories (§13.2.7). Interval and ratio data *permit* means, standard deviations (SDs, a measure of spread, §3.7) and tests built on them, but whether such a test is appropriate also depends on its own assumptions, for example about the shape of the data's distribution (§9.2.4, §10.1).
- A **single Likert item** (strongly disagree … strongly agree) is ordinal. A **summed Likert scale** (the total of many items) takes many values and is routinely analysed with means and parametric methods (methods that assume a distribution shape, such as the normal; §9.2). Norman (2010) reviewed studies going back to the 1930s showing that such methods are robust for Likert data. Know the strict position for exams (median for an ordinal item) and the defensible practice for summed scales.
- An interval scale has an *arbitrary* zero (ratios invalid); a ratio scale has a *true* zero (ratios valid). That single distinction separates the two.

1.1.2.1 Nominal Scale

DEFINITION

A **nominal scale** classifies values into named categories with no inherent order. The only meaningful operation is checking whether two units fall in the *same* category or *different* ones.

SIMPLIFIED EXPLANATION

A nominal value is a label. Knowing someone's blood group is A tells you nothing about "more" or "less" than group B; it tells you only that they differ from a person with B and match a person with A.

Numbers assigned for convenience in a database (Male = 1, Female = 2) are still labels. The "2" does not mean twice anything, and averaging the codes gives a number with no direct meaning as a measurement.

ILLUSTRATION / EXAMPLE

Blood group, religion, marital status, sex and diagnosis codes are nominal. In the community-based health insurance (CBHI) household survey, "reason for not joining" (lack of knowledge / lack of money / lack of trust / bad image / other) is nominal: five distinct categories with no ranking.

A nominal variable can be **binary** (dichotomous: two categories, e.g. CBHI member yes/no) or **polychotomous** (more than two, e.g. religion).

MUST-KNOW NOTES (MD-LEVEL)

- Only valid operation: equality or inequality. No ordering, no arithmetic.
- The **mode** (the most frequent category, §2.4) is the only meaningful measure of central tendency.
- Summarise with frequencies, proportions and percentages.
- Association between two nominal variables is tested with the chi-square test of independence (§10.6.1); when *expected* cell counts are small (the rule is given in §10.6.1), Fisher's exact test (§10.7) is used instead. The criterion is small expected counts, not a small sample as such.
- Numeric codes carry no magnitude. The one exception worth knowing: for a binary variable coded 0/1, the mean of the codes equals the proportion coded 1.

1.1.2.2 Ordinal Scale

DEFINITION

An **ordinal scale** classifies values into categories with a meaningful *order*, but the gaps between adjacent ranks are unequal or undefined.

SIMPLIFIED EXPLANATION

Ordinal data add one capability to nominal data: you can say one value is higher or lower than another. You still cannot say *by how much*. Runners finishing 1st, 2nd and 3rd are ordered, but the ranks do not tell you whether 1st beat 2nd by a second or an hour, nor that the 1st–2nd gap equals the 2nd–3rd gap.

ILLUSTRATION / EXAMPLE

Cancer staging (Stage I–IV) is ordinal: Stage III is worse than Stage II, but the jump from I to II need not equal the jump from III to IV. Socioeconomic class on the modified Kuppuswamy scale (classes such as upper-middle, lower-middle, upper-lower and lower) is ordinal (Naaz et al., 2025). So is a single Likert-type response such as the CBHI survey's 1-to-5 satisfaction item.

MUST-KNOW NOTES (MD-LEVEL)

- Valid operations: equality and ordering. Subtraction is not strictly valid, because gaps are not guaranteed equal.
- The **median** (the middle value, §2.3) and the mode are the strictly appropriate measures of central tendency. Means of single ordinal items are common in applied work; know that this is a compromise (§1.1.2 on summed scales).
- Rank-based methods (Mann–Whitney, Wilcoxon signed-rank, Kruskal–Wallis, Spearman correlation; §10.2–10.4, §11.3) use only the order of the values. Ordinal (cumulative-link) logistic regression (§13.2.7) models ordered categories directly and is often preferred once several predictors are involved.

1.1.2.3 Interval Scale

DEFINITION

An **interval scale** has equal, meaningful gaps between values, so subtraction is valid, but its zero point is *arbitrary*, not a true absence of the quantity.

SIMPLIFIED EXPLANATION

Interval data fix the unequal gaps of ordinal data: differences are now real quantities. The catch is zero. On an interval scale, zero is a reference point someone chose, so ratios stop making sense even though differences still do.

In Celsius, 0°C is the freezing point of water. So 20°C − 10°C = 10°C is meaningful, but "20°C is twice as hot as 10°C" is not. Convert to Fahrenheit (68°F vs 50°F) and the "ratio" changes, which shows it was never a property of the heat itself.

ILLUSTRATION / EXAMPLE

Celsius and Fahrenheit temperatures are the textbook interval examples. The **Kelvin** scale measures the same quantity from absolute zero, so Kelvin is a ratio scale: 20°C and 10°C are 293.15 K and 283.15 K, a ratio of 1.035, nowhere near 2. The zero, not the quantity, makes the scale. IQ is another interval example: it is anchored at a population mean of 100, not at "zero intelligence", so IQ 150 is not "1.5 times as intelligent" as IQ 100, though the 50-point difference is meaningful.

MUST-KNOW NOTES (MD-LEVEL)

- Valid operations: equality, order, subtraction. Division is not valid.
- Means and SDs are legitimate, because differences are meaningful.
- Interval-only variables are uncommon in clinical and community data. Almost every field measurement (weight, height, blood pressure, counts, durations) has a true zero and is therefore ratio-scaled.

1.1.2.4 Ratio Scale

DEFINITION

A **ratio scale** has equal, meaningful gaps *and* a true zero representing a genuine absence of the quantity, so division and ratio statements ("twice as much") are meaningful.

SIMPLIFIED EXPLANATION

Ratio data fix the one remaining problem of interval data. Because 0 kg truly means no mass, 20 kg really is twice 10 kg. Every operation from the lower scales (equality, order, subtraction) is available, plus division. There is no fifth level.

ILLUSTRATION / EXAMPLE

Body weight, height, blood pressure, haemoglobin, number of children, age and duration of illness are ratio-scaled: in each, zero means none of the quantity. A 20 kg child weighs twice as much as a 10 kg child, a statement that fails for Celsius temperature.

MUST-KNOW NOTES (MD-LEVEL)

- Valid operations: equality, order, subtraction, division.
- Percentage-change statements become meaningful ("Group A's mean weight is 20% higher than Group B's").
- Being ratio-scaled permits means, SDs and methods such as the t-test (§9.2), ANOVA (§9.5), correlation (Chapter 11) and regression (Chapter 12); it does not guarantee they are appropriate. Each method has its own assumptions (independence, distribution shape, linearity, variance structure), listed where the method is taught (e.g. the assumptions of linear regression, §12.3).

CHECKPOINT 1.2

1. Without naming the four scales, describe the one extra "permission" each step up the measurement ladder unlocks.
2. A colleague says "0°C means there is no heat." Which scale is she confusing Celsius with? Convert 10°C and 20°C to kelvin and compute the true ratio of the two temperatures.
3. In a dataset of 10 children, sex is coded Male = 1, Female = 2, and the "mean sex" comes out as 1.4. What, if anything, does 1.4 tell you?

(Answers in the Appendix at the end of the book.)

1.2 Population vs Sample

DEFINITION

A **population** is the complete, explicitly defined set of units about which you want to draw conclusions. A **sample** is the subset of that population you actually observe. **Primary data** are data you collect yourself for your question; **secondary data** were collected by someone else, usually for another purpose (a register, a survey dataset, a programme report), and you reuse them.

SIMPLIFIED EXPLANATION

Suppose you want the mean haemoglobin of every pregnant woman in Chhattisgarh. Examining all of them is too slow and too expensive, so you examine a manageable subset and use it to infer about the whole. That gap between what you want to know and what you can measure is why inferential statistics exists. A **census** (every unit measured) removes the need for sampling-based inference, but not for statistics: measurement error, change over time and questions of cause remain.

Whether primary or secondary, data become useful in stages. **Data** are raw recorded values (each antenatal care (ANC) visit's haemoglobin). **Information** is data organised and summarised so that it an-

swers a question ("38% of women seen this quarter had haemoglobin (Hb) below the cut-off"). **Intelligence** is information interpreted and combined with context so that someone can act on it ("the rise is concentrated in two blocks where iron–folic acid supply broke down; redirect stock there").

ILLUSTRATION / EXAMPLE

Population: all children under 5 in a district during a given year. Sample: the children who happened to visit one sub-centre this month. This sample may represent the population badly, because children who never visit a facility are excluded by design. Whether a sample represents its population is the subject of Chapter 6.

A resident who weighs 200 children in a village survey collects primary data. A resident who analyses the district's HMIS reports (§1.4.4) uses secondary data: cheap and large, but the variables, definitions and quality were fixed by someone else.

MUST-KNOW NOTES (MD-LEVEL)

- Define the population before sampling: "all children under 5 in District X during 2025" is defined; "children" is not.
- Numbers computed from a sample estimate the corresponding (usually unknown) population numbers; §1.3 names them.
- A census is the case where the sample equals the population. It is rare because of cost, though national censuses are done (§1.4.4).
- Secondary data: check who collected them, for what purpose, with which definitions and denominators, and how complete they are, before analysing them.

1.3 Parameter vs Statistic

DEFINITION

A **parameter** is a numerical characteristic of a population: fixed and usually unknown. A **statistic** is a numerical characteristic computed from a sample: known, and different from one sample to the next.

SIMPLIFIED EXPLANATION

Once you study a sample instead of the population, you need two names: the true number you want but cannot see (the parameter), and the number you can calculate from the data in front of you (the statistic). By convention, parameters get Greek letters (μ for a population mean, σ for a population standard deviation) and statistics get Roman letters or a "hat" ($\bar{x}$, s, $\hat{p}$). The population proportion is the usual exception: some texts write it π, but this book, like many, writes it p, with $\hat{p}$ for the sample proportion, because π is also the constant 3.14....

ILLUSTRATION / EXAMPLE

The true mean birthweight of every newborn in Raipur district this year is a parameter, μ. No one knows it exactly, because no one has weighed every newborn. The mean computed from 200 newborns you

measured is a statistic, $\bar{x}$. A different 200 newborns would give a slightly different $\bar{x}$, but there is only one μ.

MUST-KNOW NOTES (MD-LEVEL)

- **Parameter**: describes the population; fixed; usually unknown; Greek letters (μ, σ), except the proportion, p in this book.
- **Statistic**: computed from a sample; known; varies from sample to sample; Roman letters or hats ($\bar{x}$, s, $\hat{p}$).
- Statistics estimate parameters. Estimation (Chapter 7) and hypothesis testing (Chapter 8) quantify how far a statistic can be trusted as a stand-in for its parameter.
- The sample-to-sample variability of a statistic is formalised as its sampling distribution (§6.5).

CHECKPOINT 1.3

1. A colleague says "if you sampled correctly, the sample mean should equal the population mean exactly." Is that true?
2. Give a population and a sample for a study you might run as a Community Medicine resident, and name which of your numbers is the parameter and which the statistic.
3. In the CBHI survey, 150 of 500 sampled households are members. Compute the membership proportion, give its symbol, and say whether it is a parameter or a statistic. Is the survey dataset primary or secondary data for a resident who reanalyses it five years later?

(Answers in the Appendix at the end of the book.)

1.4 Organising Raw Data

DEFINITION

Raw data are data exactly as recorded: an unsorted, unsummarised list of observations. **Organising** raw data means sorting them and counting how many fall at each value or in each range, usually as a frequency distribution (§1.4.1).

SIMPLIFIED EXPLANATION

Here are the weights (kg) of the 30 children seen at the under-five clinic today, in the order they walked in:

11.2, 9.8, 14.1, 10.0, 12.3, 13.6, 11.1, 11.9, 12.6, 9.4, 13.8, 15.3, 13.4, 12.9, 9.1, 9.7, 11.6, 10.3, 11.5, 8.6, 14.8, 10.5, 11.0, 12.4, 11.8, 10.8, 12.0, 13.1, 11.4, 14.4.

Their mean is 11.8 kg, but the list itself is hard to read. Sorting it and counting how many fall in each range is the first step towards seeing its shape.

ILLUSTRATION / EXAMPLE

Sorted, the same 30 weights run from 8.6 kg to 15.3 kg. Counting them into 2-kg ranges (§1.4.1) turns a wall of 30 numbers into a four-row table.

MUST-KNOW NOTES (MD-LEVEL)

- Raw data = data as collected, with no summarising applied.
- Many statistics (the mean, the median) can be computed straight from the raw list; a frequency table is a convenient precursor, and it is *required* only for grouped-data formulas and for charts such as the histogram (§4.3).

1.4.1 Frequency Distribution

DEFINITION

A **frequency distribution** is a table of values (or ranges of values, called **class intervals**) alongside the number of observations, the **frequency**, in each.

SIMPLIFIED EXPLANATION

Instead of listing 30 weights, count how many fall in each 2-kg class. Patterns invisible in the raw list, such as where most values cluster, become visible at a glance.

ILLUSTRATION / EXAMPLE

Weight (kg)	Frequency
8 to <10	5
10 to <12	12
12 to <14	9
14 to <16	4
Total	30

Each class includes its lower end and excludes its upper end, so the child weighing exactly 10.0 kg is counted in "10 to <12", and the child weighing 12.0 kg in "12 to <14". Most children today weighed between 10 and 14 kg.

DERIVATION

How many classes? Sturges (1926) started from an ideal distribution of $N = 2^{k-1}$ items spread over k classes in proportion to the numbers in a row of Pascal's triangle (the binomial coefficients, §5.7.1.2): 1; 1, 1; 1, 2, 1; 1, 3, 3, 1; 1, 4, 6, 4, 1; ..., each number the sum of the two above it (for 16 items and 5 classes: 1, 4, 6, 4, 1). Each number feeds two numbers in the row below, so each row adds up to twice the row above: 1, 2, 4, 8, 16. The row with k numbers therefore adds up to 2^{k-1}, so

$$N = 2^{k-1} \Longrightarrow k - 1 = \log_2 N \Longrightarrow k = 1 + \log_2 N = 1 + \log_{10} N \div \log_{10} 2 = 1 + 3.322 \log_{10} N$$

where ⟹ reads "which gives". The class width is roughly the range divided by k.

WORKED CALCULATION

1. $\log_{10}30 = 1.477$.
2. $3.322 \times 1.477 = 4.907$.
3. $k = 1 + 4.907 = 5.91$, so Sturges suggests about 6 classes.
4. Range = 15.3 − 8.6 = 6.7 kg; width = 6.7 ÷ 5.91 = 1.13 kg.

Result: Sturges' rule suggests 6 classes of about 1.1 kg; the table above uses 4 classes of 2 kg. Both are defensible. With only 30 children, 6 classes would leave several classes with 1–3 children each, and 1.13-kg limits are awkward to read; 2-kg classes with whole-number limits are easier to read and compare with other clinics. Sturges' rule is a starting point, not a requirement.

MUST-KNOW NOTES (MD-LEVEL)

- Two forms exist: **ungrouped** (a frequency for each distinct value) and **grouped** (a frequency for each class interval); see §1.4.2.
- Classes must be mutually exclusive (no value fits two classes) and exhaustive (every value fits one). Equal widths are the usual convention and make a histogram easy to read; unequal classes, such as the WHO anaemia bands of §1.4.2, are legitimate when a clinical classification demands them, but a histogram must then plot frequency density (§4.3.1).
- Sturges' rule, $k = 1 + 3.322 \log_{10}N$ (Sturges, 1926), is a convention; it assumes a roughly symmetric, bell-shaped spread, and readability or a clinical classification may justify a different k.
- A frequency distribution is the input for the histogram (§4.3) and for grouped-data formulas for the mean, median and mode (Chapter 2).

1.4.2 Grouped vs Ungrouped Data

DEFINITION

Ungrouped data list every distinct value with its own frequency. **Grouped data** put ranges of values into class intervals and report the frequency per interval, giving up some detail for a readable table. Each class interval has **class limits** (the smallest and largest values that can be *recorded* in it) and **class boundaries** (the points on the continuous scale that separate adjacent classes, halfway between one class's upper limit and the next class's lower limit).

SIMPLIFIED EXPLANATION

If a variable takes a handful of values, such as number of ANC visits (0, 1, 2, 3, 4 ...), list each value: that is ungrouped. If it is continuous or has many distinct values, such as weight to 0.1 kg across 500 children, bucket the values into ranges: that is grouped.

Limits and boundaries differ because measurements are rounded. Haemoglobin recorded to 0.1 g/dL might be grouped as 7.0–9.9 and 10.0–10.9: those are class *limits*. A true value of 9.96 g/dL is recorded as 10.0, so the real dividing line between the two classes is 9.95: that is the class *boundary*. The class 7.0–9.9 has boundaries 6.95 and 9.95, width 3.0, and **class mark** (mid-point) (6.95 + 9.95) ÷ 2 = 8.45 g/dL.

ILLUSTRATION / EXAMPLE

Ungrouped: number of diarrhoeal episodes in the past month among the 100 children of the clinic register: 0 episodes, 41 children; 1, 35; 2, 16; 3, 5; 4, 3.

Grouped: haemoglobin grouped into WHO's severity bands for children aged 24–59 months: severe <7.0 g/dL, moderate 7.0–9.9, mild 10.0–10.9, no anaemia ≥11.0 (WHO, 2024). Here grouping mirrors a clinical classification. The bands depend on the population: WHO's 2024 guideline gives different cut-offs for children 6–23 months (anaemia below 10.5 g/dL), for non-pregnant women (below 12.0) and for men (below 13.0), and by trimester in pregnancy. Always state the population when you quote a cut-off.

MUST-KNOW NOTES (MD-LEVEL)

- Choose grouped or ungrouped by how many distinct values the variable takes, not by whether it is qualitative or quantitative.
- Grouping loses information: from the table alone you know a reading lies in 7.0–9.9, not its exact value.
- Grouped-data calculations use the **class mark** as a stand-in for every value in the class; the approximation is better when classes are narrow.
- Write classes so that no value is ambiguous: either "8 to <10" (half-open) or limits that do not overlap ("8.0–9.9, 10.0–11.9"). A table reading "8–10, 10–12" leaves 10.0 with two homes.
- Strictly, a weight recorded as 10.0 kg stands for anything from 9.95 to 10.05 kg, so the true boundaries of "10 to <12" in recorded values are 9.95 and 11.95. This book's grouped calculations ignore that 0.05-kg shift and use the mid-points 9, 11, 13 and 15 kg.

1.4.3 Relative and Cumulative Frequency

DEFINITION

Relative frequency is a class's frequency divided by the total number of observations. **Cumulative frequency** is the running total of frequencies up to and including a class.

SIMPLIFIED EXPLANATION

"What fraction of the children are in this class?" is relative frequency. "How many are in this class or below?" is cumulative frequency. Relative frequency lets you compare a sample of 30 with a sample of 3,000; cumulative frequency is the basis of percentiles and the ogive (§4.3.3, a cumulative frequency curve).

ILLUSTRATION / EXAMPLE

Weight (kg)	f	Relative f (%)	Cumulative f	Cumulative relative f (%)
8 to <10	5	16.7	5	16.7

Weight (kg)	f	Relative f (%)	Cumulative f	Cumulative relative f (%)
10 to <12	12	40.0	17	56.7
12 to <14	9	30.0	26	86.7
14 to <16	4	13.3	30	100.0

WORKED CALCULATION

1. Relative frequency of 10 to <12 kg = 12 ÷ 30 = 0.40 = 40%.
2. Cumulative frequency up to and including 10 to <12 kg = 5 + 12 = 17.
3. Cumulative relative frequency = 17 ÷ 30 = 56.7%.

Result: 17 of the 30 children (56.7%) weigh under 12 kg.

MUST-KNOW NOTES (MD-LEVEL)

- Relative frequencies sum to 1 (100%), apart from rounding.
- The last cumulative frequency equals N.
- Read cumulative frequency at a class *boundary*: "17 children weigh under 12 kg", not "12 kg or less".
- Cumulative relative frequency is the basis for the grouped median, quartiles and percentiles (§2.3, §3.4) and for the ogive.

CHECKPOINT 1.4

1. List, in order, the steps that turn a raw, unsorted list of 200 haemoglobin readings into something you could read a percentile from.
2. Why group haemoglobin readings but not "number of children per household"?
3. Using the diarrhoea table in §1.4.2 (0: 41, 1: 35, 2: 16, 3: 5, 4: 3), compute the cumulative frequency up to 1 episode and the percentage of children with 2 or more episodes.

(Answers in the Appendix at the end of the book.)

1.4.4 Sources of Health Data in India

DEFINITION

A **source of health data** is a system that routinely or periodically records health-related events or characteristics for a defined population. Sources differ in what they count, which population forms the **denominator** (the group the count is divided by to make a rate or proportion), how often they report, and how their data can go wrong. **Data quality** is judged on dimensions that include **completeness** (did every unit that should report, report?), **timeliness** (did reports arrive by the deadline?), **internal consistency** (do the numbers agree with each other and over time?) and **external**

comparison (do they agree with other sources, such as surveys?) (WHO, 2020). Validity, whether a recorded value measures what it claims to (Chapter 16), sits underneath all four.

SIMPLIFIED EXPLANATION

Most numbers a Community Medicine resident quotes come from secondary data (§1.2), so you must know each source's denominator and blind spots. A useful split: **counting systems** try to record every event (census, civil registration, surveillance); **sample surveys** measure a probability sample (units chosen by a random mechanism, Chapter 6) and estimate the rest (SRS, NFHS); **facility systems** record people who reached a facility (HMIS, hospital registries). Counting systems fail through under-registration; surveys through sampling error (chance differences between sample and population, §6.4) and non-response; facility systems through the people who never came.

ILLUSTRATION / EXAMPLE

Source	What it measures	Unit and denominator	How often	Known weakness
Census of India	Population size and composition, down to village level	Every household; the denominator for most rates	Decennial; 2011 is the latest census with published tables	Denominators for later years must be projected, and projections drift as years pass from the census
Civil Registration System (CRS)	Births, deaths, stillbirths; compulsory under the Registration of Births and Deaths (RBD) Act, 1969	Every registered event, at the place of occurrence	Continuous; annual report	Completeness and promptness vary by state
Sample Registration System (SRS)	Birth rate, death rate, infant mortality rate (IMR), other fertility and mortality indicators	Continuous enumeration plus a six-monthly independent survey in sample villages and urban blocks	Annual estimates	Sample-based: sampling error; estimates are for the nation, states and natural divisions, not districts
Medical Certification of Cause of Death (MCCD)	Medically certified cause of death, coded by the International Classification of Diseases (ICD)	Certified deaths only	Annual report	Only 22.3% of registered deaths were certified in 2022
National Family Health Survey (NFHS)	Fertility, child health, nutrition, anaemia, blood pressure and blood glucose	Probability sample of households, women 15–49, men 15–54	Periodic (NFHS-5 fieldwork June 2019–April 2021)	Survey design must be used in analysis (§15.2.6)
Health Management Information System (HMIS)	Service delivery (ANC, deliveries, immunisation) reported by facilities	Service users at reporting facilities	Monthly, web-based portal since 2008	Denominator is service users, not the population; reporting by private facilities has needed specific effort to improve

Source	What it measures	Unit and denominator	How often	Known weakness
Integrated Disease Surveillance Programme (IDSP / IHIP)	Epidemic-prone diseases: Syndromic (S), Presumptive (P) and Laboratory (L) forms	Auxiliary nurse-midwives (ANMs) at sub-centres (S), medical officers (P), laboratories (L)	Continuous, case-based entry on the IHIP portal	Captures only people who reach a sub-centre, facility or laboratory
National Cancer Registry Programme (NCRP)	Cancer incidence, mortality and patterns	Population-based registries (PBCRs: defined area) and hospital-based registries (HBCRs: patients of one hospital)	Periodic reports	Registry areas cover part of the country; HBCRs describe patients, not populations

Sources: Census of India; Office of the Registrar General, India (ORGI) (2022, CRS report; 2022, MCCD report; 2020, SRS Statistical Report); Data for India (2025, CRS and SRS); DHS Program (NFHS-5, India's Demographic and Health Survey); Meghani et al. (2022, HMIS); Singh et al. (2023, IDSP-IHIP); ICMR-NCDIR (2020).

WORKED CALCULATION

A block has 40 sub-centres expected to send their monthly HMIS report. By the deadline 29 reports have arrived; 5 more arrive late; 6 never arrive.

1. Reports received = 29 + 5 = 34. Completeness = 34 ÷ 40 = 85.0%.
2. On-time reports = 29. As a share of expected reports: 29 ÷ 40 = 72.5%. As a share of reports received: 29 ÷ 34 = 85.3%.
3. State which denominator you used; tools differ, and the two figures answer different questions.

Result: completeness 85.0%, timeliness 72.5% of expected reports. Any block total built on these reports undercounts services by whatever the 6 missing sub-centres delivered.

A second calculation shows why a denominator matters. In 2022, CRS recorded 86,49,930 deaths, and 22.3% of them were medically certified: 0.223 × 86,49,930 ≈ 19.3 lakh deaths with a certified cause. Cause-of-death proportions from MCCD describe those 19.3 lakh deaths, not the remaining 77.7% of registered deaths or any unregistered ones.

MUST-KNOW NOTES (MD-LEVEL)

- Ask of every source: who is counted, who is the denominator, how often, and who is missed.
- CRS aims to count every event (ORGI, 2022); SRS samples to give reliable annual estimates of birth rate, death rate and IMR (ORGI, 2020). Rates from the two can differ.
- HMIS and IDSP counts are facility-based: a rise can mean more disease, more care-seeking, or better reporting. Check completeness before reading a trend.
- NFHS is a complex survey (strata, clusters, weights); unweighted analysis of its data can be biased (§15.2.6).
- Registries: a PBCR can give incidence rates for its area; an HBCR cannot, because it has no population denominator.

CHECKPOINT 1.5

1. Why can a hospital-based cancer registry describe the pattern of cancers among its patients but not the incidence of cancer in the district?
2. A district's IDSP data show acute diarrhoeal disease cases doubling this month. Give three explanations other than a true doubling of disease.
3. In a month, 50 primary health centres (PHCs) should report; 44 reports arrive, 36 of them by the deadline. Compute completeness and timeliness (as a share of expected reports).

(Answers in the Appendix at the end of the book.)

1.5 Study Designs – A First Map

DEFINITION

A **study design** is the plan that decides which units are studied, how they are selected, and when exposure and outcome are measured. An **exposure** is a factor whose effect or association you study (e.g. not being exclusively breastfed); an **outcome** is the event or condition you measure (e.g. diarrhoea). **Descriptive** designs (case report, case series, cross-sectional survey) describe how much disease there is and in whom. **Analytical** designs compare groups to study an exposure–outcome association: **ecological**, **case-control**, **cohort** and the **randomised controlled trial (RCT)**. **Hybrid** designs sample from within a cohort: the **nested case-control** study and the **case-cohort** study.

SIMPLIFIED EXPLANATION

The design decides which numbers the data can give you. Four measures recur:

- **Prevalence**: the proportion of a population that has the condition at one time (existing cases ÷ population).
- **Risk** (incidence proportion): the proportion of initially disease-free people who develop the condition over a stated period (new cases ÷ people at risk at the start).
- **Rate** (incidence rate): new cases per unit of person-time (the time each person was observed while at risk, added up), e.g. per 1,000 child-years (§7.4 treats rates fully).
- **Odds**: the probability of an event divided by the probability of it not happening, p ÷ (1 − p) (used for 2 × 2 tables in §10.6.4).

A **cross-sectional** survey measures exposure and outcome at the same moment, so it gives prevalence, but it cannot tell which came first. A **cohort** study starts with disease-free people, classifies them by exposure, and follows them forward (or, in a retrospective cohort, through past records) to count new cases, so it gives risk and rate. A **case-control** study starts from the outcome: it selects people with the disease (cases) and without it (controls) and looks back for past exposure. The investigator fixes how many cases and controls there are, so the study cannot estimate prevalence or risk; it compares the *odds of exposure* between cases and controls. An **RCT** is a cohort in which the investigator assigns the exposure at random, which balances other factors between groups on average (§15.9). An **ecological** study uses groups (villages, districts) as units, relating a group's exposure level to its disease rate.

The hybrid designs save money inside a large cohort. A **nested case-control** study takes every case that arises in the cohort and, for each, samples controls from cohort members still at risk at that time. A **case-cohort** study takes a random subcohort at the start and adds all cases; the same subcohort serves as the comparison group for every outcome.

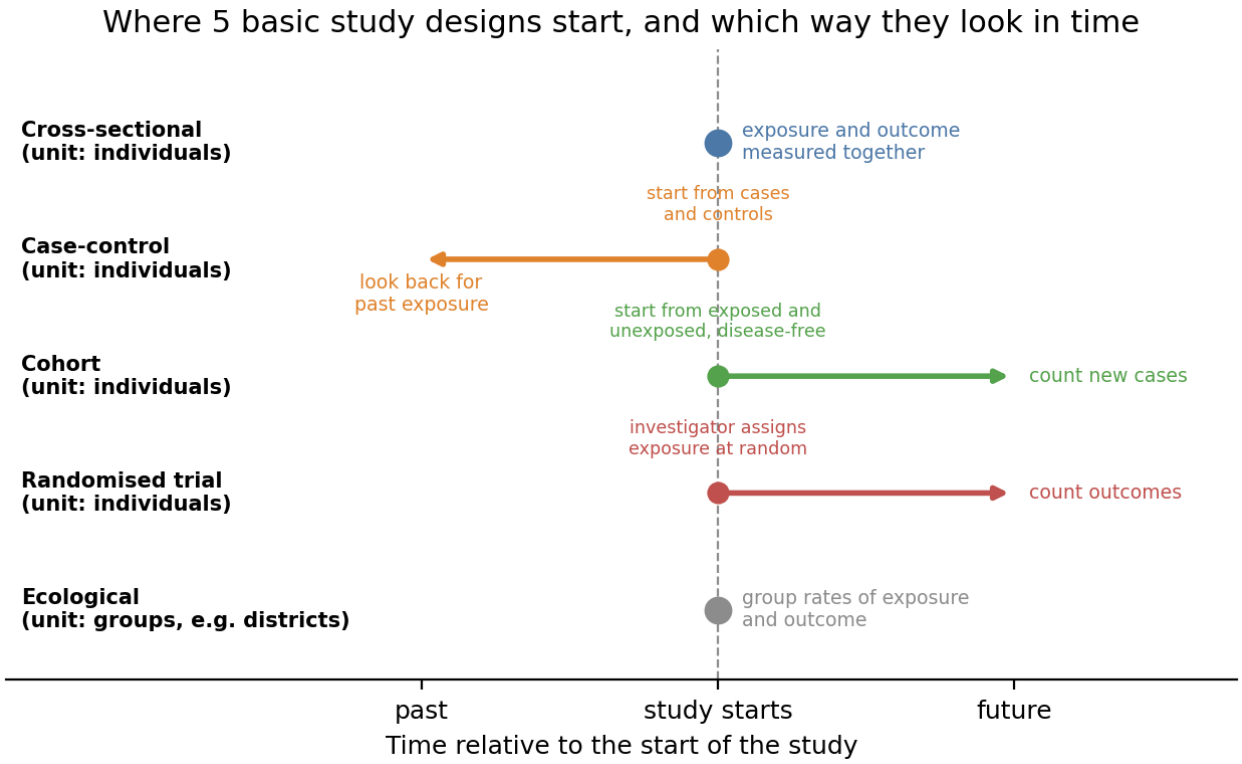


Figure 1.1. Where each basic design starts and which way it looks in time. A retrospective cohort also looks forward from exposure to outcome, but uses records made in the past.

ILLUSTRATION / EXAMPLE

Design	Unit	Sampled on	Can estimate	Analysed in
Case series	Individuals (cases only)	Outcome	Counts and features of cases; no rate	Chapters 2–4 (description)
Cross-sectional	Individuals	Neither (a population sample)	Prevalence; prevalence ratio or odds ratio	§7.2, §10.6, §13.7
Ecological	Groups	Group	Group rates and their correlation	§11.2, Chapter 12
Case-control	Individuals	Outcome	Odds ratio of exposure	§10.6.4, §10.8, §13.2
Cohort	Individuals	Exposure (or a population, then classified)	Risk, rate, risk ratio, rate ratio, hazard ratio (a ratio of instantaneous event rates, Chapter 14)	§7.4, §10.6.4, §13.5, Chapter 14
RCT	Individuals (or clusters)	Assigned exposure	Risk or rate in each arm; effect of the assigned intervention	§15.9

Design	Unit	Sampled on	Can estimate	Analysed in
Nested case-control / case-cohort	Individuals within a cohort	Outcome, within the cohort	The full cohort's ratio measures (e.g. rate or hazard ratio), with exposure measured on far fewer people	§13.2.9, Chapter 14

The ecological trap is real. Robinson (1950) found that the correlation between illiteracy and the proportion of Black residents was 0.77 across the 48 US states but only 0.20 across individuals. Group-level association cannot be read as individual-level association: the **ecological fallacy**.

WORKED CALCULATION

A district has 10,000 children, of whom 100 (1%) had dysentery this season. An investigator runs a case-control study: all 100 cases and 200 controls.

1. Proportion with dysentery among the 300 studied = 100 ÷ 300 = 33.3%.
2. True proportion affected in the district = 100 ÷ 10,000 = 1.0%.
3. Had she chosen 400 controls, the study proportion would be 100 ÷ 500 = 20.0%, with the district unchanged.

Result: in a case-control study the "proportion diseased" (33.3% or 20.0%) is set by the investigator's control-to-case ratio. It says nothing about prevalence or risk, which is why case-control studies report odds ratios of exposure instead.

COMMON MISREADING

"The design with the most statistical machinery gives the strongest evidence." The analysis must match the design, not the other way round. A logistic regression on a cross-sectional survey still cannot tell whether exposure came before outcome; a perfectly analysed ecological correlation still cannot be applied to individuals.

MUST-KNOW NOTES (MD-LEVEL)

- Descriptive designs generate hypotheses; analytical designs test associations. Only the RCT assigns exposure, so only it balances unmeasured factors between groups, on average, by design.
- Case-control: sampled on outcome; efficient for rare diseases; gives an odds ratio of exposure, not prevalence or risk.
- Cohort: sampled on exposure or as a population; gives risk and rate; efficient for rare exposures; needs follow-up.
- Cross-sectional: gives prevalence; exposure and outcome measured together, so temporality is often unknown.
- Ecological: the unit is a group; beware the ecological fallacy (Robinson, 1950).
- §8.8 extends this map into a table of design × outcome type → test or model.

CHECKPOINT 1.6

1. A resident surveys 400 randomly chosen women in a block on one day, recording current tobacco use and current hypertension. Name the design and the measure of disease frequency it can give. Why can it not show that tobacco use preceded hypertension?
2. A case-control study of 80 neonatal tetanus cases uses 240 controls. What proportion of the study sample are cases? Why is this not the incidence of neonatal tetanus?
3. Districts with more mobile phones per household have lower infant mortality. Name the design and the error of concluding that a mother's phone lowers her own infant's risk of death.

(Answers in the Appendix at the end of the book.)

1.6 Data in Software – Codebooks, Entry, Cleaning and an R Primer

DEFINITION

A **codebook** (data dictionary) lists every variable in a dataset with its name, label, type, unit, allowed values or range, codes for categories, and codes for missing values. **Data entry** is the transfer of recorded values into a computer file, ideally through a form that checks each value as it is entered. **Data cleaning** is the systematic search for, and correction of, errors before analysis: values out of range, inconsistent combinations, duplicates and missing values. A **data frame** is a table held by software, with one row per unit and one column per variable (§1.1).

The codebook, entry and cleaning habits below apply whatever software you use, from Epi Info or Excel to SPSS, Stata or R. The R primer at the end of this section (the **In R (optional)** box) is optional: if you will not use R, skip it. No later chapter needs it.

SIMPLIFIED EXPLANATION

Every statistic in this book is only as good as the file it is computed from. Four habits protect that file.

1. **Write the codebook first.** Short, lower-case variable names without spaces (`weight_kg`, not `Weight of child (kg)`); the unit in the name or label; one column per variable; one row per unit.
2. **Code missing values as missing.** Leave the cell empty (read as `NA` in R). A code such as 999 is a number, and software will average it with the real weights unless told otherwise. Distinguish "not applicable" (a child not seen today has no arrival order) from "not recorded".
3. **Enter through validating forms.** Mobile tools such as Epicollect5, KoboToolbox and ODK let you set a range (weight 5–20 kg for this clinic), a type (number, date) and skip rules, so impossible values are refused in the field. For paper forms, **double data entry** (two people enter the same forms independently, then a program lists every cell where they disagree) catches any typing error that only one of the two people made.
4. **Clean with checks, not with your eyes.** Range checks, consistency checks (a child marked "not seen today" should have no arrival order), duplicate IDs, and a count of missing values in every column. Keep the raw file untouched and do all corrections in a script, so every change is recorded.

Software differs in menus, not in logic. In SPSS you set a variable's level (nominal, ordinal, scale) and its missing-value codes in Variable View; in Epi Info the form designer holds the ranges and legal values. In R you tell the software the scale by storing a categorical variable as a **factor**, and an ordinal one as an **ordered factor** (recall §1.1.2).

ILLUSTRATION / EXAMPLE

Codebook for the clinic register (`data/clinic_children.csv`, 100 rows):

Variable	Label	Type (scale)	Values / range	Missing
child_id	Child ID	text (identifier)	C001–C100, unique	none allowed
seen_today	One of today's 30	binary (nominal)	1 = yes, 0 = no	none allowed
arrival_order	Order of arrival today	discrete (ordinal rank)	1–30	empty if seen_today = 0 (not applicable)
age_months	Age	continuous (ratio), recorded in completed months	12–36	empty
sex	Sex	nominal	F, M	empty
weight_kg	Weight	continuous (ratio)	5–20 kg, 0.1 kg precision	empty
height_cm	Length/height	continuous (ratio)	65–110 cm	empty
nutritional_status	Nutritional status	ordinal	normal < moderate wasting < severe wasting	empty
diarrhoea_episodes_month	Diarrhoeal episodes, past month	discrete (ratio)	0–10	empty
wealth_quintile	Household wealth quintile	ordinal	1 (poorest) – 5	empty
maternal_education	Mother's education	ordinal	none < primary < secondary+	empty

The ranges are the ones set for this clinic's form; they are there to catch keying errors, not to define normal growth.

WORKED CALCULATION

Suppose the last three children of the day were not weighed and their weights were keyed as 999 instead of left empty.

1. Sum of the 27 real weights = 315.5 kg.
2. Sum with the codes = 315.5 + 3 × 999 = 3,312.5.
3. "Mean" over 30 = 3,312.5 ÷ 30 = 110.42 kg.
4. Correct handling: mean of the 27 recorded weights = 315.5 ÷ 27 = 11.69 kg, reported with "weight missing for 3 children".

Result: three missing-value codes turned an 11.7-kg mean into 110.4 kg. A range check (5–20 kg) flags all three at once. A subtler code (say 0.0) would drag the mean down without looking im-

possible, which is why missing values should be empty cells, not numbers. What to do about missing values in analysis is taken up in §15.4.

IN R (OPTIONAL)

Optional R primer: skip this box if you will not use R. R is free software for statistical computing (R Core Team, 2024); the output below comes from R 4.3.3. The lines below read the clinic register, tell R which variable is ordinal, and run the cleaning checks from the Simplified Explanation.

```
cl ← read.csv("data/clinic_children.csv")
dim(cl)
vars ← c("child_id", "sex", "weight_kg", "nutritional_status",
         "arrival_order")
str(cl[, vars])
cl$nutritional_status ← factor(cl$nutritional_status,
    levels = c("normal", "moderate wasting", "severe wasting"),
    ordered = TRUE)
table(cl$nutritional_status)
summary(cl$weight_kg)
range(cl$weight_kg)
sum(duplicated(cl$child_id))
colSums(is.na(cl))[c("weight_kg", "arrival_order")]
table(seen_today = cl$seen_today,
      has_order = !is.na(cl$arrival_order))
t30 ← subset(cl, seen_today == 1)
table(cut(t30$weight_kg, breaks = c(8, 10, 12, 14, 16),
          right = FALSE))
```

```
[1] 100  11
'data.frame':   100 obs. of  5 variables:
 $ child_id          : chr  "C001" "C002" "C003" "C004" ...
 $ sex               : chr  "F" "F" "M" "M" ...
 $ weight_kg         : num  11.2 9.8 14.1 10 12.3 13.6 11.1 11.9 12.6 9.4 ...
 $ nutritional_status: chr  "normal" "moderate wasting" "normal" "normal" ...
 $ arrival_order     : int  1 2 3 4 5 6 7 8 9 10 ...

          normal moderate wasting   severe wasting
              60               25               15
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.60   11.20   12.10   12.15   13.30   15.40
[1]  8.6 15.4
[1] 0
    weight_kg arrival_order
            0            70
          has_order
seen_today FALSE TRUE
         0    70    0
         1     0   30

 [8,10) [10,12) [12,14) [14,16)
      5      12       9       4
```

`dim` shows 100 rows (children) and 11 columns (variables). `str` shows each column's type: `chr` is text, `num` a continuous number, `int` a whole number; `nutritional_status` arrives as plain text until `factor(..., ordered = TRUE)` records its order, after which `table` lists the categories in clinical order: 60 normal, 25 moderate and 15 severe wasting. `summary` gives the minimum (8.6 kg), maximum (15.4 kg), mean (12.15 kg) and three cut points you will meet as the quartiles and median (§2.3, §3.4). The checks pass: every weight lies inside 5–20 kg, no ID is duplicated, weight has no

missing values, and the 70 missing arrival orders are exactly the 70 children not seen today (not applicable, not lost). The last table rebuilds §1.4.1 from the file: `[8,10)` is R's notation for "8 to <10", and `right = FALSE` is what puts the 10.0-kg child in the second class.

MUST-KNOW NOTES (MD-LEVEL)

- Codebook before data collection; one row per unit, one column per variable; names without spaces.
- Missing values are empty cells (`NA`), never 0, 99 or 999. Separate "not applicable" from "not recorded".
- Prevent errors at entry (validating forms, double entry); detect the rest with scripted range, consistency, duplicate and missing-value checks.
- Never overwrite the raw file; correct in a script that can be rerun.
- Declare each variable's scale in the software, so that it respects the scale (§1.1.2). In R (optional), that means storing nominal variables as factors and ordinal variables as ordered factors.

CHECKPOINT 1.7

1. Write codebook rows (name, type, values, missing code) for "ever breastfed (yes/no)" and "age at first solid food in months" in an infant-feeding survey.
2. In a double-entry comparison of 250 forms with 40 fields each, 30 cells disagree. What percentage of cells disagree, and what do you do with each disagreeing cell?
3. (Optional, R) Why does `table(cl$nutritional_status)` list "moderate wasting" first if you skip the `factor(...)` line? (Hint: without levels, R sorts text alphabetically.)

(Answers in the Appendix at the end of the book.)

Descriptive Statistics – Central Tendency

Descriptive statistics summarises the dataset in hand without claiming anything about a larger population; that claim is the job of inferential statistics (recall §1.2). It has two halves that travel together: **central tendency** (a typical, central value), built in this chapter, and **dispersion** (how spread out the values are around that centre), built in Chapter 3. This chapter builds the four means (§2.2), the median (§2.3) and the mode (§2.4), each for raw and for grouped data, and ends with how to choose among them (§2.5).

2.1 Why We Need a Typical Value

DEFINITION

A **measure of central tendency** is a single value chosen to represent where a whole distribution is centred.

SIMPLIFIED EXPLANATION

Chapter 1 turned a raw list of 30 weights into a frequency table (recall §1.4.1). If a colleague asks "how are the children in today's clinic doing, weight-wise?", you do not read out the table. You say one number: "the average weight was 11.8 kg." That number answers "what is typical here?" and pays for its brevity by discarding all the individual variation. Chapter 3 recovers that variation.

ILLUSTRATION / EXAMPLE

Two blocks in a district survey report average birthweights of 2.9 kg and 3.1 kg. You compare them at a glance. Comparing two full frequency tables to reach the same conclusion is slow and error-prone, even though the tables hold strictly more information.

MUST-KNOW NOTES (MD-LEVEL)

- Three measures exist: mean, median and mode. Each answers "what is typical?" differently and suits different scales and shapes (§2.5).
- Any single summary discards spread. Reporting a mean with no SD, or a median with no IQR, is a common and misleading shortcut (SD and IQR are defined in §3.7 and §3.4).

2.2 Mean

A **mean** combines every value in a dataset into one number by a stated combining rule. "The mean" means the **arithmetic mean** (§2.2.1) unless a question says otherwise. Three variants exist for data with a different arithmetic structure: the **weighted mean** (§2.2.2) when observations count unequally, the **geometric mean** (§2.2.3) when values combine by multiplication, and the **harmonic mean** (§2.2.4) when you average rates whose numerators are fixed. Every mean needs arithmetic to be meaningful on the variable, so the data must be at least interval-scaled (recall §1.1.2.3); the geometric and harmonic means need more, namely strictly positive values on a ratio scale (recall §1.1.2.4). A mean is never valid on nominal or ordinal data.

2.2.1 Arithmetic Mean

DEFINITION

The **arithmetic mean** is the sum of all values divided by the number of values.

SIMPLIFIED EXPLANATION

Add everything up and divide by how many things you added. Why is this "the" typical value? It is the single number that, if it replaced every observation, would leave the total unchanged. It is also the balance point: place each value as a weight on a seesaw, and the seesaw balances at the mean.

DERIVATION

The balance-point property follows in two lines. Write the n values as $x_1, x_2, \ldots, x_n$ (the small number below is a label: x_1 is the first value, not a power), and write Σ ("sigma") for "add up", so $\Sigma x = x_1 + x_2 + \ldots + x_n$. The deviations from $\bar{x}$ are $(x_1 - \bar{x}), \ldots, (x_n - \bar{x})$. Their sum is

$$\Sigma(x - \bar{x}) = \Sigma x - n \cdot \bar{x} = \Sigma x - n \cdot (\Sigma x \div n) = 0.$$

So the positive and negative deviations cancel exactly, for any dataset.

ILLUSTRATION / EXAMPLE

Five children's weights: 9, 10, 11, 12, 13 kg. Sum = 55. Mean = 55 ÷ 5 = 11 kg. The deviations are −2, −1, 0, +1, +2, and they sum to 0, as the derivation promised.

WORKED CALCULATION: MEAN FROM GROUPED DATA

When only the frequency table survives, each class is represented by its **class mark** (mid-point; recall §1.4.2), and each mark is counted as many times as its class frequency f. The mean is then $\Sigma(f \cdot x) \div \Sigma f$. Today's 30 children (recall §1.4.1):

Class (kg)	Mid-point x	f	f·x
8 to <10	9	5	45
10 to <12	11	12	132
12 to <14	13	9	117

Class (kg)	Mid-point x	f	f·x
14 to <16	15	4	60
Total		30	354

1. Multiply each mid-point by its frequency (last column).
2. Add: Σf·x = 354; Σf = 30.
3. Grouped mean = 354 ÷ 30 = **11.8 kg**.

The mean of the 30 raw weights is 11.81 kg. The grouped mean is close because the values inside each class sit roughly evenly around its mid-point; with wide or lopsided classes the two can differ more.

MUST-KNOW NOTES (MD-LEVEL)

- Formulas: sample mean $\bar{x} = \Sigma x \div n$; population mean $\mu = \Sigma x \div N$; grouped mean = $\Sigma f \cdot x \div \Sigma f$.
- **Defining property:** $\Sigma(x - \bar{x}) = 0$. This is why the average deviation from the mean is useless as a measure of spread, and why dispersion measures either take absolute values or square the deviations (§3.5, §3.6).
- Every value pulls on the mean in proportion to its size, so one extreme value can drag it away from where most of the data sit. The median (§2.3) exists partly to resist this.

2.2.2 Weighted Mean

DEFINITION

The **weighted mean** is an arithmetic mean in which each value counts in proportion to an assigned **weight** w, rather than every value counting once.

SIMPLIFIED EXPLANATION

The arithmetic mean assumes every observation deserves equal say. When some observations stand for more units of the thing you are averaging, that assumption is wrong. The weighted mean multiplies each value by its weight before adding, and divides by the total weight instead of a plain count.

ILLUSTRATION / EXAMPLE

A health worker has the average household size in three villages: Village A (25 households, average size 4), Village B (100 households, average size 5), Village C (10 households, average size 3).

The unweighted mean (4 + 5 + 3) ÷ 3 = 4 gives tiny Village C the same say as Village B. The quantity wanted is the average size of a household in the district, so each village's weight is its number of households:

Weighted mean = (25×4 + 100×5 + 10×3) ÷ (25 + 100 + 10) = 630 ÷ 135 ≈ **4.67**

The weight must match the quantity being averaged. Had the question been "what household size does a randomly chosen *person* in the district live in?", the weight would be the number of persons in each vil-

lage, because a person is more likely to live in a large household. Both are legitimate weighted means; they answer different questions.

MUST-KNOW NOTES (MD-LEVEL)

- Formula: weighted mean = $\Sigma(w \cdot x) \div \Sigma w$. The arithmetic mean is the special case with every $w = 1$. The grouped mean in §2.2.1 is a weighted mean of the mid-points, with the frequencies as weights.
- Choose the weight from the target quantity: households for a per-household quantity, persons for a per-person quantity. Never average area-level rates or percentages across areas of unequal size without asking which weight fits. Doing so gives small areas an unfair say and produces a misleading "average of averages".

2.2.3 Geometric Mean

DEFINITION

The **geometric mean (GM)** of n positive values is the nth root of their product. It averages quantities that combine by multiplication, such as growth factors and ratios, and data whose logarithms are roughly symmetric, such as antibody titres.

SIMPLIFIED EXPLANATION

A population that grows 10% one year and 20% the next has been multiplied by 1.10 and then by 1.20. The GM is the single factor that, applied every year, gives the same overall product.

DERIVATION

Taking logarithms turns products into sums: $\log(x_1 \times x_2 \times \ldots \times x_n) = \Sigma \log x$. So

$\log GM = (1/n) \times \Sigma \log x$, and $GM = \exp(\text{mean of } \log x)$, where exp undoes the natural log: $\exp(\ln x) = x$ (with logs to base 2 or 10, raise 2 or 10 to that power instead).

The GM is therefore the arithmetic mean on the log scale, converted back ("back-transformed") to the original scale. This is how software computes it, and it is why the GM needs strictly positive values: the logarithm of zero or of a negative number does not exist.

ILLUSTRATION / EXAMPLE

Growth. A population's growth factors over three years are 1.10, 1.20 and 1.05.

$GM = \sqrt[3]{(1.10 \times 1.20 \times 1.05)} = \sqrt[3]{1.386} \approx \mathbf{1.1149}$, an average annual growth of about 11.49%.

The arithmetic mean of the three percentages is $(10 + 20 + 5) \div 3 = 11.67\%$, which overstates the compounded growth: $1.11667^3 = 1.3924$, more than the true product 1.386.

WORKED CALCULATION: GEOMETRIC MEAN TITRE (GMT)

Serum antibody titres are reported as reciprocals of doubling dilutions (1:16, 1:32, 1:64 ...), and a few children respond far more strongly than the rest. Vaccine studies therefore log-transform

titres and report the **geometric mean titre (GMT)**; an Indian H1N1 vaccine trial, for example, used "log transformation of the data to calculate Geometric Mean Titers (GMTs)" (Kubavat et al., 2011). Take eight children's reciprocal titres after a vaccine dose (illustrative values): 16, 32, 32, 64, 64, 128, 256, 1024.

1. Take $\log_2$ of each (each doubling adds 1): 4, 5, 5, 6, 6, 7, 8, 10.
2. Mean of the logs = 51 ÷ 8 = 6.375.
3. Back-transform: GMT = 2^6.375 (2 raised to the power 6.375) ≈ **83** (a titre of about 1:83).
4. Compare: arithmetic mean = 1616 ÷ 8 = 202; median = 64.

The single child at 1024 more than doubles the arithmetic mean. Without that child the arithmetic mean is 84.6, and the GMT falls from 83 to 58. The GMT sits near the middle of the doubling steps, where most children are. Any log base gives the same GMT; base 2 just makes the steps whole numbers.

MUST-KNOW NOTES (MD-LEVEL)

- Formula: GM = $(x_1 \times x_2 \times \ldots \times x_n)^{(1/n)} = \exp(\text{mean of } \ln x)$.
- Uses: average growth factors (population growth, year-on-year rise in incidence), ratios, and positively skewed positive data such as antibody titres, parasite counts and many laboratory values, where the GM is the back-transformed mean of the logs.
- Needs strictly positive, ratio-scale values. A GM of temperatures in °C, or of a variable that includes zeros, is meaningless.
- **GM ≤ arithmetic mean** for the same positive data, equal only when all values are identical. This is a useful check when computing by hand.

2.2.4 Harmonic Mean

DEFINITION

The **harmonic mean (HM)** of n positive values is the reciprocal of the arithmetic mean of their reciprocals: HM = n ÷ Σ(1/x).

SIMPLIFIED EXPLANATION

The HM is the right average for rates when the **numerator** of the rate is the same for every observation. The classic case is speed (distance ÷ time) over equal distances: you spend more time at the slower speed, so the slower speed deserves more weight than a plain average gives it.

DERIVATION

The honest average of any set of rates is total numerator ÷ total denominator. For speeds v_i over distances d_i, the time on leg i is d_i/v_i, so

overall speed = Σd ÷ Σ(d/v).

If every leg has the same distance d, the d's cancel: overall speed = n·d ÷ (d·Σ(1/v)) = n ÷ Σ(1/v), which is the HM. If instead every leg has the same *time* t, overall speed = Σ(v·t) ÷ (n·t) = Σv ÷ n, the plain arithmetic mean.

ILLUSTRATION / EXAMPLE

You drive 60 km at 30 km/h, then another 60 km at 60 km/h. The arithmetic mean of the speeds is 45 km/h, but the trip took 60/30 + 60/60 = 3 h for 120 km, so the true average is 120 ÷ 3 = **40 km/h**. The HM gives 2 ÷ (1/30 + 1/60) = 2 ÷ 0.05 = **40 km/h**. Had you instead driven one hour at each speed, you would have covered 90 km in 2 h, and the arithmetic mean, 45 km/h, would be correct.

MUST-KNOW NOTES (MD-LEVEL)

- **Rule:** equal numerators (same distance, same number of doses) → harmonic mean; equal denominators (same time, same population) → arithmetic mean. In general, use total numerator ÷ total denominator, which is a weighted mean of the rates with the denominators as weights.
- For the same positive data, **HM ≤ GM ≤ arithmetic mean**, all equal only when every value is identical. These three are the "Pythagorean means".
- The HM is rarely named in community medicine, but the rule behind it recurs whenever rates are combined across groups.

CHECKPOINT 2.1

1. A district has three villages. Village P has 1,000 eligible children with 60% measles vaccination coverage; villages Q and R have 100 eligible children each, both with 95% coverage. Compute the unweighted mean of the three coverages and the district's true coverage, and explain which weight you used.
2. A population grew 10% one year and 20% the next. Compute the average annual growth rate with the geometric mean and compare it with the simple average of 10% and 20%.
3. An ASHA walks 12 km to a sub-centre at 4 km/h and returns the same 12 km by shared auto at 12 km/h. What is her average speed for the round trip, and which mean gives it directly?

(Answers in the Appendix at the end of the book.)

2.3 Median

DEFINITION

The **median** is the middle value of a dataset arranged in order: as many observations lie below it as above it.

SIMPLIFIED EXPLANATION

The mean is pulled by every value's size, including extreme ones. The median depends only on *position*: which value sits in the middle once everything is sorted. One very large or very small value shifts the mean a lot but moves the middle position little or not at all.

ILLUSTRATION / EXAMPLE

Five children's weights, sorted: 9, 10, 11, 12, 13 kg. Median = 11 kg, the same as the mean because these data are symmetric.

Now one child is severely malnourished and weighs 2 kg instead of 9: sorted weights 2, 10, 11, 12, 13. The mean drops to 9.6 kg. The median stays at **11 kg**, because the middle value did not move. For an even number of values, the median is the average of the two middle values: today's 30 sorted weights have 11.6 and 11.8 kg in positions 15 and 16, so the raw median is **11.7 kg**.

DERIVATION

With grouped data the individual values are gone, so you locate the **median class** (the class containing the (N/2)th observation) from the cumulative frequencies (recall §1.4.3) and assume its f values are spread evenly across its width h. Let L be the lower boundary of the median class and cf the cumulative frequency of all classes below it. You need N/2 − cf more observations after L. Each observation in the median class occupies h/f of the width, so walk (N/2 − cf) × h/f into the class:

$$\text{Median} = L + [(N/2 - cf) \div f] \times h$$

WORKED CALCULATION: GROUPED MEDIAN

Today's 30 children: cumulative frequencies 5, 17, 26, 30.

1. N/2 = 15. The 15th child lies in the class whose cumulative frequency first reaches 15: 10 to <12 (cumulative 17).
2. L = 10, cf = 5 (the class below), f = 12, h = 2.
3. Median = 10 + [(15 − 5) ÷ 12] × 2 = 10 + 1.667 = **11.67 kg**.

The raw median was 11.7 kg. The small gap comes from the even-spread assumption.

MUST-KNOW NOTES (MD-LEVEL)

- Odd n: the middle value after sorting, position (n + 1)/2. Even n: the average of the two middle values.
- Grouped data: Median = L + [(N/2 − cf) ÷ f] × h. Use the lower class *boundary* for L (for "10 to <12" this book uses 10, ignoring the 0.05-kg rounding shift noted in §1.4.2), and cf of the classes *below* the median class.
- The median is **robust**: extreme values barely move it. This is its main practical advantage over the mean.
- The median suits ordinal data and skewed quantitative data; the mean needs interval or ratio data and describes roughly symmetric distributions best (§2.5).

- The median is the value that minimises the sum of absolute distances $\Sigma|x - c|$; the mean minimises the sum of squared distances. Both facts are derived in §3.5 and §3.6.

2.4 Mode

DEFINITION

The **mode** is the value or category that occurs most often in a dataset.

SIMPLIFIED EXPLANATION

Sometimes the question is neither "where does the data balance?" (mean) nor "what is in the middle?" (median) but "what shows up most often?". That is the mode. It is the only one of the three that works for nominal data: you cannot average or rank blood groups, but you can count which one is commonest.

ILLUSTRATION / EXAMPLE

Blood groups recorded for 20 patients: O 9, A 6, B 4, AB 1. Mode = **O**. Mean and median are undefined, because blood group has no numeric value and no order.

A dataset can be **unimodal** (one peak), **bimodal** (two peaks) or **multimodal** (more than two).

DERIVATION

For grouped data, the **modal class** is the class with the highest frequency f_1. Let f_0 and f_2 be the frequencies of the classes just below and just above it. The peak should lean toward the heavier neighbour. Draw a line from the top-left corner of the modal bar to the top-left corner of the next bar, and another from the top-right corner of the modal bar to the top-right corner of the previous bar (the grey lines in the figure below). By similar triangles, they cross at a distance d from L where $d \div (h - d) = (f_1 - f_0) \div (f_1 - f_2)$. Solving for d gives

$$\text{Mode} = L + [(f_1 - f_0) \div (2f_1 - f_0 - f_2)] \times h$$

WORKED CALCULATION: GROUPED MODE

1. Modal class: 10 to <12 ($f_1 = 12$). $L = 10$, $h = 2$, $f_0 = 5$ (class below), $f_2 = 9$ (class above).
2. Mode $= 10 + [(12 - 5) \div (24 - 5 - 9)] \times 2 = 10 + (7 \div 10) \times 2 =$ **11.4 kg**.

The peak leans right of the class mid-point (11) because the class above (9 children) is heavier than the class below (5 children).

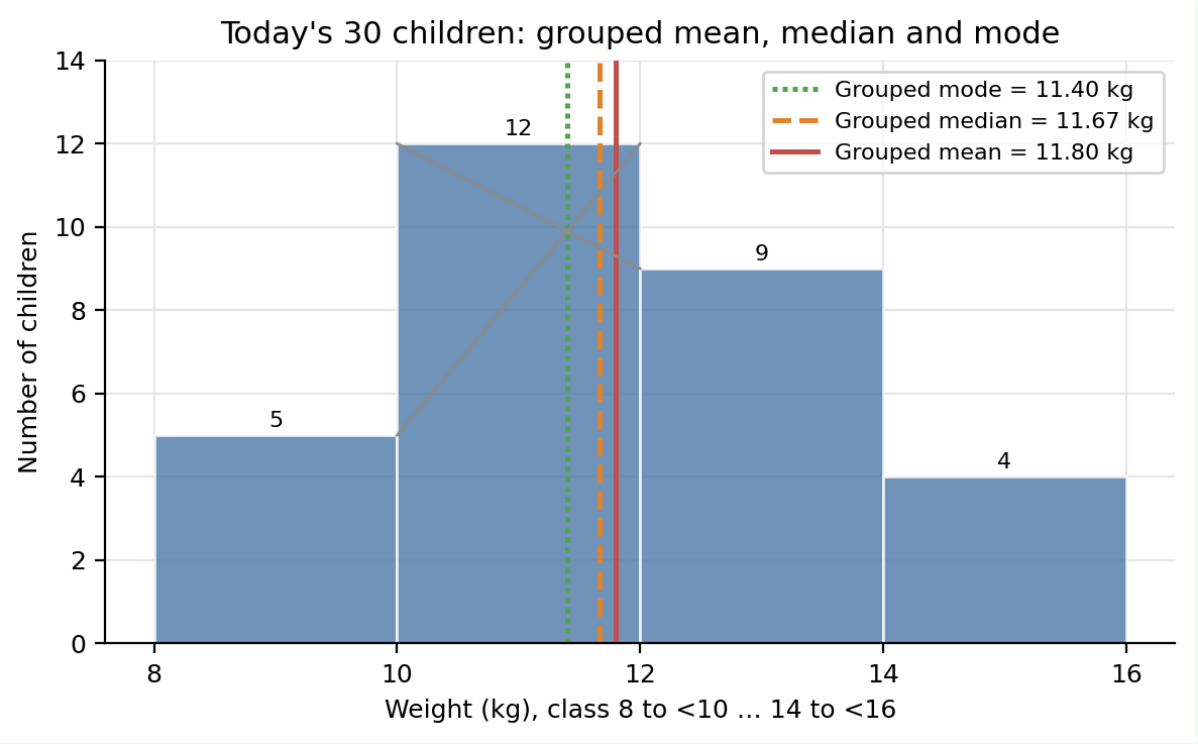


Figure 2.1. Histogram of today's 30 children. The grey lines are the construction behind the grouped-mode formula; they cross at the grouped mode, 11.40 kg. The grouped median (11.67 kg) and grouped mean (11.80 kg) lie to its right.

MUST-KNOW NOTES (MD-LEVEL)

- The mode is the only measure of central tendency valid for nominal data.
- Grouped data: Mode = $L + [(f_1 - f_0) \div (2f_1 - f_0 - f_2)] \times h$. It needs equal class widths around the modal class.
- **Empirical relation.** A classical rule of thumb for moderately skewed, single-peaked distributions: mode ≈ 3 × median − 2 × mean, equivalently mean − mode ≈ 3(mean − median). It is an approximation, not an identity; §3.9 checks it on a smooth skewed curve (Figure 3.1). For today's 30 children: 3 × 11.67 − 2 × 11.8 = 35.0 − 23.6 = 11.4 kg, which matches the grouped mode. Use it when the mode is hard to estimate, and do not trust it for strongly skewed or multimodal data.
- Bimodality in health data can signal two subgroups mixed together, for example two disease subtypes with different typical ages of onset.
- In small samples the mode is the least stable of the three measures: a single changed value can shift which value is commonest.

2.5 Choosing the Right Measure of Central Tendency

DEFINITION

The choice among mean, median and mode rests on two things: the variable's scale of measurement (recall §1.1.2) and the shape of its distribution.

SIMPLIFIED EXPLANATION

Given a real dataset, which one do you report? First ask what scale the variable is on; that rules measures in or out. Then, for interval or ratio data, ask whether the distribution is roughly symmetric or skewed (**skewed** means one tail is longer than the other; §3.9 measures it).

ILLUSTRATION / EXAMPLE

- **Nominal data** (blood group) → mode only; mean and median are undefined.
- **Ordinal data** (cancer stage, Kuppuswamy class) → median (and mode). The mean is not valid, though it is sometimes used pragmatically.
- **Interval/ratio data, roughly symmetric** (height in a healthy population) → mean. It uses every value, and for Normal data (the bell-shaped curve of §5.7.2.2) it varies less from sample to sample than the median does (the property called efficiency in §7.1.1). With heavier tails than the Normal the advantage can shrink or reverse.
- **Interval/ratio data, skewed or with outliers** (household income, hospital length of stay) → median, which resists the distortion the mean suffers.
- **Positive, right-skewed data whose logs are roughly symmetric** (antibody titres) → geometric mean (§2.2.3).

MUST-KNOW NOTES (MD-LEVEL)

- Two independent checks drive the decision: scale of measurement (nominal → mode; ordinal → median; interval/ratio → mean usually appropriate) and, for interval/ratio data, shape (symmetric → mean; skewed → median).
- Classic exam and thesis trap: reporting a mean (and SD) for an ordinal or a strongly skewed variable (income, length of stay, many laboratory values) when the median (and IQR, §3.4) is the better summary.
- A quick shape check: in a unimodal distribution, a mean noticeably away from the median usually signals skew. The converse is weaker. Mean and median can be close in a distribution that is bimodal, or skewed in a way that balances out, so look at a histogram (§4.3.1) before trusting either number alone (von Hippel, 2005).
- The same two checks, scale and shape, return when choosing between parametric tests (tests that assume a particular distribution for the data, such as the t-test in §9.2) and non-parametric tests (§10.1).

CHECKPOINT 2.2

1. Hospital length of stay is right-skewed. In one ward the mean stay is 6.2 days and the median is 4 days. Which would you report as the typical stay, and why could the mean alone mislead?

2. For a dataset of blood groups (A, B, AB, O), which single measure of central tendency is valid, and why do the other two fail?
3. Haemoglobin of 50 pregnant women at a PHC is grouped as 7 to <9 g/dL (6 women), 9 to <11 (14), 11 to <13 (20) and 13 to <15 (10). Compute the grouped mean, median and mode. Then check the mode against the empirical relation mode ≈ 3 × median − 2 × mean.

(Answers in the Appendix at the end of the book.)

Descriptive Statistics – Dispersion

Dispersion (variability, spread) is the second half of descriptive statistics: how scattered the values are around their centre. Two datasets can share a mean and still describe very different realities (§3.1). This chapter builds the measures of spread in rising order of refinement: the range (§3.2), quartiles, percentiles and the interquartile range (§3.3–3.4), the mean deviation (§3.5), variance and standard deviation (§3.6–3.7), and the coefficient of variation (§3.8). It ends with two measures of shape, skewness (§3.9) and kurtosis (§3.10).

3.1 Why Average Alone Is Not Enough

DEFINITION

A measure of central tendency, however well chosen, can hide very different underlying realities. A **measure of dispersion** reports how far the values lie from one another or from their centre.

SIMPLIFIED EXPLANATION

Two clinics both report a mean waiting time of 20 minutes. In Clinic X almost everyone waits close to 20 minutes. In Clinic Y half the patients wait 5 minutes and half wait 35. The mean cannot tell the clinics apart, yet their problems, and the fixes they need, are completely different.

ILLUSTRATION / EXAMPLE

Clinic X's waits: 18, 19, 21, 22 minutes (mean 20; range 4 minutes). Clinic Y's waits: 5, 5, 35, 35 minutes (mean 20; range 30 minutes). Clinic Y probably has a subset of patients held up at one bottleneck, such as a particular test or specialist; Clinic X has a modest delay that affects everyone.

MUST-KNOW NOTES (MD-LEVEL)

- Identical central tendency can describe very different situations. Dispersion measures tell them apart.
- Report a centre with a spread: mean with SD (§3.7), or median with IQR (§3.4). A mean reported alone hides exactly the difference between Clinic X and Clinic Y.

3.2 Range

DEFINITION

The **range** is the difference between the largest and smallest values: Range = Maximum − Minimum.

SIMPLIFIED EXPLANATION

The crudest answer to "how spread out is this?" is to subtract the smallest value from the largest. It gives the span the data cover and says nothing about how values are arranged inside that span.

ILLUSTRATION / EXAMPLE

Weights of five children: 9, 10, 11, 12, 13 kg → range = 13 − 9 = **4 kg**. A different five children: 9, 11, 11, 11, 13 → range is *still* 4 kg, although these values cluster tightly at 11 apart from the two ends. The range uses only 2 of the n values.

Today's 30 children (recall §1.4.1) run from 8.6 to 15.3 kg: range = **6.7 kg**.

MUST-KNOW NOTES (MD-LEVEL)

- Simple, but it uses only the two extremes, so a single outlier changes it completely, and it tends to grow as the sample grows (more chances to catch an extreme value).
- Useful as a first data check: an age range of "−5 to 200 years" flags an entry error at once. It is rarely the main reported measure of spread in research.

3.3 Quartiles Deciles and Percentiles

DEFINITION

Quartiles, **deciles** and **percentiles** are values that divide sorted data into 4, 10 and 100 equal-sized parts. The pth percentile is the value below which a proportion p of the data lie. The median is the 50th percentile, which is also the 2nd quartile (Q2).

SIMPLIFIED EXPLANATION

The median splits sorted data into two halves. Split into four and you get quartiles: Q1 (25th percentile), Q2 (the median) and Q3 (75th percentile). Deciles and percentiles are the same idea at finer resolution.

WORKED CALCULATION: QUARTILES OF RAW DATA

Several rules exist for placing a quartile between two observations, and software differs; R alone offers nine (Hyndman & Fan, 1996, via the R documentation). This book uses R's default (type 7): the pth quantile sits at position 1 + (n − 1)p in the sorted data, interpolating between neighbours.

1. Five children, sorted 9, 10, 11, 12, 13. Q1 position = 1 + 4 × 0.25 = 2 → Q1 = **10 kg**. Q3 position = 1 + 4 × 0.75 = 4 → Q3 = **12 kg**.

2. Today's 30 children, sorted (recall §1.4.1). Q1 position = 1 + 29 × 0.25 = 8.25: the 8th value is 10.5 and the 9th is 10.8, so Q1 = 10.5 + 0.25 × (10.8 − 10.5) = **10.575 kg**.
3. Q3 position = 1 + 29 × 0.75 = 22.75: the 22nd value is 12.9 and the 23rd is 13.1, so Q3 = 12.9 + 0.75 × 0.2 = **13.05 kg**.

SPSS's default rule (R type 6, position (n + 1)p) gives 10.45 and 13.175 kg for the same data. State the rule you used when you report quartiles from small samples.

WORKED CALCULATION: QUARTILES FROM GROUPED DATA

Replace N/2 in the grouped-median formula (recall §2.3) with the target position: Qk = L + [(kN/4 − cf) ÷ f] × h. Today's 30 children (cumulative frequencies 5, 17, 26, 30):

1. Q1: N/4 = 7.5 → class 10 to <12 (cf below = 5, f = 12). Q1 = 10 + [(7.5 − 5) ÷ 12] × 2 = **10.42 kg**.
2. Q3: 3N/4 = 22.5 → class 12 to <14 (cf below = 17, f = 9). Q3 = 12 + [(22.5 − 17) ÷ 9] × 2 = **13.22 kg**.

ILLUSTRATION / EXAMPLE

A child whose weight sits at the 10th percentile of a growth chart weighs more than 10% of children of the same age and sex in the reference population, and less than 90%. The WHO Child Growth Standards are built by modelling the reference distribution at each age (the Box-Cox power exponential method) and are presented as both percentile curves and z-score curves (WHO Multicentre Growth Reference Study Group, 2006). A z-score counts standard deviations from the reference median; it is defined in §5.7.2.2.1 and applied to growth standards in §5.7.2.2.3.

MUST-KNOW NOTES (MD-LEVEL)

- Q1 = 25th percentile, Q2 = median = 50th percentile, Q3 = 75th percentile. Deciles D1…D9; percentiles P1…P99.
- Like the median, these are positional measures: robust to outliers and valid for ordinal as well as skewed quantitative data.
- Raw data: state the quartile rule. Grouped data: Qk = L + [(kN/4 − cf) ÷ f] × h, and any percentile Pp = L + [(pN/100 − cf) ÷ f] × h.

3.4 Interquartile Range

DEFINITION

The **interquartile range (IQR)** is Q3 − Q1, the span of the middle 50% of the data. Half of it is the **quartile deviation (QD)**, or semi-interquartile range. The **coefficient of quartile deviation** is (Q3 − Q1) ÷ (Q3 + Q1), a unit-free version.

SIMPLIFIED EXPLANATION

The range was ruined by a single extreme value. The IQR discards the lowest and highest quarter of the data and reports the spread of the middle half, the part least affected by extremes.

WORKED CALCULATION

1. Five children: IQR = 12 − 10 = **2 kg**, against a range of 4 kg.
2. Today's 30 children (type 7 quartiles from §3.3): IQR = 13.05 − 10.575 = **2.475 kg**; QD = 2.475 ÷ 2 = **1.24 kg**; coefficient of QD = 2.475 ÷ (13.05 + 10.575) = 2.475 ÷ 23.625 = **0.105**.
3. Grouped quartiles for the same children: IQR = 13.222 − 10.417 = 2.81 kg; QD = **1.40 kg**; coefficient of QD = 2.81 ÷ 23.64 = **0.119**. The grouped values differ from the raw ones because grouping spreads values evenly across each class.

MUST-KNOW NOTES (MD-LEVEL)

- IQR = Q3 − Q1; QD = (Q3 − Q1) ÷ 2; coefficient of QD = (Q3 − Q1) ÷ (Q3 + Q1).
- The IQR is the usual companion of the median, as the SD is the companion of the mean (recall §2.5).
- **Outlier fences (a convention).** Values below Q1 − 1.5 × IQR or above Q3 + 1.5 × IQR are flagged as possible outliers. This is Tukey's convention for box-plot whiskers (Tukey, 1977; 1.5 is also the default multiplier in R's box-plot documentation). NIST calls points beyond these "inner fences" mild outliers and points beyond wider "outer fences" extreme outliers. It is one convention among several: box-plot software may use "hinges" that differ slightly from the quartiles, and a flagged value is a prompt to check, not a licence to delete. For today's 30 children the fences are 10.575 − 3.71 = 6.86 kg and 13.05 + 3.71 = 16.76 kg, so no child is flagged. Box plots are drawn in §4.3.4.

CHECKPOINT 3.1

1. Two datasets have the same range. Explain how they could still look completely different, and which measure from §3.3–3.4 would catch the difference.
2. A child's weight is "at the 10th percentile" on a growth chart. Explain what that means without using the word "percentile".
3. For six women's haemoglobin values, 9.0, 10.5, 11.0, 11.5, 12.0 and 14.0 g/dL, find Q1 and Q3 by the type 7 rule, the IQR, the quartile deviation and the coefficient of quartile deviation.

(Answers in the Appendix at the end of the book.)

3.5 Mean Deviation

DEFINITION

The **mean deviation (MD)**, or mean absolute deviation, is the average absolute distance of the values from a chosen centre, usually the mean: MD = $\Sigma|x - \bar{x}| \div n$. It can also be taken about the median: MD = $\Sigma|x - \text{median}| \div n$.

SIMPLIFIED EXPLANATION

The range and IQR use only a couple of positions in the sorted data. The MD is the first measure here to use every value. Raw deviations from the mean always sum to zero (recall §2.2.1), so the MD drops their signs and averages the distances.

DERIVATION

About which centre c is $\Sigma|x - c|$ smallest? Sort the data and move c a small step δ to the right. Every value above c gets δ closer; every value below c gets δ farther. The total changes by (number below − number above) × δ. While more than half the values lie above c, moving right reduces the total; once more than half lie below, moving right increases it. The total is therefore smallest where the counts balance, at the **median**. So the MD about the median is never larger than the MD about the mean, or about any other point.

ILLUSTRATION / EXAMPLE

Five children, 9, 10, 11, 12, 13 kg, mean 11. Absolute deviations: 2, 1, 0, 1, 2. MD = 6 ÷ 5 = **1.2 kg**: on average a child's weight lies 1.2 kg from the mean. Here mean = median, so the MD about the median is also 1.2 kg.

Replace the 9 kg child with a 2 kg child: 2, 10, 11, 12, 13 (mean 9.6, median 11).

- About the mean: $|2 - 9.6| + |10 - 9.6| + \ldots = 7.6 + 0.4 + 1.4 + 2.4 + 3.4 = 15.2$; MD = **3.04 kg**.
- About the median: $9 + 1 + 0 + 1 + 2 = 13$; MD = **2.6 kg**, the smaller, as the derivation promised.

For today's 30 children the MD is 1.41 kg about either the mean or the median, because the two centres almost coincide (11.81 and 11.7 kg).

MUST-KNOW NOTES (MD-LEVEL)

- MD = $\Sigma|x - \bar{x}| \div n$ (about the mean) or $\Sigma|x - \text{median}| \div n$ (about the median). The MD about the median is the minimum possible.
- It uses every value and is easy to explain. Its weakness is algebraic: absolute distances do not split cleanly into parts. Squared distances do (the identity in §3.6), which is why variance, not MD, became the base of later methods.

3.6 Variance

DEFINITION

Variance is the average squared deviation from the mean. The **sum of squares (SS)** is $\Sigma(x - \bar{x})^2$, the total of the squared deviations.

SIMPLIFIED EXPLANATION

Squaring, like taking absolute values, stops positive and negative deviations cancelling. Squaring also weights large deviations more heavily, and it gives variance an algebra that absolute values lack: sums of

squares split into meaningful parts. That property, shown below, underlies standard errors, ANOVA (literally "analysis of variance", §9.5) and regression.

DERIVATION

For any number c, write $x - c = (x - \bar{x}) + (\bar{x} - c)$, square and add over all n values:

$\Sigma(x - c)^2 = \Sigma(x - \bar{x})^2 + 2(\bar{x} - c)\cdot\Sigma(x - \bar{x}) + n(\bar{x} - c)^2$

The middle term is zero, because $\Sigma(x - \bar{x}) = 0$ (recall §2.2.1). So

$\Sigma(x - c)^2 = SS + n(\bar{x} - c)^2$

Two consequences follow. First, the sum of squared deviations is smallest when $c = \bar{x}$: the mean is the centre that minimises squared distance, as the median minimises absolute distance (§3.5). Second, the total squared distance from any point splits exactly into "spread around the mean" plus "distance of the mean from that point". No such split exists for absolute distances.

Check on the five children ($\bar{x} = 11$, SS = 10): about c = 12, $\Sigma(x - 12)^2 = 9 + 4 + 1 + 0 + 1 = 15 = 10 + 5 \times 1^2$. Expanding the square the same way also gives the hand-computing formula $SS = \Sigma x^2 - (\Sigma x)^2 \div n$: here $615 - 55^2 \div 5 = 615 - 605 = 10$.

ILLUSTRATION / EXAMPLE

Five weights 9, 10, 11, 12, 13, mean 11. Deviations −2, −1, 0, 1, 2; squared 4, 1, 0, 1, 4; SS = 10. Treating these five as the whole population, variance = 10 ÷ 5 = **2 kg²**.

MUST-KNOW NOTES (MD-LEVEL)

- Variance is in squared units (kg^2), which is hard to picture. The standard deviation (§3.7) takes the square root to return to kg.
- Two versions exist: the population variance divides SS by N (§3.6.1); the sample variance divides by n − 1 (§3.6.2).
- Identity to remember: $\Sigma(x - c)^2 = SS + n(\bar{x} - c)^2$. It is the first instance of the partitioning of sums of squares that ANOVA and regression use.

3.6.1 Population Variance

DEFINITION

Population variance (σ^2) is the average squared deviation of every value in a population from the population mean μ: $\sigma^2 = \Sigma(x - \mu)^2 \div N$.

SIMPLIFIED EXPLANATION

When every unit of the population is measured, μ is known exactly and nothing is estimated. Dividing by N gives the true average squared deviation.

ILLUSTRATION / EXAMPLE

If you weighed all 50,000 under-five children in a district, summed their squared deviations from the district mean and divided by 50,000, you would have σ^2. It is one fixed number, a parameter (recall §1.3), even though nobody may ever compute it.

MUST-KNOW NOTES (MD-LEVEL)

- $\sigma^2 = \Sigma(x - \mu)^2 \div N$. It is a parameter, written with a Greek letter.
- Research almost never has the whole population, so σ^2 is estimated from a sample (§3.6.2).

3.6.2 Sample Variance and Bessel's Correction

DEFINITION

Sample variance (s^2) estimates σ^2 from a sample of n values by dividing the sum of squares by **n − 1**: $s^2 = \Sigma(x - \bar{x})^2 \div (n - 1)$. Dividing by n − 1 instead of n is **Bessel's correction**.

SIMPLIFIED EXPLANATION

In a sample you do not know μ, so you measure deviations from $\bar{x}$. But $\bar{x}$ was computed from the same values, and §3.6 showed that $\bar{x}$ is the point that makes the sum of squared deviations smallest. Deviations from $\bar{x}$ are therefore, as a rule, a little smaller than the deviations from μ that σ^2 is built on. Divide by n and you underestimate σ^2 **on average** over repeated samples; any single sample can still land above σ^2. Dividing by n − 1 enlarges the estimate by exactly the right amount, on average. The derivation shows why the factor is exactly n − 1.

DERIVATION

Put $c = \mu$ in the identity from §3.6:

$$SS = \Sigma(x - \mu)^2 - n(\bar{x} - \mu)^2$$

Now average both sides over all possible samples of n independently drawn values. (This long-run average over all possible samples is called the **expected value**, written E[·].)

1. Each $(x - \mu)^2$ averages σ^2, by the definition of σ^2. So $E[\Sigma(x - \mu)^2] = n\sigma^2$.
2. $\bar{x} - \mu = (1/n)\Sigma(x - \mu)$. Its square is $(1/n^2) \times [\Sigma(x - \mu)^2$ + the cross-products $(x_i - \mu)(x_j - \mu)$ for $i \neq j]$. For independent draws, the average of a product is the product of the averages, and each deviation from μ averages 0, so every cross-product averages 0. Hence $E[(\bar{x} - \mu)^2] = (1/n^2) \times n\sigma^2 = \sigma^2/n$. (Its square root, $\sigma/\sqrt{n}$, is the standard error of the mean, §6.7.)
3. Therefore $E[SS] = n\sigma^2 - n \times \sigma^2/n =$ **$(n - 1)\sigma^2$**.

So $E[SS \div n] = (n - 1)\sigma^2/n$, which is too small, and $E[SS \div (n - 1)] = \sigma^2$. An estimator whose average over repeated samples equals the parameter is called **unbiased** (the formal definition is in §7.1.1): s^2 is unbiased for σ^2, and $SS \div n$ is not.

WORKED CALCULATION: CHECKING THE DERIVATION EXACTLY

Treat the five weights 9, 10, 11, 12, 13 as a whole population: $\mu = 11$, $\sigma^2 = 2$ (§3.6). Draw samples of n = 2 children independently (with replacement). There are $5 \times 5 = 25$ equally likely ordered samples. For each, $SS = \Sigma(x - \bar{x})^2$:

SS	SS ÷ n (n = 2)	SS ÷ (n − 1)	Number of samples
0	0	0	5
0.5	0.25	0.5	8
2	1	2	6
4.5	2.25	4.5	4
8	4	8	2

1. Average SS = $(0\times5 + 0.5\times8 + 2\times6 + 4.5\times4 + 8\times2) \div 25 = 50 \div 25 =$ **2**, which equals $(n - 1)\sigma^2 = 1 \times 2$.
2. Average of SS ÷ n = **1**: dividing by n gets $\sigma^2 = 2$ wrong by half, on average.
3. Average of SS ÷ (n − 1) = **2** = σ^2: Bessel's correction is exactly right on average.
4. Yet in 6 of the 25 samples (SS = 4.5 or 8), even SS ÷ n exceeds σ^2. The bias is a property of the average, not of every sample.

The same enumeration with n = 3 (125 samples) gives average SS ÷ n = 1.33 = (2/3) × 2, and average SS ÷ (n − 1) = 2 again.

ILLUSTRATION / EXAMPLE

The five weights as a **sample**: SS = 10, so $s^2 = 10 \div 4 =$ **2.5 kg²**, larger than the population-formula value of 2. For today's 30 children, $s^2 = 88.95 \div 29 =$ **3.07 kg²**.

MUST-KNOW NOTES (MD-LEVEL)

- $s^2 = \Sigma(x - \bar{x})^2 \div (n - 1)$, and $E[\Sigma(x - \bar{x})^2] = (n - 1)\sigma^2$.
- The divisor n − 1 is the **degrees of freedom** of s^2 (§3.6.3). The degrees-of-freedom story is intuition; the derivation above is the proof.
- The correction matters in small samples (4 against 5 is a 25% difference) and hardly at all in large ones (999 against 1000).
- s^2 is unbiased for σ^2, but **$s = \sqrt{s^2}$ is not unbiased for σ**: it underestimates σ slightly on average, because the square root of an average exceeds the average of square roots. In the n = 2 enumeration above, the average s is 1.13, against $\sigma = \sqrt{2} = 1.41$.

3.6.3 Degrees of Freedom

DEFINITION

Degrees of freedom (df) is the number of independent pieces of information left for estimating a quantity after the constraints already imposed on the data are counted: usually, the number of observations minus the number of parameters estimated from them first.

SIMPLIFIED EXPLANATION

Once $\bar{x}$ is calculated from n values, the n deviations from it must sum to zero (recall §2.2.1). Know any n − 1 of them and the last is fixed. So only n − 1 deviations are free to vary, and s^2 has n − 1 degrees of freedom. The derivation in §3.6.2 reaches the same n − 1 by algebra.

ILLUSTRATION / EXAMPLE

Five weights with mean 11. If four deviations are −2, −1, 0 and +1, the fifth must be +2 to make the sum zero. Only 4 of the 5 deviations carry new information; s^2 divides by 4.

MUST-KNOW NOTES (MD-LEVEL)

- General rule: df = number of independent observations − number of parameters estimated from them before the quantity in question.
- df is not the sample size. The same dataset can give different df in different calculations, depending on how many parameters each had to estimate first. The same subtraction returns in the t-distribution (§5.7.2.3), the chi-square test (§10.6) and regression (Chapter 12).

3.7 Standard Deviation

DEFINITION

The **standard deviation (SD)** is the square root of the variance: $\sigma = \sqrt{\sigma^2}$ for a population, $s = \sqrt{s^2}$ for a sample. It is the **root-mean-square** deviation from the mean, expressed in the original units.

SIMPLIFIED EXPLANATION

Variance squares the units (kg^2), and nobody has a feel for "2.5 kg^2 of spread". The square root returns to kg. Read an SD as a typical distance from the mean that gives large deviations extra weight. It is *not* the average distance from the mean: that is the mean deviation (§3.5), and the SD is always at least as large as the MD.

ILLUSTRATION / EXAMPLE

Five weights: $s = \sqrt{2.5} \approx$ **1.58 kg**. This is the root-mean-square distance from the mean. It exceeds the mean deviation (1.2 kg) because squaring gives the two children 2 kg from the mean more weight than the children 1 kg away.

Today's 30 children: $s = \sqrt{3.07} =$ **1.75 kg** (MD 1.41 kg).

WORKED CALCULATION: SD FROM GROUPED DATA

With only the table, use mid-points x and frequencies f (grouped mean 11.8 kg, recall §2.2.1): $s^2 = \Sigma f(x - \text{mean})^2 \div (N - 1)$.

Mid-point x	f	x − 11.8	$(x - 11.8)^2$	$f(x - 11.8)^2$
9	5	−2.8	7.84	39.20
11	12	−0.8	0.64	7.68
13	9	1.2	1.44	12.96
15	4	3.2	10.24	40.96
Total	30			100.80

1. $s^2 = 100.80 \div 29 = 3.48\ \text{kg}^2$.
2. $s = \sqrt{3.48} =$ **1.86 kg**, against 1.75 kg from the raw data. Grouping moves every child to a class mid-point, and with these classes that slightly exaggerates the spread.

MUST-KNOW NOTES (MD-LEVEL)

- SD = √variance. The SD with divisor n is at least as large as the MD, equal only when every deviation has the same size; s, with divisor n − 1, is larger still.
- The usual summary for roughly symmetric interval or ratio data is mean ± SD.
- A large or small SD means something only relative to the variable's scale: an SD of 5 is large for haemoglobin in g/dL and small for age in years. The coefficient of variation (§3.8) makes that comparison explicit.
- The SD sets the spread (scale) of the Normal distribution (the bell-shaped curve of §5.7.2.2); every Normal curve has the same shape. It is also the building block of the standard error (§6.7) and of confidence intervals (§7.2).

CHECKPOINT 3.2

1. Explain why variance squares the deviations instead of averaging them directly, and name one property squared deviations have that absolute deviations lack.
2. For the six haemoglobin values in Checkpoint 3.1 (9.0, 10.5, 11.0, 11.5, 12.0, 14.0 g/dL), compute the mean, the sum of squares, s^2 and s. Then compute SS ÷ n and say why it is not the usual estimate of σ^2.
3. A student writes: "SD = 1.66 g/dL, so on average each woman's haemoglobin differs from the mean by 1.66 g/dL." Correct the sentence, using the mean deviation of the same data.

(Answers in the Appendix at the end of the book.)

3.8 Coefficient of Variation

DEFINITION

The **coefficient of variation (CV)** is the SD expressed as a percentage of the mean: CV = (SD ÷ mean) × 100%. It is a unit-free measure of relative spread.

SIMPLIFIED EXPLANATION

An SD makes sense only against the scale of the variable. You cannot say whether height (SD in cm) or weight (SD in kg) is "more variable" by comparing raw SDs. Dividing by the mean removes the units.

ILLUSTRATION / EXAMPLE

Heights with mean 150 cm and SD 10 cm → CV = 10 ÷ 150 = **6.7%**. Weights with mean 40 kg and SD 8 kg → CV = 8 ÷ 40 = **20%**. Weight's raw SD is smaller, yet weight is proportionally three times as variable.

Today's 30 children: CV = 1.75 ÷ 11.81 × 100 = **14.8%**.

MUST-KNOW NOTES (MD-LEVEL)

- CV = (SD ÷ mean) × 100%. Unit-free, so it compares variability across different variables, units or studies.
- Meaningful only for ratio-scale data with a positive mean. It fails for interval data such as temperature in °C, where the mean depends on an arbitrary zero, and it becomes unstable when the mean is near zero.
- In the laboratory, the CV of repeated measurements of one sample is used to judge an assay's precision (Reed et al., 2002).

3.9 Skewness

DEFINITION

Skewness measures the asymmetry of a distribution: whether one tail stretches further than the other. **Positive (right) skew** means a long tail toward high values; **negative (left) skew**, a long tail toward low values.

SIMPLIFIED EXPLANATION

The measures so far say *how much* the data spread, not the *shape* of the spread. Two datasets can share a mean and SD while one is symmetric and the other has a long tail. Skewness puts a number on that asymmetry.

ILLUSTRATION / EXAMPLE

Hospital length of stay is a classic right-skewed variable: most patients stay 2–4 days, a few stay weeks. The few long stays pull the mean up, which is why length of stay is usually summarised by the median (recall §2.5).

In a unimodal right-skewed distribution the order is **usually** mean > median > mode, reversed for left skew, and all three coincide in a symmetric distribution. This holds for the smooth, single-peaked curves drawn in textbooks, and is guaranteed for broad families of them, but it can fail, and "violations are not at all unusual if the distribution is discrete" (von Hippel, 2005). Treat it as a rule of thumb for continuous, single-peaked data.

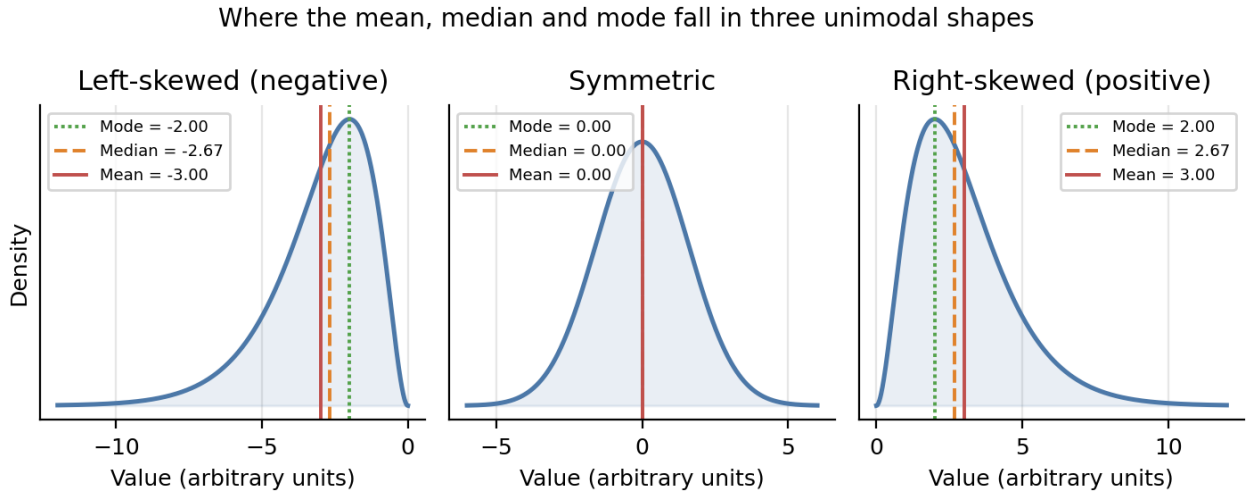


Figure 3.1. Three unimodal shapes. In the right-skewed curve (a gamma distribution with shape 3) the mode is 2.00, the median 2.67 and the mean 3.00; the left-skewed curve is its mirror image. In the symmetric curve all three coincide. The empirical relation of §2.4 gives 3 × 2.674 − 2 × 3 = 2.02 for the right-skewed mode, close to the true 2.00.

DERIVATION

Three families of coefficient are in use.

1. **Pearson's coefficients** use the gap between centres, scaled by the SD. First coefficient: (mean − mode) ÷ SD. The mode is hard to estimate, so substitute the empirical relation mean − mode ≈ 3(mean − median) (recall §2.4) to get the second coefficient: **Sk = 3(mean − median) ÷ SD**, which NIST/SEMATECH calls the "Pearson 2 skewness coefficient".
2. **Bowley's quartile coefficient** uses positions only: (Q3 + Q1 − 2Q2) ÷ (Q3 − Q1). It compares the upper half of the box, Q3 − Q2, with the lower half, Q2 − Q1. It is zero when they are equal. NIST/SEMATECH lists it as Galton skewness, "also known as Bowley's skewness".
3. **Moment coefficient.** Let $m_k = \Sigma(x - \bar{x})^k \div n$ (the kth moment about the mean; m_2 is the variance with divisor n). Cubing keeps the sign of each deviation and exaggerates large ones, so a long right tail makes m_3 positive. Divide by m_2^(3/2) to remove the units: $\mathbf{g_1 = m_3 \div m_2^{(3/2)}}$, the Fisher–Pearson coefficient of skewness (NIST/SEMATECH). It is 0 for a normal distribution. Many programs report an adjusted version for samples.

WORKED CALCULATION: TODAY'S 30 CHILDREN

1. Pearson: Sk = 3 × (11.81 − 11.7) ÷ 1.75 = **0.19**, slightly positive.
2. Bowley (type 7 quartiles): (13.05 + 10.575 − 2 × 11.7) ÷ 2.475 = 0.225 ÷ 2.475 = **0.09**.
3. Moment: m_2 = 88.95 ÷ 30 = 2.965; m_3 = 0.709; g_1 = 0.709 ÷ 2.965^1.5 = 0.709 ÷ 5.106 = **0.14**.

All three agree that these weights are nearly symmetric, with a faint right lean. For contrast, the five weights with the 2 kg child (2, 10, 11, 12, 13; §2.3) give Pearson Sk = 3 × (9.6 − 11) ÷ 4.39 = **−0.96** and g_1 = **−1.26**: clear negative skew from one low value.

MUST-KNOW NOTES (MD-LEVEL)

- Formulas: Pearson Sk = 3(mean − median) ÷ SD (the first coefficient uses mean − mode); Bowley = (Q3 + Q1 − 2Q2) ÷ (Q3 − Q1); moment $g_1 = m_3 \div m_2$^(3/2). Sign: positive = right skew, negative = left skew, 0 = symmetric.
- A mean noticeably above the median usually signals right skew in unimodal data, and suggests reporting median and IQR instead of mean and SD.
- Right-skewed variables in community medicine include length of stay, income, many laboratory values (CRP, bilirubin) and times to an event.
- Many parametric tests, such as the t-test (§9.2), assume roughly Normal data. Marked skew is a reason to consider a transformation (§9.4) or a non-parametric method (§10.1).

3.10 Kurtosis

DEFINITION

Kurtosis measures the **tailedness** of a distribution: how prone it is to produce extreme values, relative to the Normal distribution. It is computed as $b_2 = m_4 \div m_2^2$, which equals 3 for a Normal distribution; **excess kurtosis** is $b_2 - 3$ (NIST/SEMATECH).

SIMPLIFIED EXPLANATION

Skewness captures lean to one side. Kurtosis captures something else: how heavy both tails are. Raising deviations to the fourth power makes the few values far from the mean dominate m_4, so kurtosis is driven almost entirely by the tails. Older books describe kurtosis as "peakedness". That is misleading: kurtosis "tells you virtually nothing about the shape of the peak"; its only clear meaning is tail extremity (Westfall, 2014). A symmetric distribution, with zero skew, can be light-tailed or heavy-tailed.

ILLUSTRATION / EXAMPLE

Heavy-tailed (**leptokurtic**, excess kurtosis > 0) distributions produce more extreme values than a Normal distribution with the same SD, which matters when you judge whether a Normal-based method is safe. Light-tailed (**platykurtic**, excess < 0) distributions produce fewer; they often, but not necessarily, look flatter near the centre. **Mesokurtic** distributions have Normal-like tails (excess ≈ 0).

WORKED CALCULATION: TODAY'S 30 CHILDREN

1. $m_4 = \Sigma(x - 11.81)^4 \div 30 = 19.61$; $m_2 = 2.965$, so $m_2^2 = 8.792$.
2. $b_2 = 19.61 \div 8.792 =$ **2.23**; excess kurtosis = 2.23 − 3 = **−0.77**.

The weights are mildly platykurtic: the lightest and heaviest children (8.6 and 15.3 kg) are only about 1.8 and 2.0 SDs from the mean, so there are no long tails. Software often reports an adjusted sample excess kurtosis (−0.68 here), so check which definition a program uses.

MUST-KNOW NOTES (MD-LEVEL)

- $b_2 = m_4 \div m_2^2$; Normal = 3; excess kurtosis = $b_2 - 3$. Many sources and programs say "kurtosis" when they mean excess kurtosis.
- Leptokurtic = heavier tails than Normal; platykurtic = lighter; mesokurtic = Normal-like. Kurtosis describes tails, not the peak.
- Skewness and kurtosis are descriptive signals, not a pass/fail test of normality. Combine them with a histogram, a Q-Q plot and, where relevant, a formal normality test (all in §9.3).

CHECKPOINT 3.3

1. You want to compare how variable height is with how variable weight is in the same children. Why can't you compare raw SDs, and what do you compare instead? Compute it for heights with mean 85 cm, SD 4 cm and weights with mean 11.8 kg, SD 1.75 kg.
2. Describe a right-skewed distribution of hospital length of stay, and state the usual order of mean, median and mode. In one ward the mean stay is 6 days, the median 5 days and the SD 4 days: compute Pearson's coefficient of skewness and the mode implied by the empirical relation.
3. A colleague says a platykurtic distribution "must have a flat top". Is that right? Explain what kurtosis actually measures.

(Answers in the Appendix at the end of the book.)

Data Visualisation

Chapters 2 and 3 reduced a dataset to a few numbers: a centre, a spread, a shape. A picture often shows what those numbers hide. Two variables can share the same mean, SD and skewness while a histogram shows that one of them has two peaks, perhaps because two subpopulations were mixed. This chapter covers tables (§4.1), diagrams for categorical data (§4.2), diagrams for the distribution of a numeric variable (§4.3), the scatter diagram for two numeric variables (§4.4), and the diagrams community medicine uses for time, place and population (§4.5). The chart you pick follows from the variable's scale of measurement (recall §1.1.2) and from the question you want the reader to answer.

4.1 Tables for Data Presentation

DEFINITION

A **table** arranges data in rows and columns under clear headings, so that it can be read and compared without any graphical element. A frequency distribution (recall §1.4.1) is one kind of table; a **cross-tabulation** (two-way table) is another, showing how two categorical variables are distributed jointly.

SIMPLIFIED EXPLANATION

A well-labelled grid is already a large gain in readability over a raw list. Tables also do something charts cannot: they keep exact values. A chart trades exactness for visual impact.

ILLUSTRATION / EXAMPLE

A table listing several districts down the rows and their CBHI membership percentages across the columns lets you compare exact figures. A chart of the same numbers would let you judge them only by eye.

MUST-KNOW NOTES (MD-LEVEL)

- A good table has a title that says what, where and when; labelled rows and columns with units; totals where they help; and a footnote naming the data source. Give all four in every thesis or manuscript table.
- Use a table when exact values matter (for example, when a reviewer must check your figures); use a chart when the shape or trend matters more than the exact numbers.
- The cross-tabulation is the starting point for the chi-square test of independence (§10.6.1, a test of whether two categorical variables are associated).

4.2 Diagrams for Qualitative Data

Categorical data, such as blood group, nutritional status or CBHI membership, have no continuous scale, so each category gets its own separate bar or wedge. The same diagrams also suit ungrouped discrete counts, such as the number of ANC visits (recall §1.4.2); continuous data need a histogram instead (§4.3.1). By convention the bars stand apart with gaps, which tells the reader that the categories are separate classes, whether they are unordered (water source) or ordered (normal, moderate, severe wasting). The two standard forms are the bar diagram (§4.2.1) and the pie diagram (§4.2.2).

4.2.1 Bar Diagram

DEFINITION

A **bar diagram** represents each category by a rectangular bar whose length (or height) is proportional to the frequency, or to some other value, of that category.

SIMPLIFIED EXPLANATION

To compare "how many" across categories, draw one bar per category, make its length proportional to the count, and let the eye compare lengths. Three variants are used:

- **Simple bar**: one bar per category of one variable.
- **Multiple (grouped) bar**: bars for subgroups stand side by side within each category, for example boys and girls within each age group.
- **Component (stacked) bar**: each bar is divided into segments for the subgroups, so it shows the total and its make-up at once. When every bar is scaled to 100%, it becomes a **percentage (100%) component bar**, which compares the make-up of groups of different sizes.

DERIVATION

Why must the value axis start at zero? A bar of value a drawn from a baseline c has length a − c. For two bars, a and b:

drawn ratio = (b − c) ÷ (a − c)

This equals the true ratio b ÷ a only when c = 0. If 0 < c < a < b, the drawn ratio is always larger than the true one, so a raised baseline always exaggerates the difference. The CDC's teaching text states that breaks in the scale should not be used in bar charts (CDC, 2012a). §4.5.6 works through a numeric example.

ILLUSTRATION / EXAMPLE

In the under-five clinic register of 100 children, 60 have normal nutritional status, 25 moderate wasting and 15 severe wasting. Three bars show at once that severe wasting is the smallest group.

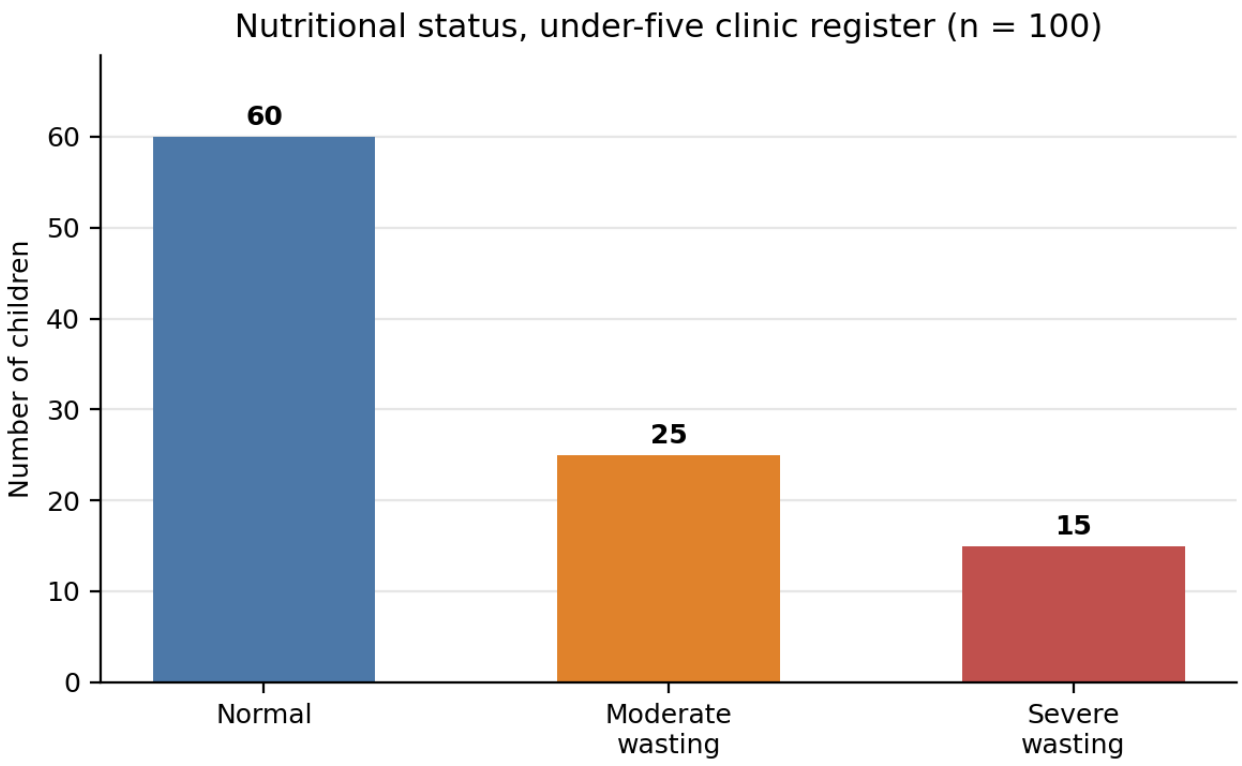


Figure 4.1. Simple bar diagram: nutritional status of the 100 children in the under-five clinic register.

Now split the same 100 children by wealth quintile (Q1 poorest).

WORKED CALCULATION

1. Quintile sizes are unequal: Q1 13, Q2 26, Q3 21, Q4 20, Q5 20 children. In a component bar of counts (Figure 4.2, left) the Q2 bar is twice as tall as the Q1 bar, so the eye compares totals, not make-up.
2. Convert each segment to a percentage of its own bar. In Q1, 1 of 13 children is normal: 1 ÷ 13 × 100 = 7.7%. Severe wasting in Q1: 6 ÷ 13 × 100 = 46.2%.
3. In Q2, 9 of 26 are normal: 34.6%. In Q3, 15 of 21: 71.4%. In Q4, 15 of 20: 75.0%. In Q5, 20 of 20: 100%.
4. Result: in the percentage bar (Figure 4.2, right) every bar has the same height, so the gradient in normal status, from 7.7% in the poorest quintile to 100% in the richest, is plain to see.

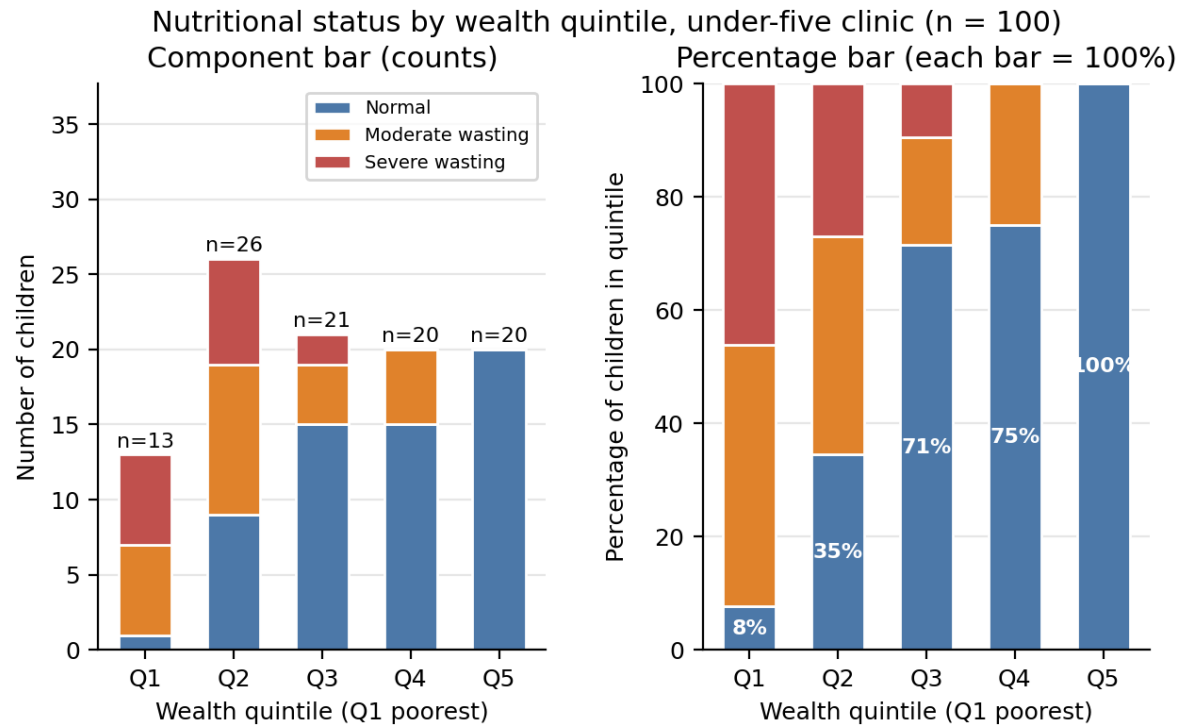


Figure 4.2. Component bar (counts, left) and percentage component bar (right): nutritional status by wealth quintile, under-five clinic register (n = 100).

MUST-KNOW NOTES (MD-LEVEL)

- Use a component bar of counts when the totals matter, and a percentage bar when you want to compare composition across groups of different sizes. The CDC describes the 100% component bar as useful "for comparing the contribution of different subgroups within the categories" (CDC, 2012a).
- Only the bottom segment of a stacked bar sits on a common baseline. The other segments float, so they are harder to compare across bars; put the segment you most want compared at the bottom.
- With many categories a bar diagram becomes cluttered. Order the bars by size (unless the categories have a natural order), draw them horizontally, or use a table.

4.2.2 Pie Diagram

DEFINITION

A **pie diagram** is a circle divided into wedges whose angles (and therefore areas) are proportional to each category's share of the total. It shows how one whole divides into parts.

SIMPLIFIED EXPLANATION

A bar diagram compares magnitudes; a pie diagram answers "what share of the whole?". Each wedge's central angle is:

angle = (category frequency ÷ total) × 360°

So the full circle is always 100%.

ILLUSTRATION / EXAMPLE

The pie in Figure 4.3 shows the water source of surveyed households: piped water 45%, well 30%, river 15% and other 10%. The "other" wedge is what makes the categories exhaustive; without it the wedges would not add to the whole circle.

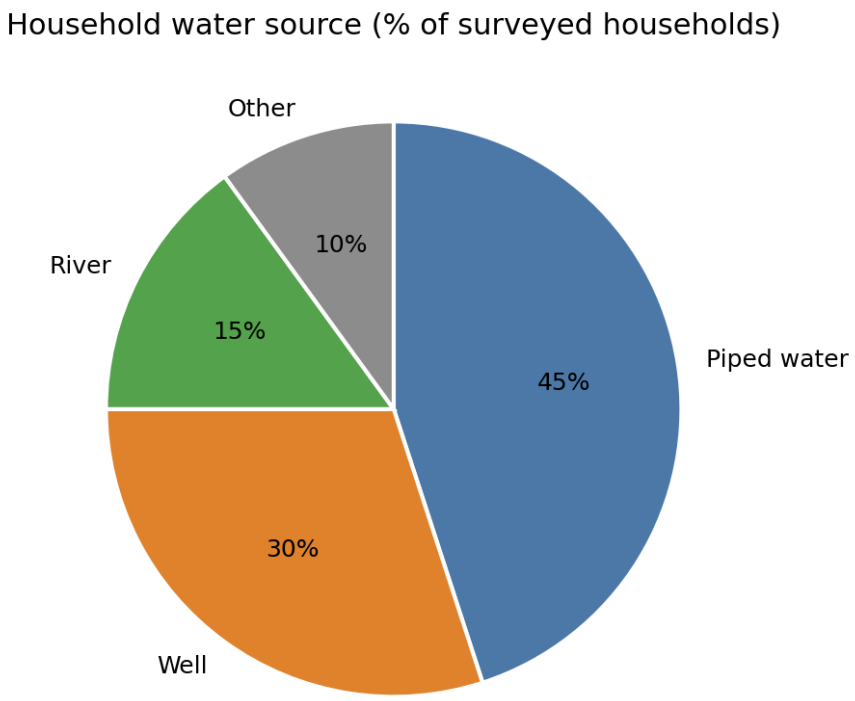


Figure 4.3. Pie diagram of household water source (% of surveyed households), drawn from 12 o'clock, largest to smallest clockwise.

WORKED CALCULATION

In the CBHI survey, 350 households did not join the scheme. Their reasons were lack of knowledge 112, lack of money 98, lack of trust 67, bad image 46, other 27.

1. Lack of knowledge: 112 ÷ 350 × 360° = 115.2°.
2. Lack of money: 98 ÷ 350 × 360° = 100.8°.
3. Lack of trust: 67 ÷ 350 × 360° = 68.9°.
4. Bad image: 46 ÷ 350 × 360° = 47.3°.
5. Other: 27 ÷ 350 × 360° = 27.8°.
6. Check: 115.2 + 100.8 + 68.9 + 47.3 + 27.8 = 360.0°. The shares are 32.0%, 28.0%, 19.1%, 13.1% and 7.7%.

MUST-KNOW NOTES (MD-LEVEL)

- A pie needs categories that are mutually exclusive and exhaustive, adding to 100%. A bar diagram does not have this requirement.
- People judge angles and areas less accurately than lengths or positions on a common scale. Cleveland and McGill ranked "position along a common scale" first and angle and area lower, and concluded that "a pie chart can always be replaced by a bar chart" (Cleveland & McGill, 1984). Pie diagrams remain common in reports, so you must be able to read and draw them.
- Choose between the two by the question, not only by the number of categories. If the reader must compare categories with each other (is lack of money as common as lack of knowledge?), use bars. If the single message is the share of one whole, a pie with a few wedges can work.
- A common construction convention: start at 12 o'clock, order wedges from largest to smallest clockwise, and label each wedge with its percentage (CDC, 2012a).

CHECKPOINT 4.1

1. The five reasons for not joining CBHI (112, 98, 67, 46 and 27 households) are to go in your thesis. You want the reader to compare the reasons with one another. Would a bar diagram or a pie diagram serve better, and why?
2. Why must a bar diagram's value axis start at zero? Use the drawn-ratio formula to show what happens if it starts at 50 for two bars of 60 and 80.
3. In Figure 4.2, Q3 has 15 normal, 4 moderately wasted and 2 severely wasted children. What height does the moderate-wasting segment have in the percentage bar?

(Answers in the Appendix at the end of the book.)

4.3 Diagrams for Quantitative Data

A numeric variable, measured on an interval or ratio scale (recall §1.1.2), usually has too many distinct values to give each its own bar. The charts in this section therefore start from the grouped frequency distribution (recall §1.4.2) or from the quartiles (recall §3.3). They show the distribution's shape, centre and spread: the histogram (§4.3.1), the frequency polygon (§4.3.2), the ogive (§4.3.3), the box plot (§4.3.4) and the stem-and-leaf plot (§4.3.5).

4.3.1 Histogram

DEFINITION

A **histogram** draws each class interval of a grouped frequency distribution as a bar on a continuous numeric axis, with adjoining bars and no gaps between them, and with bar **area** proportional to the class frequency. With equal class widths, bar height is proportional to frequency too.

SIMPLIFIED EXPLANATION

Take the grouped weights of today's 30 clinic children (recall §1.4.1): 8 to <10 kg, 5 children; 10 to <12 kg, 12; 12 to <14 kg, 9; 14 to <16 kg, 4. Draw each class as a bar from its lower to its upper boundary. The bars touch because weight flows continuously from one class into the next; there is no real gap between 9.99 kg and 10.00 kg. The CDC describes the histogram as using "adjoining columns" for the class intervals (CDC, 2012b).

ILLUSTRATION / EXAMPLE

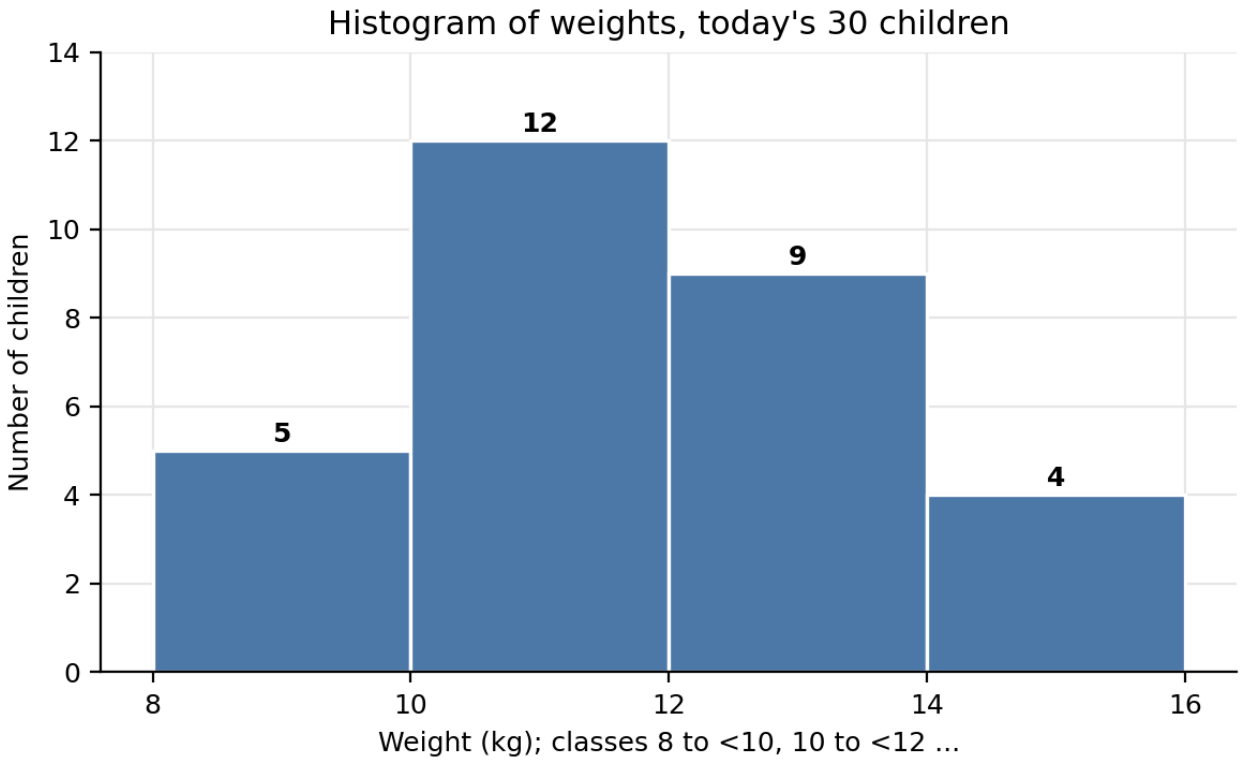


Figure 4.4. Histogram of the weights of today's 30 clinic children, classes 8 to <10, 10 to <12, 12 to <14 and 14 to <16 kg.

The histogram peaks in the 10 to <12 kg class and is only mildly asymmetric, consistent with the small Pearson skewness of today's weights (recall §3.9).

WORKED CALCULATION

Suppose you merged the last two classes into one class, 12 to <16 kg, with 9 + 4 = 13 children.

1. If you drew its bar at height 13, it would stand taller than the 10 to <12 kg bar (12) and be twice as wide, so it would look far more crowded.
2. Use **frequency density** instead: frequency ÷ class width. For 12 to <16 kg: 13 ÷ 4 = 3.25 children per kg.
3. Compare: 8 to <10 kg, 5 ÷ 2 = 2.5 per kg; 10 to <12 kg, 12 ÷ 2 = 6.0 per kg.
4. Result: drawn at height 3.25, the merged bar has area 3.25 × 4 = 13, the right share, and correctly shows the 12 to <16 kg range as about half as crowded per kg as 10 to <12 kg.

MUST-KNOW NOTES (MD-LEVEL)

- Adjoining bars on a continuous axis distinguish a histogram from a bar diagram.
- With unequal class widths, plot frequency density so that area, not height, carries the frequency.
- The histogram is the quickest informal look at skewness (recall §3.9) and at whether a variable looks roughly bell-shaped, before you use the Normal distribution (§5.7.2.2) or formal tests of normality (§9.3).
- The number of classes changes the picture: too few hides detail, too many shows noise (recall Sturges' rule, §1.4.1).

4.3.2 Frequency Polygon

DEFINITION

A **frequency polygon** plots each class frequency at the class mid-point and joins the points with straight lines. It is closed by adding a point of zero frequency one class width beyond each end.

SIMPLIFIED EXPLANATION

Put a dot at the top-centre of each histogram bar and join the dots. The polygon shows the same shape as the histogram, as an outline rather than blocks. It is not smoothed: the lines are straight segments between mid-points. Its practical use is comparison: two polygons, for example urban and rural clinic weights, can share one set of axes, where two overlapping histograms would be cluttered.

DERIVATION

With equal class widths h, the closed polygon has exactly the same area as the histogram. Take any segment joining two adjacent mid-points. At the class boundary halfway between them, the segment cuts a small triangle off the corner of one bar and adds a triangle of the same base (h/2) and the same height to the neighbouring region. The two triangles are congruent, so the area lost equals the area gained. The two closing segments do the same at each end: the triangle added outside the first class equals the triangle cut from its corner. Summing over all segments, polygon area = histogram area = $h \times N$. The CDC states the same result: a frequency polygon "contains the same area under the line as does a histogram of the same data" (CDC, 2012b).

WORKED CALCULATION

1. Histogram area = $\Sigma f \times h = (5 + 12 + 9 + 4) \times 2 = 60$ child-kg.
2. Polygon points: (7, 0), (9, 5), (11, 12), (13, 9), (15, 4), (17, 0).
3. Area by trapezoids, each of width 2: $2 \times [(0 + 5)/2 + (5 + 12)/2 + (12 + 9)/2 + (9 + 4)/2 + (4 + 0)/2] = 2 \times (2.5 + 8.5 + 10.5 + 6.5 + 2.0) = 2 \times 30 = 60$ child-kg.
4. Result: the two areas are equal, as the derivation predicts.

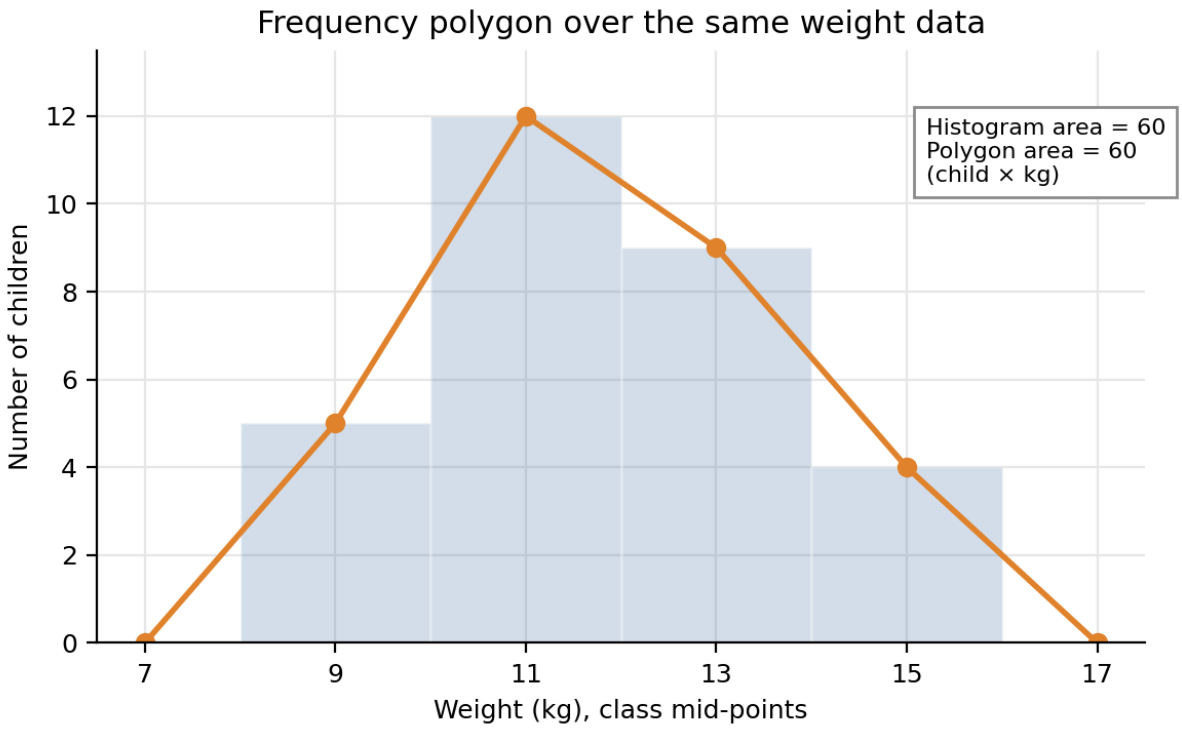


Figure 4.5. Frequency polygon (line) over the histogram of today's 30 weights (shaded). Both enclose 60 child-kg.

MUST-KNOW NOTES (MD-LEVEL)

- The polygon uses the same mid-points and frequencies as the histogram. It is a redrawing of the same data, not new data.
- As the classes narrow and the number of observations grows, the polygon drawn on the frequency-density scale (recall §4.3.1), with total area rescaled to 1, approaches a smooth curve. That is the idea behind the probability density curve of a continuous random variable (§5.6.2, a curve whose area over an interval gives a probability).

4.3.3 Ogive

DEFINITION

An **ogive** is a cumulative frequency curve. The usual "less than" ogive plots the cumulative frequency (or cumulative percentage) against the upper boundary of each class. A "more than" ogive plots the number of observations at or above each lower class boundary.

SIMPLIFIED EXPLANATION

Cumulative frequency is a running total (recall §1.4.3); the ogive draws that running total as a line. Because a running total can only grow or stay the same, the "less than" ogive never falls. It is usually S-shaped: steep where the data are dense, flat in the tails. The "more than" ogive is its mirror image and never rises.

ILLUSTRATION / EXAMPLE

For today's 30 weights, the "less than" ogive passes through (8, 0), (10, 5), (12, 17), (14, 26), (16, 30). The "more than" ogive passes through (8, 30), (10, 25), (12, 13), (14, 4), (16, 0).

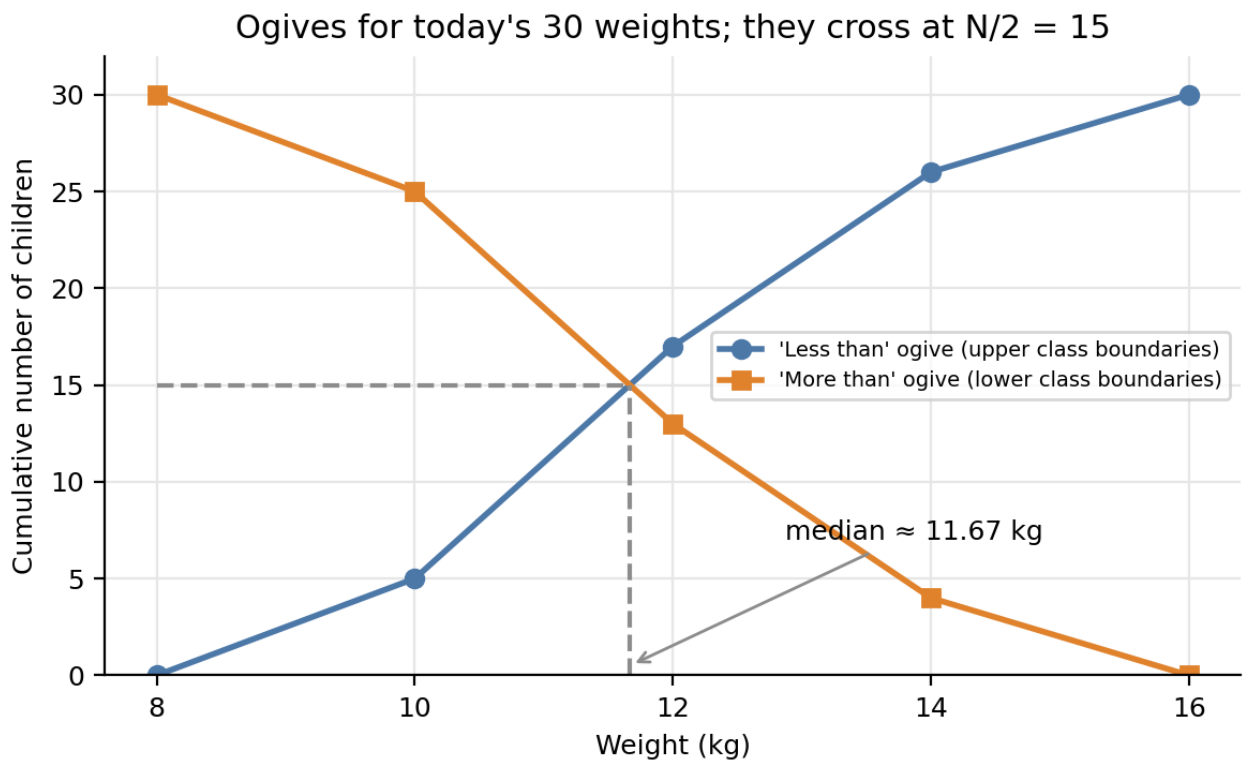


Figure 4.6. "Less than" and "more than" ogives for today's 30 weights. They cross at a cumulative count of 15 (N/2), above the median.

WORKED CALCULATION

Read the median from the ogive.

1. The median is the value with N/2 = 30/2 = 15 children below it.
2. On the "less than" ogive, 15 lies between (10, 5) and (12, 17).
3. Interpolate along the straight segment: 10 + (15 − 5) ÷ (17 − 5) × 2 = 10 + 10/12 × 2 = 11.67 kg.
4. Check with the "more than" ogive: on the segment from (10, 25) to (12, 13), the count 15 is reached at 10 + (25 − 15) ÷ (25 − 13) × 2 = 11.67 kg. The two curves cross there.
5. Result: graphical median = 11.67 kg, the same value as the grouped-data median formula of §2.3. The median of the 30 raw weights is 11.7 kg; grouping costs very little here.

MUST-KNOW NOTES (MD-LEVEL)

- Any percentile can be read the same way: find the cumulative count (or percentage) on the y-axis, go across to the curve, then down to the x-axis.
- The "less than" and "more than" ogives always cross at the median, because at that point the counts below and above are both N/2.

- Joining the plotted points with straight lines is the same assumption the grouped-median formula makes: values are spread evenly within each class.

CHECKPOINT 4.2

1. Name the one visual feature that tells you at once whether you are looking at a histogram or a bar diagram, and say why that feature exists.
2. Using the "less than" ogive points for today's 30 weights, estimate the first quartile (the weight with 7.5 children below it).
3. A histogram of haemoglobin uses classes 7 to <9, 9 to <11 and 11 to <15 g/dL, with 10, 30 and 40 women. What height should each bar have, and which class is the most crowded per g/dL?

(Answers in the Appendix at the end of the book.)

4.3.4 Box Plot

DEFINITION

A **box plot** draws a box from the first quartile (Q1) to the third quartile (Q3), with a line at the median, and whiskers out from the box. Two versions are in use:

- **Five-number-summary box plot**: whiskers run to the minimum and the maximum.
- **Tukey box plot** (R's default box plot): compute the **fences** Q1 − 1.5 × IQR and Q3 + 1.5 × IQR. Each whisker runs to the most extreme observation that lies inside its fence, and every observation beyond a fence is plotted as a separate point, a potential outlier. The whisker ends at a data value, not at the fence itself.

SIMPLIFIED EXPLANATION

The box is the IQR, the middle 50% of the data (recall §3.4); the line inside it is the median; the whiskers show how far the rest of the data reach. The box plot was introduced by John Tukey as part of exploratory data analysis. Variants exist: some use other multipliers than 1.5, some draw whiskers to fixed percentiles, and software differs in how it computes quartiles (Wickham & Stryjewski, 2011). The 1.5 × IQR rule is therefore a convention, and a figure legend should say which version is drawn.

ILLUSTRATION / EXAMPLE

In the CBHI survey, distance from home to the nearest health facility was recorded for all 500 households (Figure 4.7).

WORKED CALCULATION

Non-members (n = 350):

1. Q1 = 6.6 km, median = 8.5 km, Q3 = 10.2 km, so IQR = 10.2 − 6.6 = 3.6 km.
2. 1.5 × IQR = 5.4 km. Lower fence = 6.6 − 5.4 = 1.2 km; upper fence = 10.2 + 5.4 = 15.6 km.
3. The maximum, 15.3 km, is inside the upper fence, so the upper whisker ends at 15.3 km.

4. Fourteen households live less than 1.2 km away (from 0.0 to 1.1 km). They are plotted as separate points. The nearest value inside the fence is 1.3 km, so the lower whisker ends at 1.3 km, not at the fence (1.2 km) and not at the minimum (0.0 km).

Members (n = 150):

1. Q1 = 3.5 km, median = 6.9 km, Q3 = 9.2 km (9.175 before rounding), so IQR = 5.675 km.
2. Fences: 3.5 − 8.5 = −5.0 km and 9.175 + 8.5 = 17.7 km. The minimum (0.2 km) and maximum (14.7 km) both lie inside, so the whiskers reach them. For members, the Tukey and five-number box plots are identical.

Result: members tend to live nearer the facility (median 6.9 vs 8.5 km). The non-member distribution is skewed to the left, with a tail of households living next to the facility.

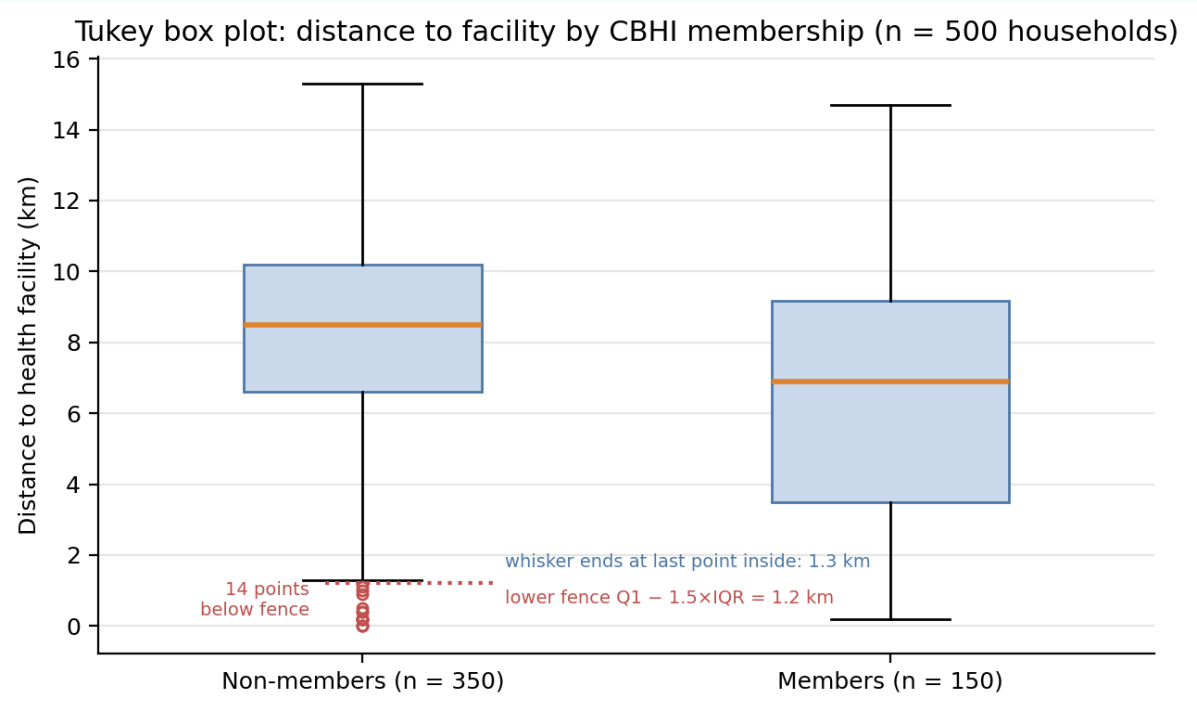


Figure 4.7. Tukey box plots of distance to the health facility, CBHI survey (350 non-members, 150 members). Whiskers end at the last observation inside the fences; the 14 points below the non-members' lower fence are drawn separately.

MUST-KNOW NOTES (MD-LEVEL)

- The box plot is at its best when it compares one numeric variable across groups on the same axis, for example birthweight across several blocks.
- Asymmetry shows skewness without calculation: the median off-centre in the box, or one whisker longer than the other. In Figure 4.7 the non-members' lower tail, with its 14 outliers, reaches 8.5 km below the median (to 0.0 km), while the upper tail reaches only 6.8 km above it (to 15.3 km): a left skew.
- A point beyond a fence is a *potential* outlier to check against the source records. It is not an error by definition, and it is not deleted for being unusual. In a right-skewed variable such as hospital length of stay, the long stays beyond the upper fence are real and important.

4.3.5 Stem-and-Leaf Plot

DEFINITION

A **stem-and-leaf plot** splits each value into a **stem** (the leading digits) and a **leaf** (the last digit), and lists the leaves beside their stems. It shows the shape of the distribution and keeps every value.

SIMPLIFIED EXPLANATION

The histogram, frequency polygon, ogive and box plot give up individual values: once a weight goes into a class or a quartile, you cannot read it back. Of the five charts in this section, the stem-and-leaf plot is the only one that keeps every value. Turned on its side, it looks like a histogram. The scatter diagram (§4.4) also keeps every value, but it plots two variables.

ILLUSTRATION / EXAMPLE

Today's 30 weights, with whole kilograms as the stem and the first decimal as the leaf:

Stem (kg)	Leaves (0.1 kg)
8	6
9	1 4 7 8
10	0 3 5 8
11	0 1 2 4 5 6 8 9
12	0 3 4 6 9
13	1 4 6 8
14	1 4 8
15	3

The row lengths (1, 4, 4, 8, 5, 4, 3, 1; total 30) trace the shape, with the peak at 11 kg, and you can still read off every value: the lightest child weighed 8.6 kg and the heaviest 15.3 kg.

MUST-KNOW NOTES (MD-LEVEL)

- It keeps exact values, so medians and quartiles can be counted straight off the display.
- It suits small datasets. With hundreds of values the rows become too long, and a histogram shows the shape more efficiently.
- Choose the stem unit so that you get a sensible number of rows, much as you choose class width for a histogram.

4.4 Scatter Diagram

DEFINITION

A **scatter diagram** plots paired values of two quantitative variables as points on a two-dimensional graph, one variable on each axis, to show how they relate.

SIMPLIFIED EXPLANATION

The charts in §4.2–4.3 show one variable at a time. The scatter diagram shows two measurements per person and lets the eye judge whether they rise together, move in opposite directions, or show no pattern.

ILLUSTRATION / EXAMPLE

Figure 4.8 plots weight against age for the 100 children in the under-five clinic register (ages 12–36 months, weights 8.6–15.4 kg). The cloud rises from lower left to upper right: older children tend to weigh more. Correlation (Chapter 11) puts a number on this and regression (Chapter 12) models it.

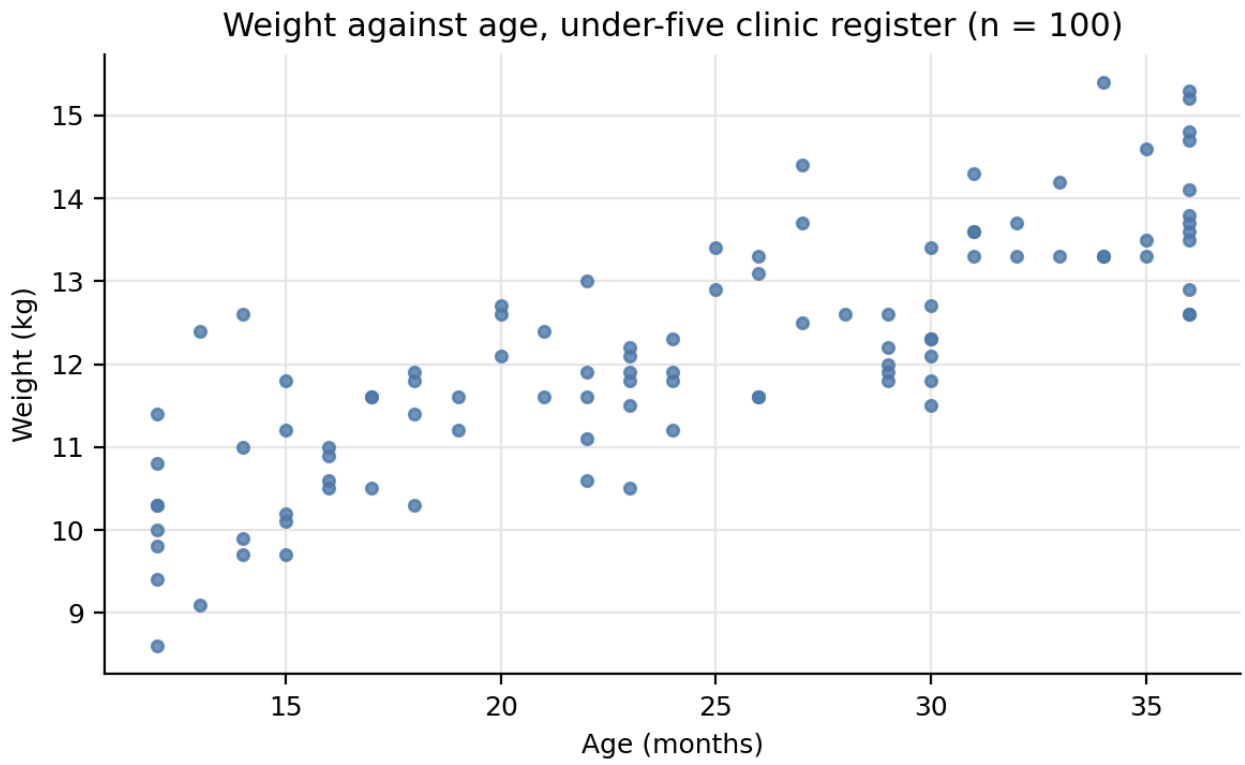


Figure 4.8. Scatter diagram of weight against age, under-five clinic register (n = 100).

MUST-KNOW NOTES (MD-LEVEL)

- Look at the scatter diagram before you compute a correlation coefficient. A single number can mislead badly when the relationship is curved or driven by a few outliers.
- Both variables must be quantitative. For a numeric outcome across the categories of a categorical variable, use side-by-side box plots (§4.3.4).

- Put the explanatory variable (here, age) on the x-axis and the outcome (weight) on the y-axis.

CHECKPOINT 4.3

1. A box plot's median line sits much closer to the top of the box than the bottom. What does that tell you about the shape of the middle half of the data?
2. For CBHI members, Q1 = 3.5 km and Q3 = 9.2 km, and distances run from 0.2 to 14.7 km. Compute the Tukey fences and say whether any member household would be drawn as a separate point.
3. Why might you prefer a stem-and-leaf plot to a histogram for a small dataset?

(Answers in the Appendix at the end of the book.)

4.5 Diagrams for Time, Place and Population

Descriptive epidemiology asks when, where and in whom. Each question has its own diagrams: the line diagram and the epidemic curve for time (§4.5.1–4.5.2), the spot map and the choropleth map for place (§4.5.3), and the pictogram and population pyramid for population (§4.5.4–4.5.5). §4.5.6 then collects the ways graphs mislead. The data in this section are synthetic teaching data, invented for the examples.

4.5.1 Line Diagram and Semi-Logarithmic Graph

DEFINITION

A **line diagram** plots a quantity against time (or another ordered variable), with successive points joined by straight lines. On an **arithmetic-scale** graph equal distances on the y-axis stand for equal amounts. On a **semi-logarithmic** (semi-log) graph the y-axis is logarithmic, so equal distances stand for equal ratios, while the x-axis stays arithmetic.

SIMPLIFIED EXPLANATION

The line diagram is the standard way to show a trend: monthly malaria cases, annual infant deaths, coverage by year. The joined line invites the eye to follow the change from one period to the next. Use it when the x-axis is ordered, usually time; do not join points for unordered categories.

On a semi-log graph the distance from 10 to 100 equals the distance from 100 to 1,000 (CDC, 2012b). That makes the slope show the rate of change (the percentage change per period) rather than the absolute change. The CDC's teaching text puts it this way: "Two or more lines following parallel paths show identical rates of change" (CDC, 2012b).

DERIVATION

Suppose a count grows by the same factor $(1 + r)$ each year: $y_t = y_0 \times (1 + r)^t$.

Take logarithms: $\log y_t = \log y_0 + t \times \log(1 + r)$.

This is a straight line in t with slope log(1 + r). So on a semi-log graph:

- a constant percentage change gives a straight line;
- two series changing by the same percentage give parallel lines, however different their sizes;
- the steeper line is the faster-growing series in relative terms.

On an arithmetic scale, the same series curves upwards, and the larger series tends to look as if it is rising faster, because its absolute changes are larger.

ILLUSTRATION / EXAMPLE

Figure 4.9 shows monthly malaria cases in one district over three years (synthetic data). Each year peaks in August (256, 198 and 159 cases), after the monsoon, and the annual totals fall from 1,738 in 2023 to 1,319 in 2024 and 1,043 in 2025. The line diagram shows both the seasonal cycle and the decline.

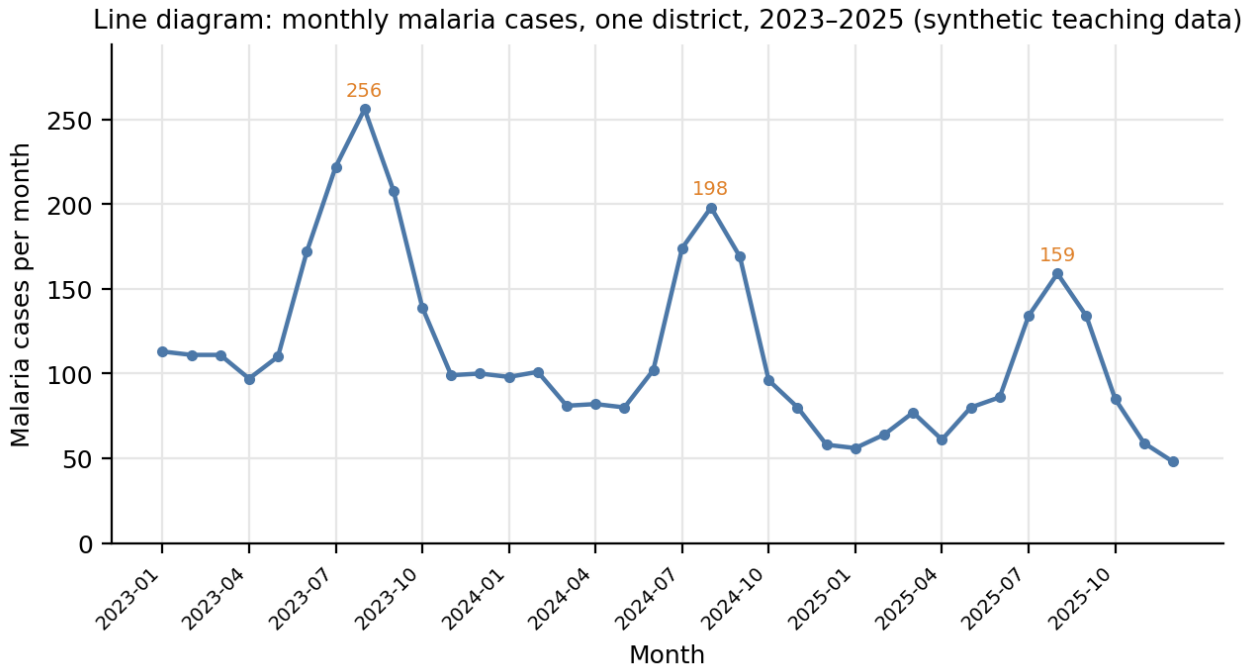


Figure 4.9. Line diagram: monthly malaria cases in one district, January 2023 to December 2025 (synthetic teaching data). August peaks are labelled.

Figure 4.10 shows annual dengue cases in a city and in a small town, 2016–2025 (synthetic data).

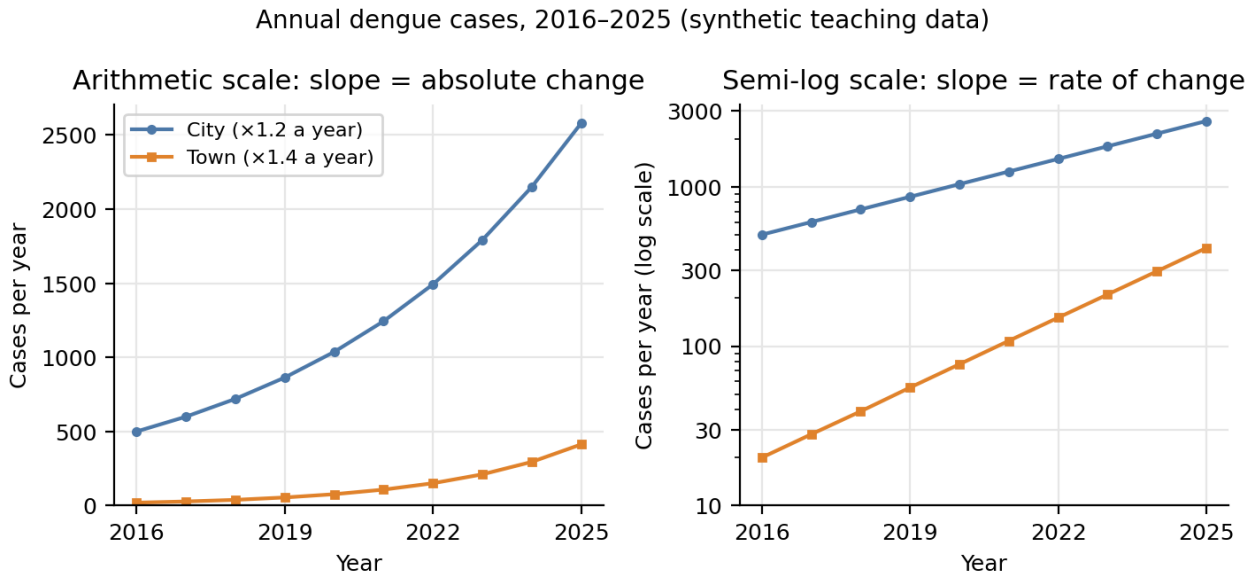


Figure 4.10. The same two series on an arithmetic scale (left) and a semi-logarithmic scale (right). Synthetic teaching data.

WORKED CALCULATION

1. Absolute change, 2016 to 2025. City: 500 to 2,580, a rise of 2,080 cases. Town: 20 to 413, a rise of 393 cases. On the arithmetic panel the city's line climbs steeply and the town's line hugs the axis.
2. Relative change. City: 2,580 ÷ 500 = 5.16-fold. Town: 413 ÷ 20 = 20.65-fold.
3. Annual growth factor (the ninth root, since there are 9 yearly steps). City: 5.16^(1/9) = 1.20, that is 20% a year. Town: 20.65^(1/9) = 1.40, 40% a year.
4. Semi-log slopes ($\log_{10}$ per year). City: $\log_{10}$ 1.2 = 0.079. Town: $\log_{10}$ 1.4 = 0.146. The town's line is almost twice as steep.
5. Doubling time: solve $(1 + r)^t = 2$ for t, giving $t = \log 2 \div \log(1 + r)$. City: 0.301 ÷ 0.079 = 3.8 years. Town: 0.301 ÷ 0.146 = 2.1 years.
6. Result: the city adds more cases each year, but the town's epidemic is growing twice as fast in relative terms. Only the semi-log graph shows this at a glance.

MUST-KNOW NOTES (MD-LEVEL)

- Use an arithmetic scale to show absolute numbers (how many beds, how many cases to treat). Use a semi-log scale to compare rates of change, or to show data spanning several orders of magnitude (Bland, 2015, ch. 5, treats line graphs and logarithmic scales together).
- A logarithmic axis has no zero. Label it clearly as a log scale, because many readers assume an arithmetic axis.
- Seasonal data need several complete years before you judge a trend. Comparing August with December tells you about the season, not the trend.

4.5.2 Epidemic Curve

DEFINITION

An **epidemic curve** (epi curve) is a histogram of the number of cases in an outbreak by time of onset of illness (CDC, 2012b).

SIMPLIFIED EXPLANATION

Time is continuous, so the bars adjoin, as in any histogram (recall §4.3.1). The x-axis interval matters. The CDC's convention is an interval of about one-quarter of the incubation period, and anywhere between one-eighth and one-third of it (CDC, 2012b). Too wide an interval merges the peaks; too narrow scatters single cases.

The shape suggests how the outbreak spread (CDC, n.d.-a):

- **Point source** (everyone exposed at one time, such as a shared meal): "The number of cases rises rapidly to a peak and falls gradually. The majority of cases occur within one incubation period."
- **Continuous common source** (exposure goes on, such as a contaminated well): the curve "rises gradually and might plateau".
- **Propagated** (person-to-person spread): "progressively taller peaks, each being one incubation period apart".

ILLUSTRATION / EXAMPLE

Figure 4.11 (left) shows 64 cases of gastroenteritis after a wedding feast; onsets run from 16:00 on day 1 to 14:00 on day 3. For this exercise, the suspected agent's incubation period is taken as 12 to 72 hours, averaging 36 hours. An 8-hour interval is 8/36 ≈ 0.22 of the average incubation period, inside the one-eighth to one-third range. The right panel shows 38 cases of a person-to-person infection in a village: clusters of 1, 3, 8 and 17 cases about 12 days apart, then 9 cases after control measures began.

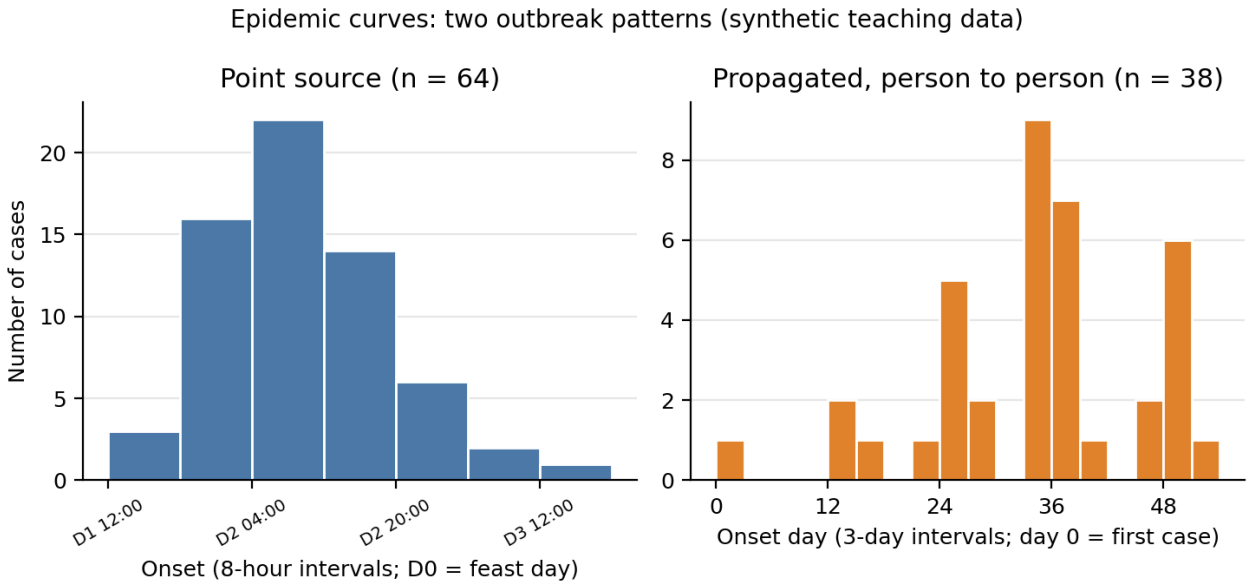


Figure 4.11. Epidemic curves. Left: point-source outbreak, 8-hour intervals (D0 = feast day). Right: propagated outbreak, 3-day intervals (synthetic teaching data).

WORKED CALCULATION

Estimate when the wedding guests were exposed, using the CDC's three rules (CDC, n.d.-b).

1. Latest possible exposure = earliest onset minus the minimum incubation period, because even the quickest case needed at least 12 hours: day 1 16:00 − 12 h = day 1 04:00.
2. Earliest possible exposure = last onset minus the maximum incubation period, because even the slowest case took at most 72 hours: day 3 14:00 − 72 h = day 0 14:00.
3. So exposure most likely fell between day 0 14:00 and day 1 04:00.
4. Cross-check from the peak: the tallest bar runs from day 2 04:00 to 12:00, with its midpoint at 08:00. Count back the average incubation period: day 2 08:00 − 36 h = day 0 20:00.
5. Result: the window from day 0 14:00 to day 1 04:00 contains the feast at 20:00 on day 0, and the peak estimate points straight at it. Your next step is to compare food-specific attack rates among the guests.

MUST-KNOW NOTES (MD-LEVEL)

- An epi curve is a histogram of onsets, so the bars adjoin; mark the interval width on the axis label.
- The back-calculation is only as good as the incubation period you assume for the agent. With the wrong agent in mind, the window will be wrong.
- A single late case far to the right may be a secondary (person-to-person) case, or a case with an unrelated exposure; it widens the window from rule 2.
- The shapes are typical, not guaranteed. Small outbreaks give ragged curves, and a point-source outbreak can seed secondary spread, giving a mixed pattern.

CHECKPOINT 4.4

1. A district's measles cases rose from 50 to 200 in five years, while a neighbouring district's rose from 500 to 1,000. On a semi-log graph, which line is steeper? Give each district's annual growth factor.
2. In a point-source outbreak the first case fell ill 30 hours after a community lunch and the last 90 hours after it. The agent's incubation period is taken as 24 to 96 hours. Using the first- and last-case rules, where does the exposure window fall relative to the lunch?
3. An epi curve rises steadily for three weeks and stays high. Which of the three patterns does this suggest, and what kind of source would you look for?

(Answers in the Appendix at the end of the book.)

4.5.3 Spot Map and Choropleth Map

DEFINITION

A **spot map** marks each case with a dot at the place where the person lived or was exposed (CDC, 2012a). A **choropleth map** (area map) shades each area, such as a ward, block or district, by a rate or proportion, with darker shades for higher burden.

SIMPLIFIED EXPLANATION

A spot map shows where cases are, and can reveal a cluster around a single water source or construction site. It shows only numerators. A dense cluster of dots may simply mean many people live there. The CDC's teaching text states that a spot map "does not show risk of disease", and that the risk "may not be particularly high if that area is densely populated" (CDC, 2012a). A choropleth map divides cases by the population of each area, so it shows risk. The CDC advises that the "intensity of shade or color reflects increasing disease burden" (CDC, 2012a).

ILLUSTRATION / EXAMPLE

Figure 4.12 shows 185 dengue cases in a town of nine wards (synthetic data). On the spot map, ward W5 in the centre has the most dots (54). On the choropleth map W5 is pale, and the dark wards are W9 and W3 on the eastern edge.

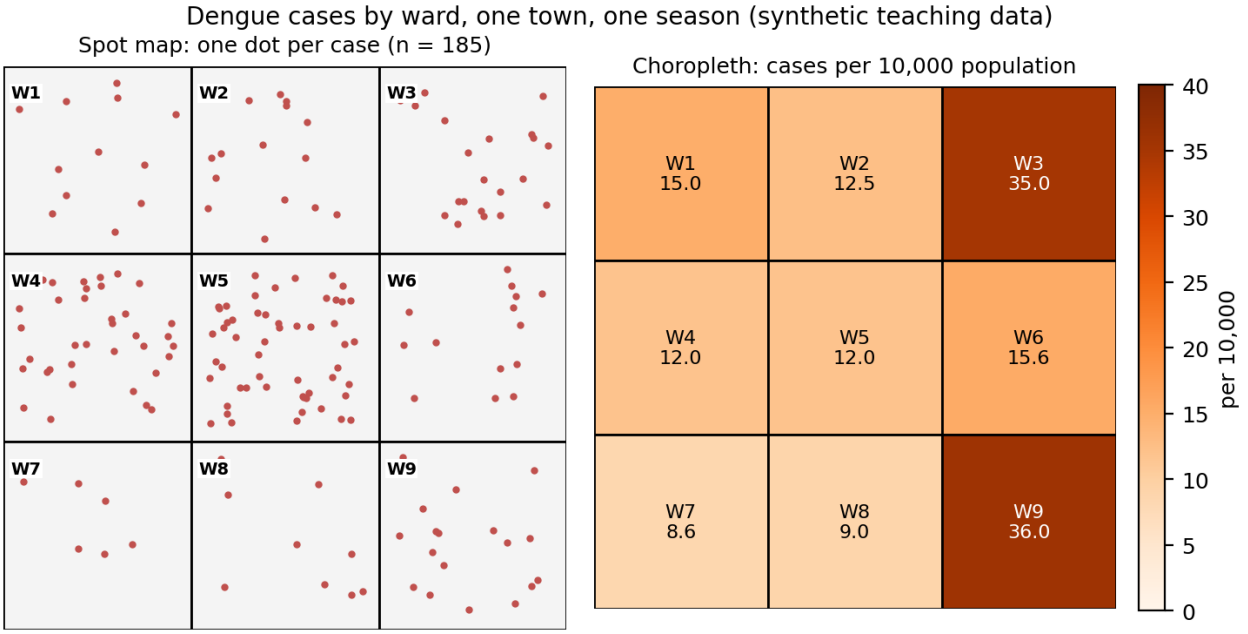


Figure 4.12. Spot map (left, one dot per case) and choropleth map (right, cases per 10,000 population) for the same 185 dengue cases in nine wards (synthetic teaching data).

WORKED CALCULATION

Rate per 10,000 = cases ÷ population × 10,000.

1. W5: 54 ÷ 45,000 × 10,000 = 12.0 per 10,000.
2. W9: 18 ÷ 5,000 × 10,000 = 36.0 per 10,000.
3. W3: 21 ÷ 6,000 × 10,000 = 35.0 per 10,000.

4. Whole town: 185 ÷ 132,000 × 10,000 = 14.0 per 10,000.
5. Result: W5 has three times as many cases as W9 (54 vs 18) but one-third of W9's risk (12.0 vs 36.0 per 10,000). Target breeding-site control by rate (W9, W3), and plan treatment capacity by count (W5).

MUST-KNOW NOTES (MD-LEVEL)

- The spot map answers "where are the cases?"; the choropleth map answers "where is the risk highest?". A report on an outbreak usually needs both.
- A choropleth rate in a small area rests on few cases, so it is unstable: one extra case in a ward of 5,000 moves its rate by 2 per 10,000.
- The choropleth map shades the whole area uniformly and hides variation within it. The picture also depends on where the boundaries are drawn.
- Plot cases by residence or by place of exposure, and say which. The two can point to different places.

4.5.4 Pictogram

DEFINITION

A **pictogram** represents a quantity by a row of identical small pictures, each standing for a fixed number of units, with a fraction of a picture for a remainder.

SIMPLIFIED EXPLANATION

Pictograms are built for lay audiences, such as a village health and sanitation meeting or a poster in a PHC, where a row of child figures is easier to take in than an axis. Their accuracy depends on one rule: change the number of symbols, never their size.

ILLUSTRATION / EXAMPLE

Figure 4.13 shows fully immunised children in four PHCs over one year (synthetic data), with one figure = 50 children. PHC A: 300 = 6 figures. PHC B: 175 = 3.5 figures. PHC C: 250 = 5 figures. PHC D: 125 = 2.5 figures.

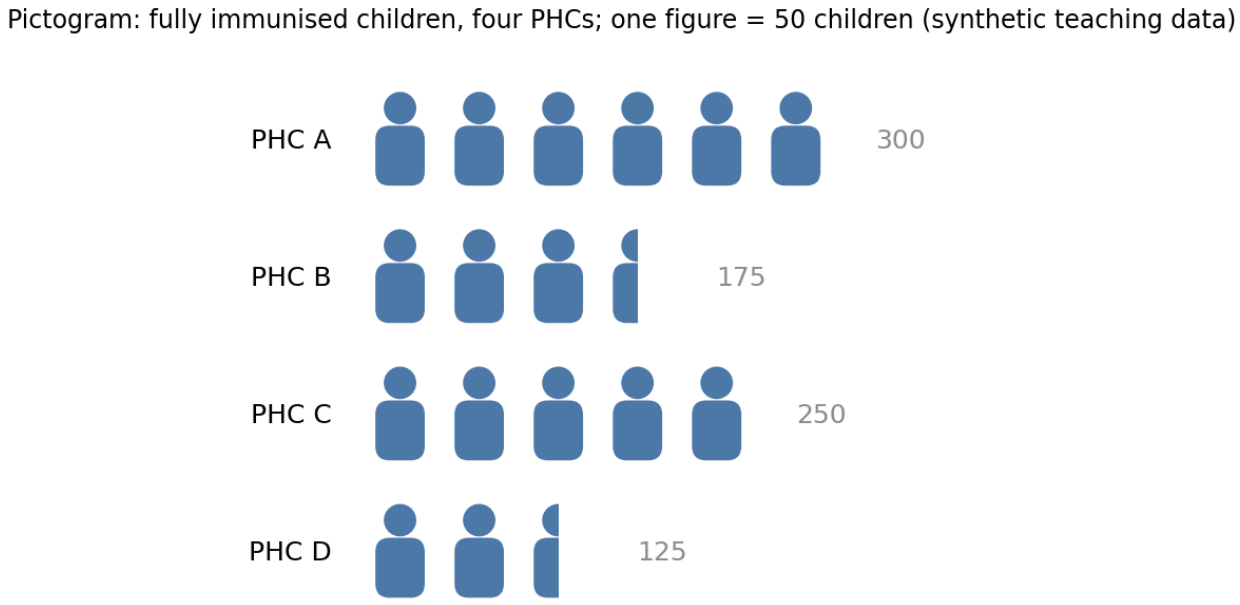


Figure 4.13. Pictogram: fully immunised children in four PHCs; one figure = 50 children (synthetic teaching data).

MUST-KNOW NOTES (MD-LEVEL)

- State the value of one symbol in the title or a key.
- A pictogram gives only approximate values; a fraction of a symbol is hard to judge by eye. Print the numbers beside the rows when exactness matters.
- Showing a larger quantity with a bigger symbol, instead of more symbols, misleads (§4.5.6).

4.5.5 Population Pyramid

DEFINITION

A **population pyramid** is a pair of back-to-back horizontal bar diagrams showing the count or percentage of a population in each age group, males on one side and females on the other (CDC, 2012b).

SIMPLIFIED EXPLANATION

The youngest age group is at the bottom and the oldest at the top. The outline records a population's history of births and deaths. Three shapes are named (Population Education, n.d.):

- **Expansive**: a broad base and a narrow top, a large share of young people. It reflects high fertility and lower life expectancy.
- **Constrictive**: narrower at the bottom than in the middle, with smaller shares in the youngest age groups, after fertility has fallen.
- **Stationary**: roughly equal shares across age groups, tapering only at the top, with low birth rates.

Pyramids can also show a disease or health characteristic by age and sex (CDC, 2012b).

DERIVATION

The **age dependency ratio** summarises a pyramid in one number. It is the ratio of dependents (people aged under 15 or over 64) to the working-age population (aged 15–64), expressed per 100 working-age people (World Bank, n.d.):

dependency ratio = (population aged 0–14 + population aged 65+) ÷ population aged 15–64 × 100

It splits into a young-age ratio (0–14 ÷ 15–64 × 100) and an old-age ratio (65+ ÷ 15–64 × 100). Because the numerator and the denominator come from the same population, you can compute it from counts or from percentages.

ILLUSTRATION / EXAMPLE

Figure 4.14 shows two hypothetical districts (synthetic data).

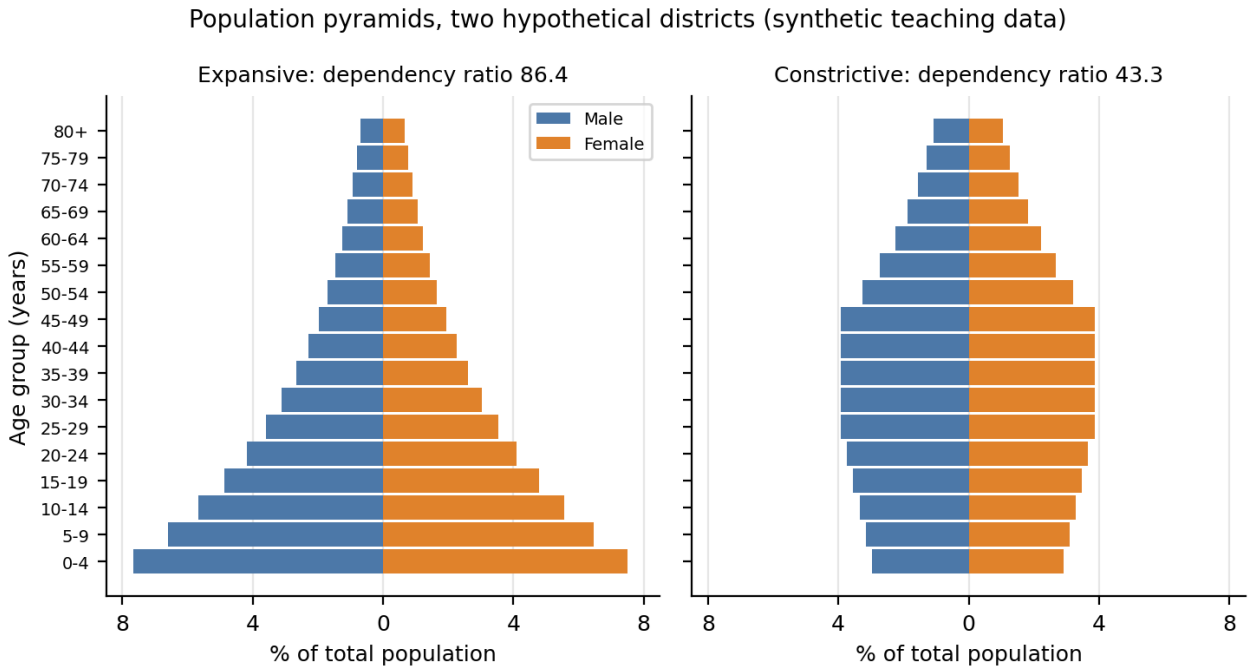


Figure 4.14. Population pyramids of two hypothetical districts, percentage of total population by age group and sex (synthetic teaching data).

WORKED CALCULATION

1. Expansive district: aged 0–14, 39.44%; aged 15–64, 53.67%; aged 65+, 6.92%.
2. Dependency ratio = (39.44 + 6.92) ÷ 53.67 × 100 = 46.36 ÷ 53.67 × 100 = 86.4 dependents per 100 working-age people. Young-age ratio 73.5; old-age ratio 12.9.
3. Constrictive district: aged 0–14, 18.74%; aged 15–64, 69.77%; aged 65+, 11.44%.
4. Dependency ratio = (18.74 + 11.44) ÷ 69.77 × 100 = 30.18 ÷ 69.77 × 100 = 43.3. Young-age ratio 26.9; old-age ratio 16.4.
5. Result: the expansive district carries twice the dependency burden, almost all of it children. Its health needs are child-health and maternal services; the constrictive district's old-age ratio is higher (16.4 vs 12.9), pointing towards non-communicable disease and geriatric care.

MUST-KNOW NOTES (MD-LEVEL)

- Plot percentages of the total population, not counts, when you compare pyramids of populations of different sizes.
- Use equal age-group widths. An open-ended top group (80+) is common; do not read its bar as a five-year band.
- A dent in one age group can come from a past event (a famine, an epidemic, a fall in births) or from migration: a district that sends young men away for work shows a pinched male side at ages 20–39.
- The dependency ratio is a demographic ratio; it counts ages, not who actually works or depends on others.

4.5.6 Misleading Graphs

DEFINITION

A **misleading graph** gives the reader a visual impression that the numbers do not support. The deception may be deliberate or accidental; the damage to the reader is the same.

SIMPLIFIED EXPLANATION

A graph encodes numbers as lengths, positions, angles or areas. It misleads when that encoding is broken. Standard texts give misleading graphs a section of their own (Bland, 2015, ch. 5). Failures to watch for:

- **Truncated baseline** on a bar diagram (recall the derivation in §4.2.1).
- **Area and volume symbols**: doubling a symbol's height and width multiplies its area by $2 \times 2 = 4$, so a two-fold difference looks four-fold.
- **3-D effects**: tilted pies and bars with depth distort angles and lengths. The near wedge of a tilted pie looks bigger than an equal far wedge.
- **Unequal intervals** on a time axis, or a histogram with unequal classes drawn by height instead of density (recall §4.3.1).
- **Two y-axes** on one chart: the apparent crossing point of the lines depends entirely on how each axis was scaled.
- **An arithmetic axis for proportional change**, or an unlabelled log axis (recall §4.5.1).

ILLUSTRATION / EXAMPLE

Two blocks report full immunisation coverage of 88% and 92% (synthetic data). Figure 4.15 draws the same two numbers twice.

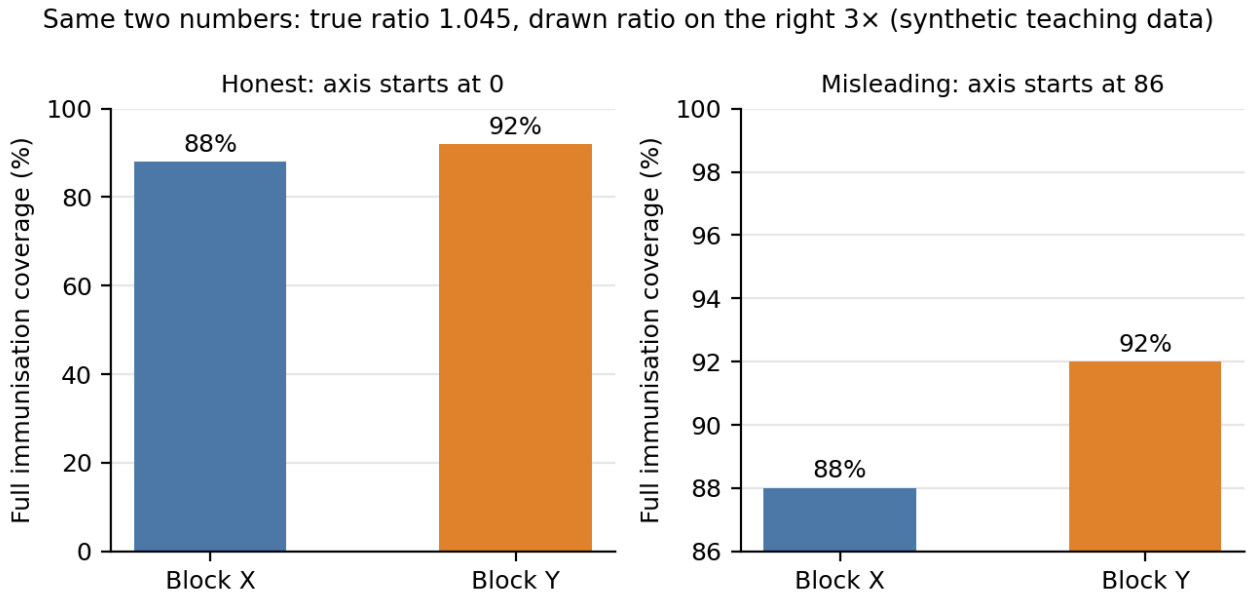


Figure 4.15. Full immunisation coverage in two blocks, drawn from zero (left) and from 86% (right) (synthetic teaching data).

WORKED CALCULATION

1. True ratio: 92 ÷ 88 = 1.045. Block Y's coverage is 4.5% higher (4 percentage points).
2. With the axis starting at 86, the drawn bars are 88 − 86 = 2 and 92 − 86 = 6 units long.
3. Drawn ratio: 6 ÷ 2 = 3.
4. Result: the truncated chart makes Block Y look three times as good as Block X, when it is only 1.045 times as high.

MUST-KNOW NOTES (MD-LEVEL)

- Bars and areas must start at zero. A line diagram, or a dot chart (one dot per category on a value axis), may use a narrower y-axis, because it encodes values by position, not length; mark the range clearly.
- One symbol, one value: change the count of symbols, not their size.
- Avoid 3-D effects on 2-D data.
- Before you trust any chart, read the axis labels, the scale and the baseline before you look at the shape.

CHECKPOINT 4.5

1. Ward P had 40 dengue cases in a population of 50,000; ward Q had 15 cases in 6,000. Which has more dots on a spot map, and which is darker on a choropleth map? Give both rates per 10,000.
2. A district's population is 34% aged 0–14, 60% aged 15–64 and 6% aged 65+. Compute its age dependency ratio. Is its pyramid more likely to be expansive or constrictive?
3. A poster shows PHC coverage of 60% by a syringe 3 cm tall and 1 cm wide, and 90% by a syringe 4.5 cm tall and 1.5 cm wide. What ratio does the poster's area suggest, and what is the true ratio?

(Answers in the Appendix at the end of the book.)

Foundations of Probability

Chapters 1–4 described data already in hand. Real questions in community medicine go beyond it: will this child respond to treatment, and how likely is it that a positive screening result is a false alarm? **Probability** is the language for that uncertainty, and every confidence interval and hypothesis test later in the book is written in it. This chapter builds probability from its axioms (§5.1), derives the rules for combining probabilities (§5.2–§5.4), turns them into Bayes' theorem and the arithmetic of screening tests (§5.5), and ends with random variables and the named distributions that inference uses (§5.6–§5.7).

5.1 From Counting to Probability

DEFINITION

An **experiment** is any process whose result is not known in advance: rolling a die, testing a child for anaemia, recording next month's case count. Each possible result is an **outcome**. The set of all outcomes is the **sample space**, written S. An **event** is any collection of outcomes, that is, any subset of S. **Probability** is a number P(A) attached to every event A that obeys three **axioms** (rules accepted without proof, from which everything else is derived):

1. **Non-negativity:** P(A) ≥ 0 for every event A.
2. **Total probability one:** P(S) = 1. Something in the sample space must happen.
3. **Additivity:** if A and B cannot happen together (they share no outcomes), then P(A or B) = P(A) + P(B).

SIMPLIFIED EXPLANATION

The sample space is the list of everything that could happen. An event is a question you can ask of the result ("is the die even?"), answered yes by some outcomes and no by others. The axioms say only three things: probabilities are never negative, the whole sample space has probability 1, and probabilities of non-overlapping events add. Every rule in this chapter (the complement rule, the addition rule, Bayes' theorem) is a consequence of these three lines. The axioms do not say where the numbers come from. Three approaches supply them: counting equally likely outcomes (§5.1.1), observing long-run frequencies (§5.1.2) and personal judgement (§5.1.3). All three obey the same axioms, so the same rules apply to all of them.

ILLUSTRATION / EXAMPLE

Rolling one die: S = {1, 2, 3, 4, 5, 6}. The event "even" is A = {2, 4, 6}; the event "4" is {4}. Testing one child for anaemia: S = {anaemic, not anaemic}. Counting malaria cases in a village next month: S = {0, 1, 2, 3, ...}, a sample space with no upper limit.

MUST-KNOW NOTES (MD-LEVEL)

- Two consequences follow at once from the axioms. Every probability lies between 0 and 1 (the upper bound is derived in §5.2.3). The probability of an impossible event is 0.
- Write the sample space down before calculating. Many errors in probability problems come from a sample space that is vague or that silently omits outcomes.
- Events are sets, so the words "or", "and" and "not" have set meanings: "A or B" means at least one of them happens; "A and B" means both happen; "not A" (the **complement**) means every outcome outside A.

5.1.1 Classical Probability

DEFINITION

Classical (a priori) probability assumes all outcomes in a finite sample space are equally likely, and defines P(A) = (number of outcomes in A) ÷ (total number of outcomes).

SIMPLIFIED EXPLANATION

This is probability by reasoning about the structure of a situation, before any data exist. A fair coin has two symmetric sides, so heads has probability 1/2 before anyone flips it. The formula satisfies the three axioms automatically: counts are never negative, the whole sample space counts n/n = 1, and counts of non-overlapping sets add.

ILLUSTRATION / EXAMPLE

A fair die has 6 equally likely outcomes. P(even) = 3/6 = 0.5. P(at least 5) = 2/6 = 0.333. No die was rolled.

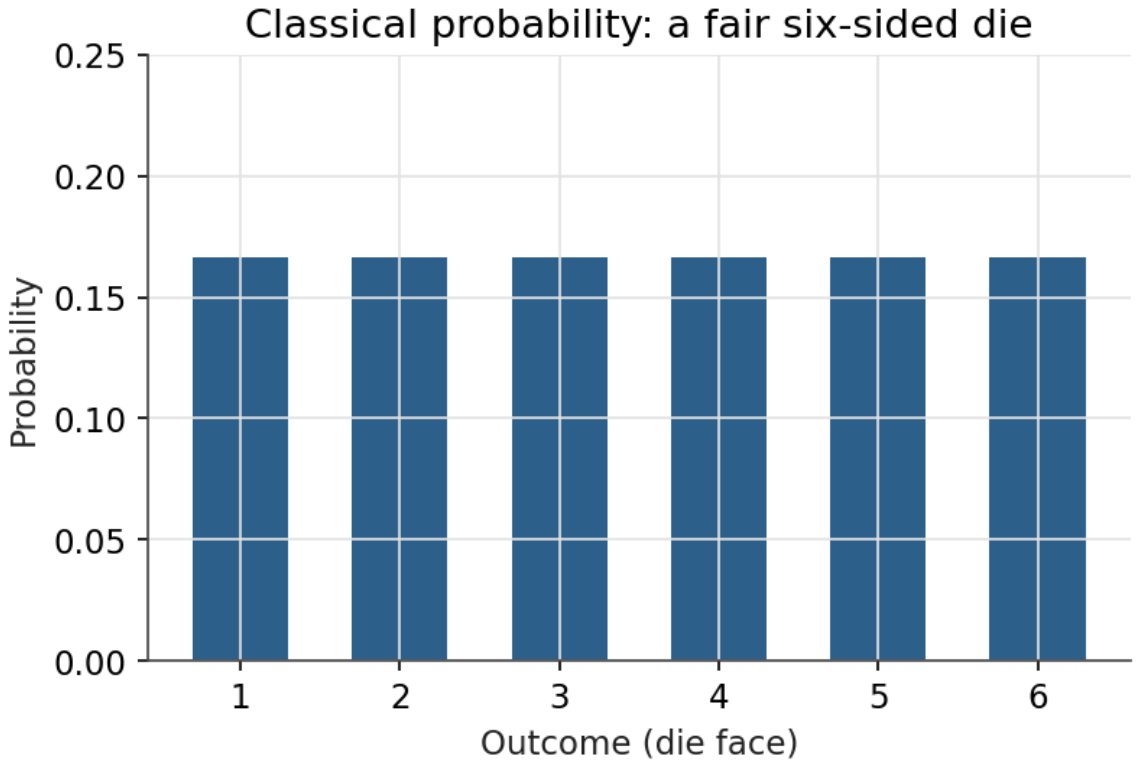


Figure 5.1. Classical probability: the six equally likely outcomes of a fair die.

MUST-KNOW NOTES (MD-LEVEL)

- Formula: P(A) = (favourable outcomes) ÷ (total equally likely outcomes).
- The equal-likelihood assumption is the whole method. It fails for a loaded die, and it has no meaning for "will this patient develop a rash?", where the outcomes are not symmetric. Then empirical probability (§5.1.2) is needed.
- Counting outcomes quickly becomes hard when there are many; the counting rules in §5.7.1.2 handle that.

5.1.2 Empirical Probability

DEFINITION

Empirical (relative frequency) probability estimates P(A) by the proportion of times A occurred in a large number of repetitions: P(A) ≈ (number of times A occurred) ÷ (number of trials).

SIMPLIFIED EXPLANATION

You cannot reason your way to the probability that a patient develops a drug rash, as you can for a fair coin. So you observe: out of many patients given the drug, what fraction developed the rash? The observed fraction is your estimate of the true probability. It is an estimate, not the probability itself. A second group of 500 patients would give a slightly different fraction, just as a second sample gives a different statistic for the same parameter (recall §1.3).

ILLUSTRATION / EXAMPLE

Of 500 patients who received an antibiotic, 15 developed a rash. Empirical P(rash) = 15/500 = 0.03. At an anganwadi, 36 of 400 children screened have pallor: empirical P(pallor) = 36/400 = 0.09.

MUST-KNOW NOTES (MD-LEVEL)

- The estimate settles down as trials accumulate. The precise statement is the **(weak) law of large numbers**: as the number of independent repetitions grows, the probability that the observed proportion differs from the true probability by more than any small amount you choose shrinks towards zero. The law says nothing about the next single trial. After ten boys, the next birth is not "due" to be a girl.
- Prevalence, treatment-response rates, sensitivity and specificity are all empirical probabilities, estimated from data.
- How far an estimate from one sample is likely to be from the truth is the subject of sampling distributions and standard errors (the spread of a statistic over repeated samples, §6.5–§6.7).

5.1.3 Subjective Probability

DEFINITION

Subjective probability is a person's degree of belief that an event will occur, based on judgement and the information available, rather than on counting or repeated observation.

SIMPLIFIED EXPLANATION

Some events are one-off and cannot be repeated or reasoned out from symmetry. An experienced clinician who says "this patient has about a 70% chance of responding" is stating a subjective probability. It is still a probability: if it is coherent, it obeys the same three axioms, so it can be combined with the rules that follow.

ILLUSTRATION / EXAMPLE

Before any case data arrive, a district epidemiologist judges that a new outbreak has "about a 20% chance" of becoming a large epidemic. The number guides immediate planning; it will be revised as cases are counted (the revision rule is Bayes' theorem, §5.5).

MUST-KNOW NOTES (MD-LEVEL)

- Two equally informed, reasonable people can hold different subjective probabilities. That is the key difference from classical and empirical probability.
- Subjective probability is the starting point of Bayesian statistics (inference that updates a prior belief with data, §15.8). A clinician's pre-test probability in §5.5 is often subjective.

CHECKPOINT 5.1

1. A clinician says "I think this patient has about a 70% chance of responding to treatment", based on years of experience with similar cases. Which approach to probability is this, and why can it not be classical probability?
2. Write down the sample space for "number of ANC visits a woman makes, capped at 4 or more". Which of the three axioms tells you the probabilities of its outcomes must add to 1?
3. In a village survey, 54 of 600 adults report a diagnosis of hypertension. What is the empirical probability of a reported diagnosis? Would a second survey of 600 adults give exactly the same number? Which law describes what happens to the estimate as the survey grows?

(Answers in the Appendix at the end of the book.)

5.2 Basic Probability Rules

Three rules combine probabilities: the **addition rule** for "A or B" (§5.2.1), the **multiplication rule** for "A and B" (§5.2.2) and the **complement rule** for "not A" (§5.2.3). Each follows from the axioms in §5.1. They are not a complete toolkit on their own: combining events that overlap or influence each other needs extra information (a joint probability, a conditional probability, or an assumption of independence), and a fourth rule, the law of total probability, needs conditional probability and is derived in §5.3.

5.2.1 Addition Rule

DEFINITION

The **addition rule** gives the probability that at least one of two events occurs: P(A or B) = P(A) + P(B) − P(A and B).

SIMPLIFIED EXPLANATION

Adding P(A) and P(B) counts the overlap, the outcomes where both happen, twice. Subtracting P(A and B) once removes the double count.

DERIVATION

Split "A or B" into three pieces that share no outcomes: A only, B only, and both. By additivity (axiom 3):

P(A or B) = P(A only) + P(B only) + P(A and B).

A itself splits into "A only" and "both", so P(A only) = P(A) − P(A and B). Likewise P(B only) = P(B) − P(A and B). Substituting:

P(A or B) = P(A) + P(B) − 2P(A and B) + P(A and B) = P(A) + P(B) − P(A and B).

ILLUSTRATION / EXAMPLE

Among adults attending a PHC NCD clinic, P(diabetes) = 0.2, P(hypertension) = 0.3 and P(both) = 0.1. Then P(at least one) = 0.2 + 0.3 − 0.1 = **0.4**, and P(neither) = 1 − 0.4 = 0.6. Without the overlap you would have claimed 0.5. Notice that the two single probabilities were not enough: the answer needed P(both).

MUST-KNOW NOTES (MD-LEVEL)

- When A and B are **mutually exclusive** (they cannot both happen, so P(A and B) = 0), the rule reduces to axiom 3: P(A or B) = P(A) + P(B). A person's ABO blood group being A and being B are mutually exclusive.
- "Or" in probability is inclusive: "A or B" includes the case where both happen.
- If P(A and B) is unknown, P(A or B) cannot be computed from P(A) and P(B) alone. You need the joint probability, or an assumption such as independence (§5.4), which supplies P(A and B) = P(A) × P(B).

5.2.2 Multiplication Rule

DEFINITION

The **multiplication rule for independent events** states that if A and B are **independent** (the occurrence of one does not change the probability of the other; made exact in §5.4), then P(A and B) = P(A) × P(B).

SIMPLIFIED EXPLANATION

If a household's water source tells you nothing about its size, then the share of households that have piped water and are large is the share with piped water times the share that is large. When one event does change the chance of the other, this simple product is wrong, and the general version (§5.3) replaces P(B) with the probability of B given that A has happened.

ILLUSTRATION / EXAMPLE

In a CBHI household survey, P(piped water) = 0.6 and P(more than 5 members) = 0.3. If water source is independent of household size, P(piped water and more than 5 members) = 0.6 × 0.3 = **0.18**. Whether that independence holds is a question to check in the data (a comparison of 0.18 with the observed joint proportion), not a fact to assume.

MUST-KNOW NOTES (MD-LEVEL)

- For independent events only: P(A and B) = P(A) × P(B). It extends to any number of independent events: $P(A_1 \text{ and } A_2 \text{ and } \ldots \text{ and } A_k) = P(A_1) \times P(A_2) \times \ldots \times P(A_k)$.
- The general form, valid with or without independence, is P(A and B) = P(A) × P(B|A), where P(B|A) is the probability of B given that A has happened, derived in §5.3 once conditional probability is defined.
- Using the product for events that are not independent is a frequent error. Two children from the same household are not independent for diarrhoea: they share water, food and hygiene.

5.2.3 Complement Rule

DEFINITION

The **complement rule** states that the probability that A does not occur is P(not A) = 1 − P(A).

SIMPLIFIED EXPLANATION

An event and its complement share no outcomes and together cover the whole sample space. So their probabilities add to 1. The rule's real power is as a shortcut: when "at least one" is hard to calculate directly, calculate "none" and subtract from 1.

DERIVATION

A and "not A" share no outcomes, and "A or not A" is the whole sample space S. By axioms 3 and 2:

P(A) + P(not A) = P(S) = 1, so P(not A) = 1 − P(A).

Because P(not A) ≥ 0 (axiom 1), it also follows that P(A) ≤ 1. This is the upper bound promised in §5.1.

ILLUSTRATION / EXAMPLE

If P(a screening test is positive) = 0.15, then P(negative) = 1 − 0.15 = 0.85.

WORKED CALCULATION: "AT LEAST ONE"

A child has ten independent laboratory tests, each with probability 0.1 of an abnormal result.

1. P(one test normal) = 1 − 0.1 = 0.9.
2. P(all ten normal) = 0.9^{10} = 0.3487 (multiplication rule for independent events).
3. P(at least one abnormal) = 1 − 0.3487 = **0.651**.

Doing it directly would mean adding P(exactly 1), P(exactly 2), … , P(exactly 10); the complement needs one line.

MUST-KNOW NOTES (MD-LEVEL)

- P(not A) = 1 − P(A); equivalently P(A) + P(not A) = 1.
- "At least one" = 1 − "none". The pattern returns in the binomial distribution (§5.7.1.2) and in the family-wise error rate for multiple tests (§9.5.3).

CHECKPOINT 5.2

1. A child has ten independent laboratory tests, each with a 5% chance of an abnormal result. What is the probability that at least one is abnormal? Show why the complement rule is the easier route.
2. Two events are mutually exclusive. Explain, without using the word "exclusive", why the addition rule simplifies for them.

3. P(A) = 0.2 and P(B) = 0.3, and A and B are independent. Find P(A or B).

(Answers in the Appendix at the end of the book.)

5.3 Conditional Probability

DEFINITION

The **conditional probability** of A given B, written P(A|B), is the probability that A occurs when B is known to have occurred:

P(A|B) = P(A and B) ÷ P(B), defined only when P(B) > 0.

SIMPLIFIED EXPLANATION

Learning that B happened shrinks the sample space from S to B. P(A|B) is the share of that smaller space in which A also happens. That is why the formula divides by P(B): it rescales the probabilities inside B so that they add to 1 again.

DERIVATION

Two results follow in one line each, and both are used throughout §5.5.

General multiplication rule. Multiply both sides of the definition by P(B): P(A and B) = P(B) × P(A|B). By symmetry, P(A and B) = P(A) × P(B|A).

Law of total probability. Let B be any event. The event A splits into two pieces that share no outcomes: "A and B" and "A and not B". By additivity and then the general multiplication rule:

P(A) = P(A and B) + P(A and not B) = P(A|B)P(B) + P(A|not B)P(not B).

The law says: to get an overall probability, average the conditional probabilities over the groups, weighting each group by its size. It extends to any set of groups B_1, B_2, … , B_k that do not overlap and together cover S.

ILLUSTRATION / EXAMPLE

In the NCD clinic of §5.2.1, P(diabetes) = 0.2, P(hypertension) = 0.3 and P(both) = 0.1. Among patients known to have diabetes, the probability of hypertension is P(H|D) = 0.1 ÷ 0.2 = **0.5**. Among patients known to be hypertensive, P(D|H) = 0.1 ÷ 0.3 = **0.333**. The two conditional probabilities differ because they condition on different groups, and neither equals the overall P(H) = 0.3.

MUST-KNOW NOTES (MD-LEVEL)

- P(A|B) and P(B|A) are different numbers. Confusing them is the root error behind misread screening results (§5.5).
- General multiplication rule: P(A and B) = P(A) × P(B|A), valid for any events.
- Law of total probability: P(A) = P(A|B)P(B) + P(A|not B)P(not B). It supplies the denominator of Bayes' theorem (§5.5).

- Conditioning is what a diagnostic test does: it moves you from P(disease) to P(disease | result).

5.4 Independence of Events

DEFINITION

Events A and B are **independent** if and only if P(A|B) = P(A). Equivalently, P(A and B) = P(A) × P(B).

SIMPLIFIED EXPLANATION

A and B are independent when learning that B happened tells you nothing new about A. The two forms of the definition are the same statement. Substitute P(A|B) = P(A) into the general multiplication rule of §5.3 and you get the product form.

ILLUSTRATION / EXAMPLE

Two flips of a fair coin are independent: a head on the first does not change the chance of a head on the second. In the NCD clinic, P(D) × P(H) = 0.2 × 0.3 = 0.06, but P(D and H) = 0.1. The product does not match, so diabetes and hypertension are **not** independent there: P(H|D) = 0.5 is higher than P(H) = 0.3. Whether the events "systolic BP ≥ 140 mmHg" and "salt intake above 5 g/day" are independent in a given population is settled the same way: compare P(BP ≥ 140 | high salt) with P(BP ≥ 140) in the data.

MUST-KNOW NOTES (MD-LEVEL)

- Test: independent if and only if P(A|B) = P(A), equivalently P(A and B) = P(A) × P(B).
- Independence and mutual exclusivity are different, and they are often confused. Mutually exclusive events cannot both happen (P(A and B) = 0). Two mutually exclusive events with non-zero probabilities are therefore always dependent: learning that one happened tells you the other did not.
- Many methods later in the book assume that observations are independent of one another. Treating repeated measurements on the same patient, or children in the same household, as independent is a frequent and serious error in applied research; clustered data need methods that allow for the within-cluster correlation (§15.2).

CHECKPOINT 5.3

1. Explain the difference between "mutually exclusive" and "independent" using a concrete pair of events.
2. How does conditional probability define independence?
3. In an under-five clinic, P(underweight) = 0.25, P(recent diarrhoea) = 0.20 and P(both) = 0.08. Find P(underweight | recent diarrhoea). Are the two events independent? Show the check.

(Answers in the Appendix at the end of the book.)

5.5 Bayes' Theorem

DEFINITION

Bayes' theorem reverses the direction of a conditional probability:

$P(A|B) = P(B|A) \times P(A) \div P(B)$,

where the denominator, by the law of total probability, is $P(B) = P(B|A)P(A) + P(B|\text{not } A)P(\text{not } A)$.

SIMPLIFIED EXPLANATION

A diagnostic test is validated in people whose true status is known, so it gives you P(positive | disease). The clinician facing a patient with a positive result wants the reverse, P(disease | positive). Bayes' theorem converts one into the other. It needs one extra ingredient: the probability of disease before the test, called the **prior** or **pre-test probability**. In a screening programme that is the prevalence. The answer, P(disease | positive), is the **posterior** or **post-test probability**.

DERIVATION

The general multiplication rule (§5.3) gives P(A and B) two ways:

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

Divide both sides by P(B):

$P(A|B) = P(B|A) \times P(A) \div P(B)$.

Expand P(B) by the law of total probability. With A = disease (D) and B = positive test (T+):

$P(D|T+) = P(T+|D)P(D) \div [P(T+|D)P(D) + P(T+|\text{not } D)P(\text{not } D)]$.

ILLUSTRATION / EXAMPLE

A screening test has $P(T+|D) = 0.95$ and $P(T+|\text{not } D) = 0.10$. In a community where $P(D) = 0.01$:

1. Numerator: $0.95 \times 0.01 = 0.0095$.
2. Denominator: $0.0095 + 0.10 \times 0.99 = 0.0095 + 0.099 = 0.1085$.
3. $P(D|T+) = 0.0095 \div 0.1085 =$ **0.088**.

Fewer than 9 in 100 people who test positive have the disease. The test is not poor. The healthy pool is 99 times larger than the diseased pool, so its 10% false-positive rate produces far more positives (0.099) than the diseased group's 95% (0.0095). §5.5.2 repeats this calculation as a table of 10,000 people.

MUST-KNOW NOTES (MD-LEVEL)

- P(disease | positive) = P(positive | disease) × P(disease) ÷ P(positive), with P(positive) expanded over the diseased and the healthy.
- The prior dominates. The same test gives very different post-test probabilities in mass community screening and in a referral clinic (§5.5.2).
- Sensitivity and specificity (§5.5.1) are the inputs, predictive values (§5.5.2) are the outputs, and likelihood ratios (§5.5.3) are the same theorem written in odds.

5.5.1 Sensitivity and Specificity

DEFINITION

Sensitivity (Se) is the probability that the test is positive in a person who has the disease: Se = P(T+| D). **Specificity** (Sp) is the probability that the test is negative in a person who does not: Sp = P(T−| not D). Both are judged against a **reference standard** (the best available method for deciding true disease status).

SIMPLIFIED EXPLANATION

The results of a test against the reference standard fill a 2×2 table:

	Disease present	Disease absent
Test positive	a (true positives, TP)	b (false positives, FP)
Test negative	c (false negatives, FN)	d (true negatives, TN)

Sensitivity reads down the first column: Se = a ÷ (a + c). Specificity reads down the second: Sp = d ÷ (b + d). Each is a conditional probability given true status, which is exactly what Bayes' theorem takes as input.

ILLUSTRATION / EXAMPLE

A new TB screening tool is applied to a consecutive series of 400 adults with cough at a district hospital, and all are also tested with the reference standard. Of 100 with TB, 95 test positive: Se = 95/100 = **95%**. Of 300 without TB, 270 test negative: Sp = 270/300 = **90%**. The 5 missed cases are false negatives; the 30 TB-free people who tested positive are false positives.

MUST-KNOW NOTES (MD-LEVEL)

- Se = TP ÷ (TP + FN) measures how well the test catches disease. A highly sensitive test suits initial screening, where a missed case is costly.
- Sp = TN ÷ (TN + FP) measures how well it clears the healthy. A highly specific test suits confirmation, where a false positive leads to harmful further intervention.
- Se and Sp do not depend on prevalence through the arithmetic (each is computed within one column). They are still not fixed properties of the test. They vary with disease severity, case mix and setting ("It is now well established that sensitivity and specificity are not fixed test properties", Bossuyt et al., 2015), and a systematic review found consistent effects of disease prevalence and severity on accuracy estimates (Whiting et al., 2013).
- Design matters. Non-consecutive inclusion inflates accuracy estimates, and studies that compared severe cases with healthy controls ("two-gate" or case-control sampling) gave the highest estimates of all, though imprecisely (Rutjes et al., 2006); a later systematic review found consistent evidence of an effect of case-control design (Whiting et al., 2013). The preferred design enrols one consecutive or random series of people in whom the test would be used, and classifies all of them by the reference standard; the STARD reporting guideline asks authors to state "whether participants formed a consecutive, random, or convenience series" (Bossuyt et al., 2015).

- The exam mnemonics **SnNout** (a highly **s**e**n**sitive test, when **n**egative, rules **out**) and **SpPin** (a highly **sp**ecific test, when **p**ositive, rules **in**) are shortcuts, and they carry a published "note of caution" (Pewsner et al., 2004). The arithmetic shows why. A test with Se 95% but Sp only 10% is negative in 5% of the diseased and in only 10% of the healthy, so a negative result is merely half as likely in the diseased: it barely rules anything out despite the high sensitivity (in the language of §5.5.3, LR− = 0.05 ÷ 0.10 = 0.5). Judge rule-in and rule-out with the likelihood ratios of §5.5.3.
- Moving a test's cut-off trades sensitivity against specificity. The full trade-off curve is the ROC curve (the plot of sensitivity against 1 − specificity over all cut-offs, §15.5.1).

5.5.2 Positive and Negative Predictive Value

DEFINITION

The **positive predictive value** is PPV = P(D|T+), the probability that a person who tests positive has the disease. The **negative predictive value** is NPV = P(not D|T−), the probability that a person who tests negative is disease-free.

SIMPLIFIED EXPLANATION

Sensitivity and specificity answer "given true status, how does the test behave?". A clinician holds a result and wants the reverse. PPV and NPV read across the rows of the 2×2 table: PPV = a ÷ (a + b), NPV = d ÷ (c + d). They are the outputs of Bayes' theorem, and because the rows mix diseased and healthy people, they depend on how many of each there are, that is, on prevalence.

WORKED CALCULATION: THE 2×2 TABLE FOR 10,000 PEOPLE

Se = 95%, Sp = 90%, prevalence 1% (the example of §5.5).

1. Diseased: 1% of 10,000 = 100. True positives = 0.95 × 100 = 95; false negatives = 5.
2. Healthy: 9,900. True negatives = 0.90 × 9,900 = 8,910; false positives = 990.

	Disease present	Disease absent	Total
Test positive	95	990	1,085
Test negative	5	8,910	8,915
Total	100	9,900	10,000

1. PPV = 95 ÷ 1,085 = **8.8%**. The row total 1,085 ÷ 10,000 = 0.1085 is P(T+) from the law of total probability.
2. NPV = 8,910 ÷ 8,915 = **99.94%**.

Of every 1,085 positives, 990 (91%) are false alarms. The table and the formula of §5.5 give the same answer because the table is the formula written out in people.

ILLUSTRATION / EXAMPLE: PPV AGAINST PREVALENCE

The same test (Se 95%, Sp 90%) used in different settings:

Setting (illustrative)	Prevalence	PPV	NPV
Mass community screening	1%	8.8%	99.94%
Screening a high-risk group	5%	33.3%	99.71%
PHC, people with symptoms	10%	51.4%	99.39%
Referral clinic	30%	80.3%	97.67%
Specialist clinic	50%	90.5%	94.74%

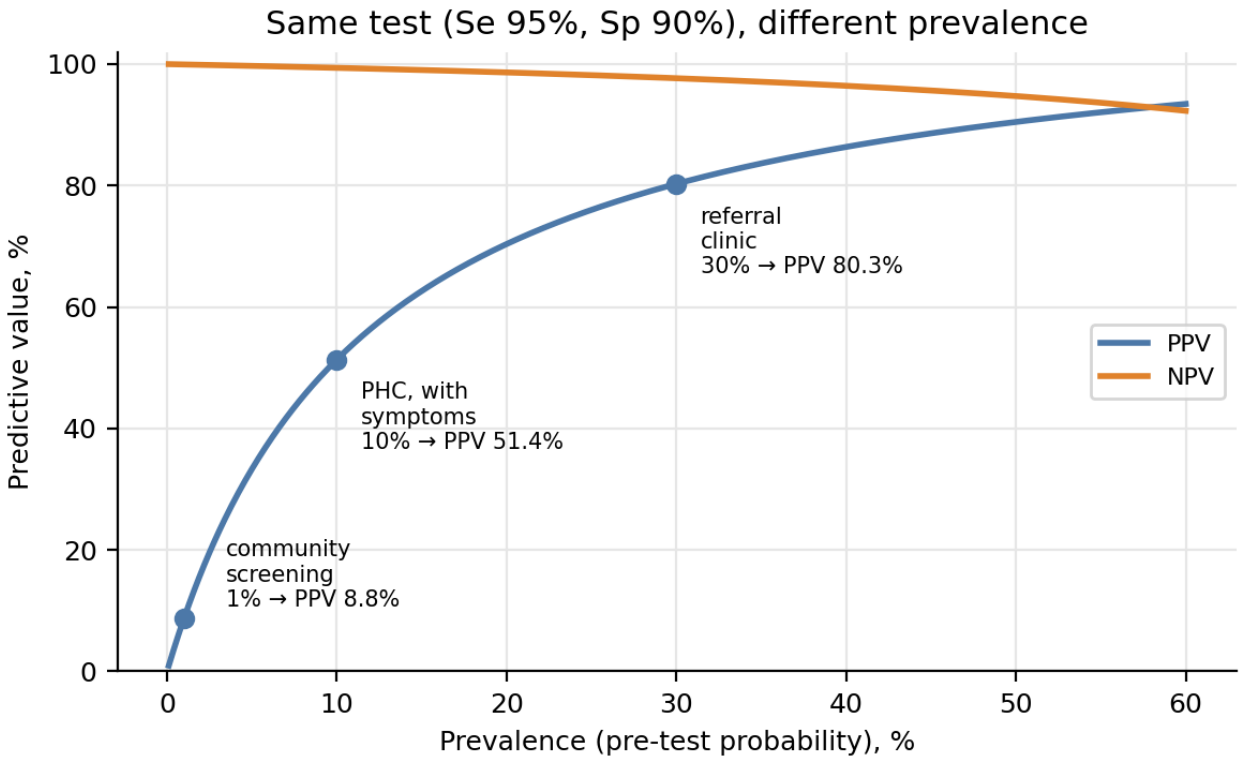


Figure 5.2. The same test gives a PPV of 8.8% at 1% prevalence, 51.4% at 10% and 80.3% at 30%. NPV falls slowly as prevalence rises.

WORKED CALCULATION: TWO TESTS IN SERIES AND IN PARALLEL

Test A (Se 0.95, Sp 0.90) and test B (Se 0.80, Sp 0.95), assuming the two tests err independently of each other within the diseased and within the healthy (conditional independence).

Series (call positive only if both are positive; in practice, run B only on A-positives):

1. Se(series) = Se_A × Se_B = 0.95 × 0.80 = **0.76**.
2. Sp(series) = 1 − (1 − Sp_A)(1 − Sp_B) = 1 − 0.10 × 0.05 = **0.995**.

Parallel (call positive if either is positive):

1. Se(parallel) = 1 − (1 − Se_A)(1 − Se_B) = 1 − 0.05 × 0.20 = **0.99**.
2. Sp(parallel) = Sp_A × Sp_B = 0.90 × 0.95 = **0.855**.

At 1% prevalence, series testing gives 76 true and 49.5 false positives per 10,000 (PPV 60.6%, up from 8.8% for A alone), but misses 24 of every 100 cases. Parallel testing catches 99 of 100 cases (NPV 99.99%) at the price of 1,435.5 false positives (PPV 6.5%).

MUST-KNOW NOTES (MD-LEVEL)

- PPV = TP ÷ (TP + FP); NPV = TN ÷ (TN + FN). Both depend on prevalence; Se and Sp do not, arithmetically.
- As prevalence rises with Se and Sp fixed, PPV rises and NPV falls. Reason it through: more of the people tested are diseased, so more positives are true and more negatives are missed cases.
- A PPV reported from a hospital study does not transfer to community screening. Recompute it for your own prevalence with Bayes' theorem or a 2×2 table.
- Series testing raises specificity and PPV and lowers sensitivity; parallel testing raises sensitivity and NPV and lowers specificity. The formulas assume conditional independence. Two tests that detect the same feature tend to err on the same people, and then the gains are smaller than the formulas predict.

5.5.3 Likelihood Ratios

DEFINITION

The **odds** of an event are its probability divided by the probability that it does not happen: odds = p ÷ (1 − p) (recall §1.5). Conversely, p = odds ÷ (1 + odds). A **likelihood ratio** compares how probable a test result is in people with the disease and in people without it:

$LR+ = P(T+|D) \div P(T+|\text{not } D) = Se \div (1 - Sp)$

$LR- = P(T-|D) \div P(T-|\text{not } D) = (1 - Se) \div Sp$

SIMPLIFIED EXPLANATION

Odds and probability describe the same uncertainty on different scales. A probability of 0.2 is odds of 0.2 ÷ 0.8 = 0.25 ("1 to 4"); a probability of 0.5 is odds of 1. Odds are not probabilities: odds can exceed 1, and the two scales agree closely only when p is small. The reason to switch scales is that on the odds scale Bayes' theorem becomes a single multiplication: post-test odds = pre-test odds × likelihood ratio. Because the LR is built only from Se and Sp, one number summarises the test and can be applied to any patient's pre-test probability.

DERIVATION

Write Bayes' theorem for D and for "not D", given a positive result. Both share the denominator P(T+):

$P(D|T+) = P(T+|D)P(D) \div P(T+)$

$P(\text{not } D|T+) = P(T+|\text{not } D)P(\text{not } D) \div P(T+)$

Divide the first line by the second. P(T+) cancels:

$P(D|T+) \div P(\text{not } D|T+) = [P(T+|D) \div P(T+|\text{not } D)] \times [P(D) \div P(\text{not } D)]$

The left side is the post-test odds, the first bracket is LR+, and the second is the pre-test odds. The same steps for a negative result give post-test odds = pre-test odds × LR−.

WORKED CALCULATION: CLOSING THE LOOP WITH §5.5.2

Se 0.95, Sp 0.90, pre-test probability 1%.

1. LR+ = 0.95 ÷ 0.10 = **9.5**. LR− = 0.05 ÷ 0.90 = **0.056** (a negative result divides the odds by 18).
2. Pre-test odds = 0.01 ÷ 0.99 = 0.0101.
3. After a positive result: post-test odds = 0.0101 × 9.5 = 0.0960; post-test probability = 0.0960 ÷ 1.0960 = **0.088**. This is the PPV of the 2×2 table.
4. After a negative result: post-test odds = 0.0101 × 0.056 = 0.00056; probability of disease = 0.00056, so NPV = **99.94%**, again as in the table.
5. The same positive result in a referral clinic with pre-test probability 30%: odds 0.30 ÷ 0.70 = 0.4286; × 9.5 = 4.071; probability = 4.071 ÷ 5.071 = **0.803**.

The LR did not change between settings; only the pre-test odds did.

ILLUSTRATION / EXAMPLE: YOUDEN'S INDEX

Youden's index J = Se + Sp − 1 summarises a test in one number (Youden, 1950). J = 0 for a test no better than chance (then Se = 1 − Sp and LR+ = 1); J = 1 for a perfect test. For the test above J = 0.95 + 0.90 − 1 = **0.85**; for test B of §5.5.2, J = 0.80 + 0.95 − 1 = 0.75. For a test measured on a continuous scale, the cut-off that maximises J is one common way to choose a positivity threshold on the ROC curve (§15.5.1). J weights a false negative and a false positive equally, which is rarely true clinically.

MUST-KNOW NOTES (MD-LEVEL)

- odds = p ÷ (1 − p); p = odds ÷ (1 + odds). Post-test odds = pre-test odds × LR.
- LR+ = Se ÷ (1 − Sp); LR− = (1 − Se) ÷ Sp. An LR of 1 means the result does not change the probability of disease.
- A widely used rough guide, from the Evidence-Based Medicine Working Group's users' guides, treats LRs above 10 or below 0.1 as generating "large and often conclusive changes" in post-test probability, 5–10 and 0.1–0.2 as moderate, and 2–5 and 0.2–0.5 as small (Bhandari et al., 2003, citing Jaeschke et al., 1994). It is a convention: whether a shift is "conclusive" depends on where the pre-test probability starts. An LR+ of 9.5 still left a 1% prior at only 8.8%.
- LRs are portable across prevalence only to the extent that Se and Sp are, and §5.5.1 showed they vary with spectrum and setting.
- Odds reappear as the basis of the odds ratio (the ratio of odds in two groups, §10.6.4) and logistic regression (§13.2).

CHECKPOINT 5.4

1. A screening test has excellent sensitivity and specificity but is used in a population with very low prevalence. Explain why most positive results could still be false alarms.
2. The test of §5.5 (Se 95%, Sp 90%) is used in a PHC where 10% of those tested have the disease. Build the 2×2 table for 10,000 people and find the PPV. Then get the same answer with pre-test odds and LR+.

3. Why can a likelihood ratio be carried from one setting to another when a PPV cannot? What limits that portability?

(Answers in the Appendix at the end of the book.)

5.6 Random Variables

DEFINITION

A **random variable** is a rule that assigns a number to each outcome of an experiment. It is written with a capital letter (X); a particular value it takes is written in lower case (x).

SIMPLIFIED EXPLANATION

So far each question was about one event: a positive test, a rolled 4. A random variable packages every numerical result of an experiment, together with its probability, into one object you can calculate with: its average, its spread, its probability of exceeding a threshold.

ILLUSTRATION / EXAMPLE

X = "number of children, out of 5 tested, who have a positive stool test for worms". X can be 0, 1, 2, 3, 4 or 5, each with its own probability. X is the random variable; "X = 3" is one event concerning it.

MUST-KNOW NOTES (MD-LEVEL)

- Random variables are **discrete** (a countable set of values, §5.6.1) or **continuous** (any value in a range, §5.6.2), mirroring the discrete/continuous split of data (recall §1.1.1).
- Every random variable has a **probability distribution**: a complete description of how likely each value, or each range of values, is (§5.7).

5.6.1 Discrete Random Variable

DEFINITION

A **discrete random variable** takes a countable set of values x_1, x_2, ... , each with probability P(X = x). The list of values and their probabilities is the **probability mass function (PMF)**. Its **expected value** (mean) and **variance** are

$$E(X) = \mu = \Sigma\, x \times P(X = x)$$

$$Var(X) = \sigma^2 = \Sigma\, (x - \mu)^2 \times P(X = x).$$

SIMPLIFIED EXPLANATION

The PMF is a relative frequency table for the whole population (recall §1.4.3) rather than for a sample. The expected value is the weighted mean of the possible values, each weighted by its probability (recall the weighted mean, §2.2.2): the long-run average of X over many repetitions, the same idea of expected value first met in §3.6.2 as the average over all possible samples. The variance is the probability-weighted

average squared deviation from μ, the population variance of §3.6.1 written with probabilities in place of relative frequencies.

ILLUSTRATION / EXAMPLE

X = number of ANC visits (capped at 4) made by a randomly chosen woman in a block:

x	0	1	2	3	4
P(X = x)	0.05	0.10	0.15	0.30	0.40

The probabilities sum to 1. $P(X \geq 2) = 0.15 + 0.30 + 0.40 = 0.85$. $E(X) = 0(0.05) + 1(0.10) + 2(0.15) + 3(0.30) + 4(0.40) =$ **2.9 visits**. $Var(X) = \Sigma(x - 2.9)^2 P(x) = 1.39$, so $\sigma = 1.18$ visits.

MUST-KNOW NOTES (MD-LEVEL)

- A PMF's probabilities are all ≥ 0 and sum to exactly 1 (axioms 1 and 2).
- The expected value need not be a possible value: 2.9 visits is not something one woman can make.
- The named discrete distributions (Bernoulli, binomial and Poisson) are in §5.7.1.

5.6.2 Continuous Random Variable

DEFINITION

A **continuous random variable** can take any value in a range. Its probabilities are given by a **probability density function (PDF)**, a curve $f(x) \geq 0$ with total area 1 under it. The probability that X falls between a and b is the area under the curve between a and b.

SIMPLIFIED EXPLANATION

Haemoglobin or diastolic blood pressure can take infinitely many values in any interval, however narrow. Probability is spread over all of them, so any single exact value gets probability zero. Only ranges get non-zero probability, measured as area. The height of the curve is a density (probability per unit of x), not a probability.

ILLUSTRATION / EXAMPLE

P(diastolic BP is exactly 80.000... mmHg) = 0. P(diastolic BP is between 78 and 82 mmHg) is the area under the BP density between 78 and 82, a non-zero number.

MUST-KNOW NOTES (MD-LEVEL)

- $P(X = x) = 0$ for every single value; so $P(X \leq 80) = P(X < 80)$ for a continuous variable.
- Total area under a PDF = 1, the continuous counterpart of a PMF summing to 1.
- Mean and variance are defined as for the discrete case, with the sum replaced by an area (an integral); you will not need to compute them by hand in this book.

CHECKPOINT 5.5

1. Explain why the probability that a continuous random variable takes one exact value is zero, and what you calculate instead.
2. Give one discrete and one continuous random variable from your own clinical work, and justify each classification.
3. For the ANC-visit PMF of §5.6.1, find $P(X \le 1)$ and check that $E(X) = 2.9$.

(Answers in the Appendix at the end of the book.)

5.7 Probability Distributions

A **probability distribution** is the complete description of how probability is spread over a random variable's values: a PMF for a discrete variable, a PDF for a continuous one. A few shapes recur so often that they have names and are fixed by one or two **parameters** (the Binomial by n and p; the Normal by μ and σ). Knowing the name and the parameters determines every probability. §5.7.1 covers the discrete family (Bernoulli, binomial, Poisson); §5.7.2 covers the continuous family (uniform, Normal, and the t, chi-square and F distributions that inference is built on).

5.7.1 Discrete Distributions

Three named distributions cover most counts in community medicine: a single yes/no trial (**Bernoulli**, §5.7.1.1), the number of successes in a fixed number of independent trials (**binomial**, §5.7.1.2), and the number of events in a fixed stretch of time or space (**Poisson**, §5.7.1.3). The binomial is built from Bernoulli trials, and the Poisson is the binomial's limit when trials are many and each is unlikely.

5.7.1.1 Bernoulli Distribution

DEFINITION

A **Bernoulli** random variable takes the value 1 ("success") with probability p and 0 ("failure") with probability $1 - p$.

SIMPLIFIED EXPLANATION

This is one yes/no outcome: CBHI member or not, vaccinated or not, survived or died. The two outcomes need not be equally likely. It is the building block of the binomial.

DERIVATION

From the definitions in §5.6.1:

$E(X) = 1 \times p + 0 \times (1 - p) = p.$

$\text{Var}(X) = (1 - p)^2 \times p + (0 - p)^2 \times (1 - p) = p(1 - p)[(1 - p) + p] = p(1 - p).$

ILLUSTRATION / EXAMPLE

One patient recovers with probability p = 0.8. Recovery is Bernoulli(0.8): mean 0.8, variance 0.8 × 0.2 = 0.16.

MUST-KNOW NOTES (MD-LEVEL)

- Mean = p; variance = p(1 − p). The variance is largest at p = 0.5.
- Any single binary outcome in a dataset is one Bernoulli trial.

5.7.1.2 Binomial Distribution

DEFINITION

A **binomial** random variable X counts the successes in n trials that are independent and share the same success probability p. Its PMF is

$P(X = k) = C(n, k) \times p^k \times (1 - p)^{n-k}$, for k = 0, 1, … , n,

where $C(n, k) = n! \div [k!(n - k)!]$ is the number of ways of choosing which k of the n trials are the successes.

SIMPLIFIED EXPLANATION

X is the sum of n independent Bernoulli trials. To get the probability of exactly k successes you need two things: the probability of any one particular sequence with k successes, and the number of such sequences. The counting rules below supply the second.

DERIVATION: COUNTING RULES

Factorial. The number of ways to arrange n distinct items in a row is $n! = n \times (n - 1) \times \ldots \times 2 \times 1$ ("n factorial"). There are n choices for the first place, n − 1 for the second, and so on. By convention 0! = 1 (there is exactly one way to arrange nothing). Example: 5! = 120.

Permutations. The number of ordered selections of r items from n is ${}^nP_r = n \times (n - 1) \times \ldots \times (n - r + 1) = n! \div (n - r)!$. Example: choosing a chairperson and a secretary from 4 ASHAs, ${}^4P_2 = 4 \times 3 = 12$.

Combinations. If order does not matter, each unordered group of r has been counted r! times among the permutations. So ${}^nC_r = C(n, r) = n! \div [r!(n - r)!]$. Example: choosing 2 ASHAs for training from 4, $C(4, 2) = 12 \div 2! = 6$.

DERIVATION: THE BINOMIAL PMF

1. Take one particular sequence with k successes and n − k failures, say S S F … F. Its trials are independent, so by the multiplication rule its probability is $p \times p \times (1 - p) \times \ldots = p^k(1 - p)^{n-k}$. Every sequence with k successes has this same probability, whatever the order.
2. Different sequences are different outcomes, so they are mutually exclusive. By the addition rule, $P(X = k) = (\text{number of such sequences}) \times p^k(1 - p)^{n-k}$.
3. A sequence is fixed by choosing which k of the n positions hold the successes: C(n, k) ways.

For n = 3, p = 0.8, k = 2: the sequences are SSF, SFS, FSS, so C(3, 2) = 3 and $P(X = 2) = 3 \times 0.8^2 \times 0.2 = 0.384$.

Mean and variance follow from the Bernoulli results. X is a sum of n Bernoulli variables, and expected values add, so E(X) = np. For independent variables the variances also add (proved in §6.5, rule 3; you can check it numerically below), so Var(X) = np(1 − p).

WORKED CALCULATION

Twenty patients each independently recover with probability 0.8. X ~ Binomial(n = 20, p = 0.8), read "X follows a binomial distribution with n = 20 and p = 0.8".

1. C(20, 18) = 20! ÷ (18! × 2!) = (20 × 19) ÷ 2 = 190.
2. $0.8^{18} \times 0.2^2 = 0.000721$.
3. P(X = 18) = 190 × 0.000721 = **0.137**.
4. P(all 20 recover) = $0.8^{20} = 0.0115$, and P(X ≥ 18) = P(18) + P(19) + P(20) = 0.206.
5. Mean = 20 × 0.8 = 16 recoveries; variance = 20 × 0.8 × 0.2 = 3.2. Computing $\Sigma(k - 16)^2 P(X = k)$ over k = 0 to 20 from the PMF also gives 3.2, the numerical check of the variance rule.

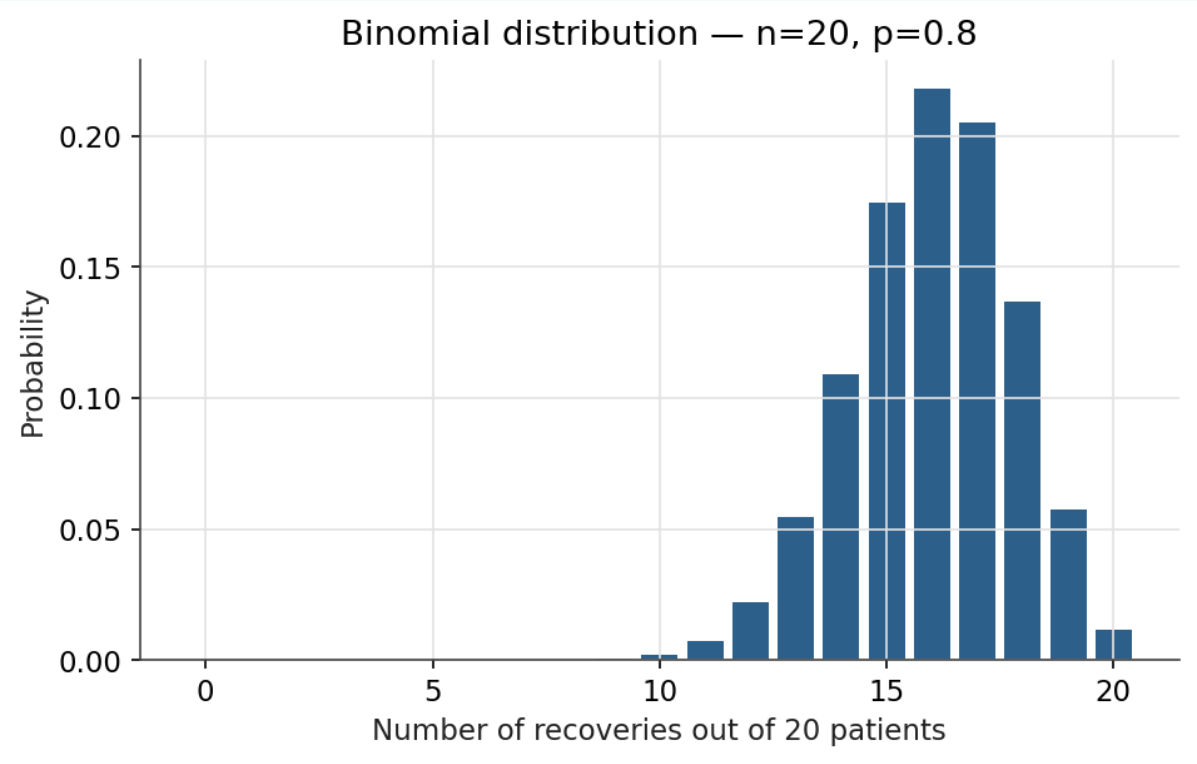


Figure 5.3. Binomial(20, 0.8): the bar at 18 has height 0.137; the distribution peaks at 16, the mean.

MUST-KNOW NOTES (MD-LEVEL)

- $P(X = k) = C(n, k)p^k(1 - p)^{n-k}$; mean np; variance np(1 − p).
- Two conditions: independent trials (§5.4) and the same p for every trial. Infections passing between patients, or recovery chances that differ by severity, break the plain binomial model.

- As n grows the binomial's shape approaches the Normal curve. §5.7.2.2.4 turns this into a working approximation, and the Central Limit Theorem (sums of many independent variables tend to a Normal shape, §6.6) explains it.
- The binomial underlies the sign test (a test that counts how many paired differences are positive, §10.1.1) and exact confidence intervals for a proportion (§7.2.4).

5.7.1.3 Poisson Distribution

DEFINITION

A **Poisson** random variable counts events in a fixed interval of time or space when events occur at a constant average rate λ ("lambda") per interval. Its PMF is

$P(X = k) = e^{(-\lambda)} \times \lambda^k \div k!$, for k = 0, 1, 2, ...

where e = 2.71828... is the constant defined in the derivation below.

SIMPLIFIED EXPLANATION

Some counts have no natural number of trials: new cancer diagnoses in a district in a month, road-traffic deaths on a highway in a week. The Poisson model needs only the average rate. It rests on three assumptions about the process: (1) events occur at a constant rate over the interval; (2) events occur independently of one another; (3) two events do not happen at exactly the same instant. "Rare" is not one of the assumptions. It matters only when the Poisson is used to approximate a binomial, as in the derivation.

DERIVATION: THE POISSON AS THE LIMIT OF THE BINOMIAL

Divide the month into n tiny slots, so small that each holds at most one event, with probability $p = \lambda/n$ in each. The count is then Binomial(n, λ/n), and

$P(X = k) = C(n, k)(\lambda/n)^k(1 - \lambda/n)^{n-k} = [n(n-1)...(n-k+1) \div n^k] \times (\lambda^k \div k!) \times (1 - \lambda/n)^n \times (1 - \lambda/n)^{-k}$.

Let n grow without limit, with k fixed:

1. $n(n-1)...(n-k+1) \div n^k \to 1$ (each of the k factors n, n − 1, ... divided by n tends to 1).
2. $(1 - \lambda/n)^{-k} \to 1$.
3. $(1 - \lambda/n)^n \to e^{(-\lambda)}$. Here **e** is defined as the limit of $(1 + 1/n)^n$ as n grows: 2.594 at n = 10, 2.705 at n = 100, 2.717 at n = 1,000, approaching 2.71828. That $(1 - \lambda/n)^n$ tends to e raised to the power −λ is a standard limit (proof beyond this book); you can check it numerically: with λ = 3, $(1 - 3/1{,}000)^{1000} = 0.0496$, against $e^{(-3)} = 0.0498$.

What remains is $e^{(-\lambda)}\lambda^k \div k!$. Because the binomial mean $np = \lambda$ stays fixed while its variance $np(1 - \lambda/n)$ tends to λ, the Poisson has **mean = variance = λ**.

WORKED CALCULATION

A district records an average of λ = 3 new childhood cancer diagnoses per month.

1. $P(X = 0) = e^{(-3)} =$ **0.050**.
2. $P(X = 5) = e^{(-3)} \times 3^5 \div 5! = 0.0498 \times 243 \div 120 =$ **0.101**.

3. P(at least one) = 1 − 0.050 = 0.950 (complement rule).
4. Binomial check: with n = 3,000 slots and p = 0.001, Binomial P(X = 5) = 0.1008, the same as the Poisson to four places; with n = 30 and p = 0.1 it is 0.1023.

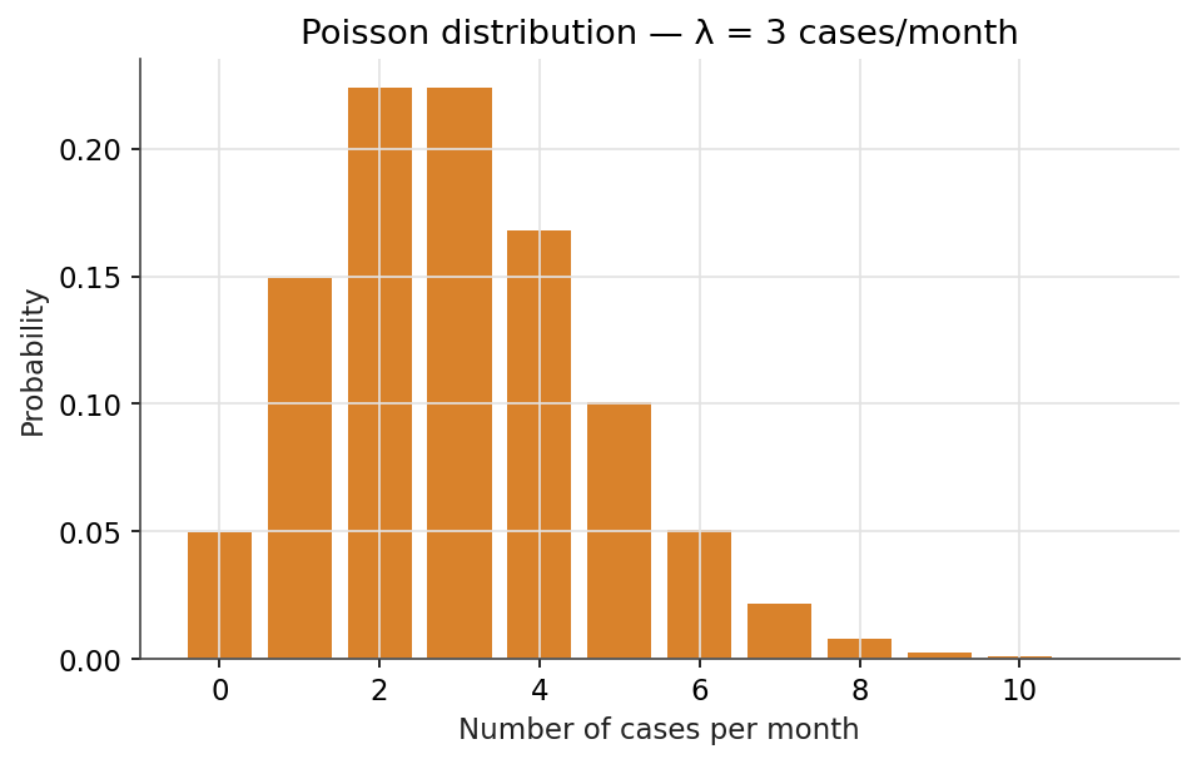


Figure 5.4. Poisson(3): P(0) = 0.050, P(5) = 0.101; mean and variance both 3.

MUST-KNOW NOTES (MD-LEVEL)

- P(X = k) = e^(−λ)λ^k ÷ k!; mean = variance = λ.
- If the observed variance of counts is much larger than the mean (**overdispersion**), the independence or constant-rate assumption has failed, and the negative binomial model (§13.6) is the standard alternative.
- **Does the model apply?** Cases in an infectious outbreak are not independent: each case can cause others, so cases arrive in clusters and counts are overdispersed. The Poisson suits non-contagious events such as cancer diagnoses or road-traffic deaths, not outbreak case counts.
- Poisson approximates the binomial when n is large and p small; one published rule of thumb asks for n ≥ 20 and p ≤ 0.05 (NIST/SEMATECH e-Handbook, n.d.).
- Poisson counts per unit of person-time are rates, the basis of §7.4 and Poisson regression (§13.5).

CHECKPOINT 5.6

1. Explain how a binomial random variable is a sum of Bernoulli trials. For n = 20, p = 0.8, state the mean and variance and show where they come from.

2. Why is the Poisson distribution a reasonable model for "number of new cancer diagnoses in a district per month" but not for "number of patients who recover out of 20 treated"? Why would "number of cases per week in a measles outbreak" usually be overdispersed?
3. Road-traffic deaths on a stretch of highway average $\lambda = 2$ per month. Find P(no deaths next month) and P(at least one).

(Answers in the Appendix at the end of the book.)

5.7.2 Continuous Distributions

The **uniform** distribution (§5.7.2.1) is the simplest continuous shape. The **Normal** distribution (§5.7.2.2) describes some measurements approximately and, more importantly, the behaviour of sample means (§6.6); its subsections turn it into z-scores, reference ranges, growth standards and approximations for counts. The **t**, **chi-square** and **F** distributions (§5.7.2.3–§5.7.2.5) are built from standard Normal variables. They rarely describe raw data; they describe statistics calculated from samples, and they supply the multipliers and cut-offs for confidence intervals and tests.

5.7.2.1 Uniform Distribution

DEFINITION

A **uniform** random variable on [a, b] has constant density $1 \div (b - a)$ between a and b and zero elsewhere.

SIMPLIFIED EXPLANATION

It is the continuous version of the fair die. Every point in [a, b] is equally likely, so the probability of a sub-interval depends only on its width: $P(c \le X \le d) = (d - c) \div (b - a)$.

ILLUSTRATION / EXAMPLE

A random-number generator producing values in [0, 1] is Uniform(0, 1). $P(0.2 \le X \le 0.3) = 0.1 = P(0.7 \le X \le 0.8)$. This is how software draws random samples and allocates participants at random.

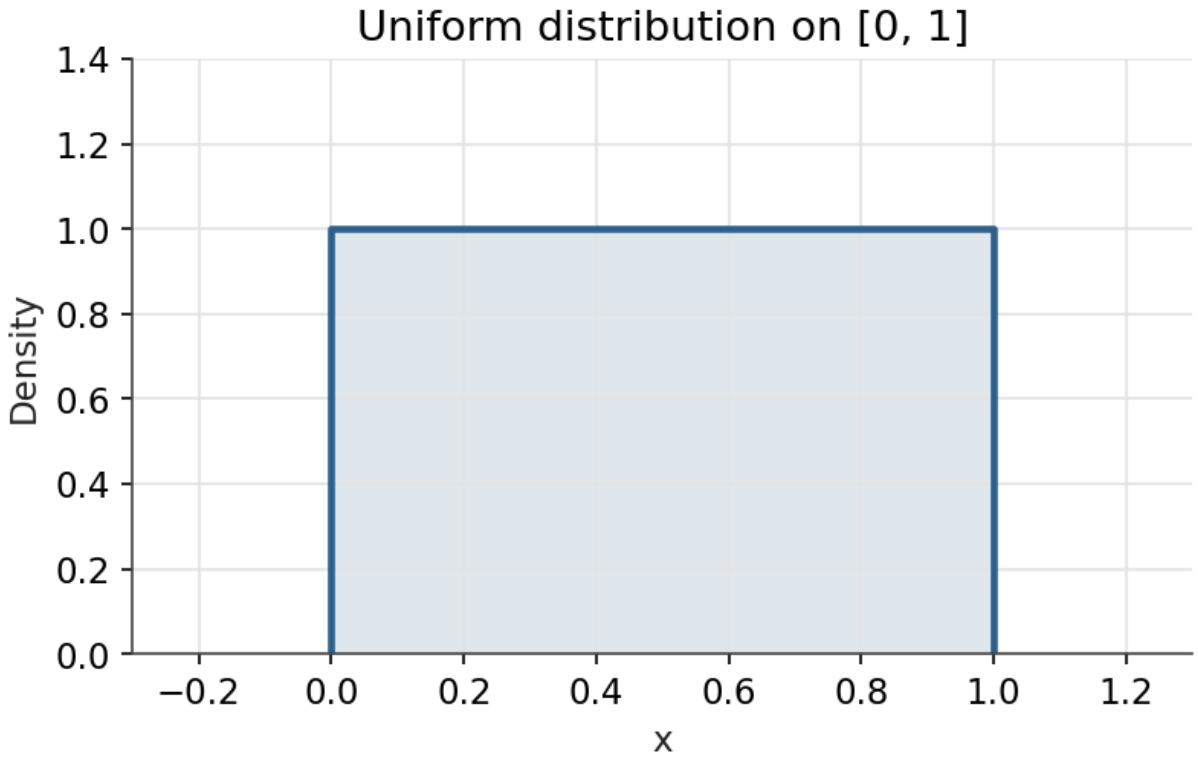


Figure 5.5. Uniform(0, 1): a flat density of height 1.

MUST-KNOW NOTES (MD-LEVEL)

- Mean = (a + b) ÷ 2; variance = (b − a)² ÷ 12.
- Biological measurements seldom look uniform; its uses are in simulation and random allocation.

5.7.2.2 Normal Distribution

DEFINITION

The **Normal (Gaussian) distribution** with mean μ and standard deviation σ is the continuous, symmetric, bell-shaped distribution with density

f(x) = [1 ÷ (σ√(2π))] × e^(−(x − μ)² ÷ (2σ²)).

SIMPLIFIED EXPLANATION

The density is highest at μ and falls away symmetrically; σ sets the width. Some measurements in a homogeneous population are close to Normal, adult height being the standard example. Many are not: many laboratory analytes are skewed, which is why laboratories often transform them before calculating reference limits (Ozarda, 2016; §5.7.2.2.3). The Normal distribution's central role comes from elsewhere. The means of samples tend towards a Normal shape as sample size grows, whatever the shape of the raw variable (the Central Limit Theorem, §6.6), and that is what most confidence intervals and tests rely on.

ILLUSTRATION / EXAMPLE

Adult height in a population is close to Normal with μ = 165 cm and σ = 7 cm: most people are near 165 cm, and very short and very tall people are equally rare.

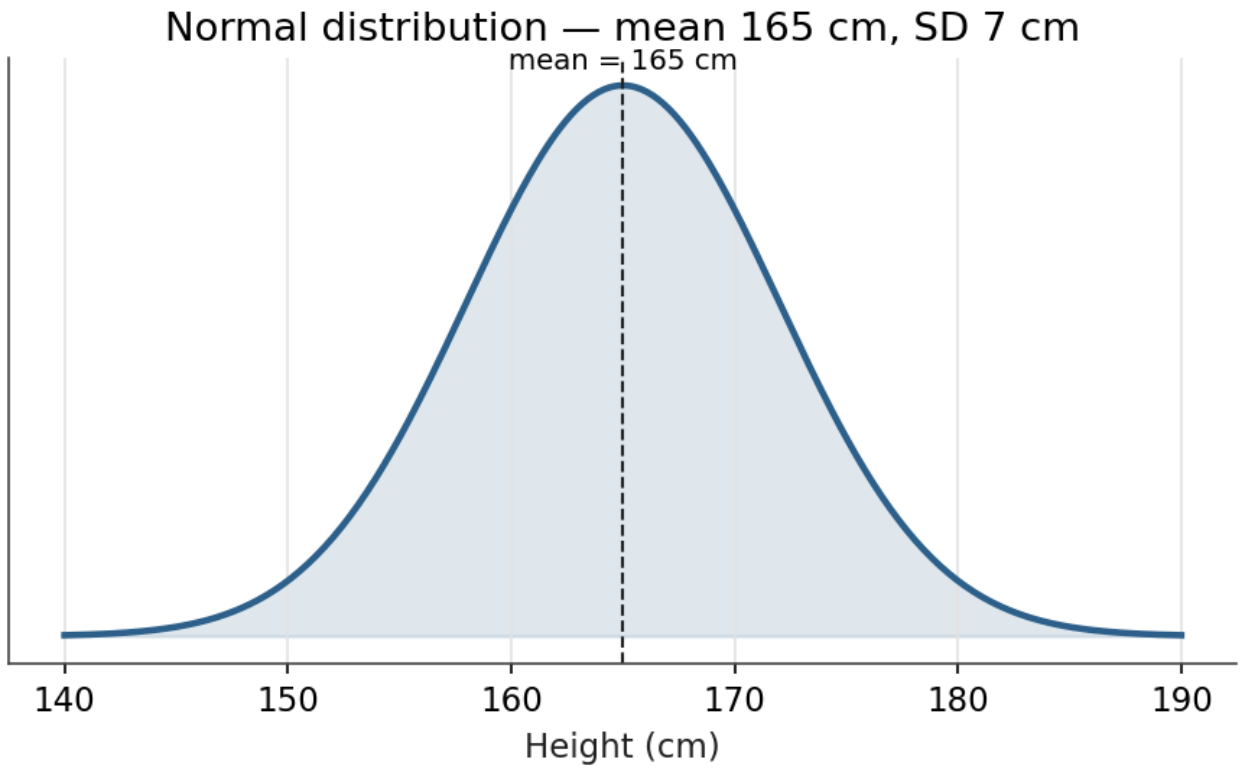


Figure 5.6. A Normal distribution with mean 165 cm and SD 7 cm.

MUST-KNOW NOTES (MD-LEVEL)

- Fully specified by μ and σ.
- Symmetric, so mean = median = mode (recall §2.5 and §3.9).
- Probabilities for any range come from converting to z-scores and reading the standard Normal table (§5.7.2.2.1). The empirical rule (§5.7.2.2.2) is a rounded summary of that table.
- Check shape before assuming normality: a histogram (recall §4.3) or a Normal plot (§9.3) of the raw variable.

5.7.2.2.1 Z-Score and Standardisation

DEFINITION

A **z-score** states how many standard deviations a value lies from the mean: $z = (x - \mu) \div \sigma$. **Standardisation** converts every value of a variable to its z-score. If X is Normal(μ, σ), then $Z = (X - \mu) \div \sigma$ is **standard Normal**: Normal with mean 0 and SD 1.

SIMPLIFIED EXPLANATION

Each Normal variable has its own mean and SD, so a raw value on its own tells you little. Subtracting the mean and dividing by the SD puts every variable on one unit-free scale, "SDs from the mean". A single table of the standard Normal, the **z-table**, then answers probability questions for any Normal variable.

ILLUSTRATION / EXAMPLE

In a hypothetical population of children with mean weight 11 kg and SD 1 kg, a child weighing 13 kg has z = (13 − 11) ÷ 1 = **+2**. The z-table gives P(Z < 2) = 0.977, so the child is heavier than about 97.7% of the population. If the same population's height has mean 86 cm and SD 3.5 cm, the child's height of 90 cm gives z = (90 − 86) ÷ 3.5 = **+1.14** (P(Z < 1.14) = 0.873). Kilograms and centimetres cannot be compared directly, but on the z scale the child is more unusual for weight (+2) than for height (+1.14). The comparison into percentages is valid only if both variables are close to Normal; for skewed variables, §5.7.2.2.3 shows the adjustment used for children's weight.

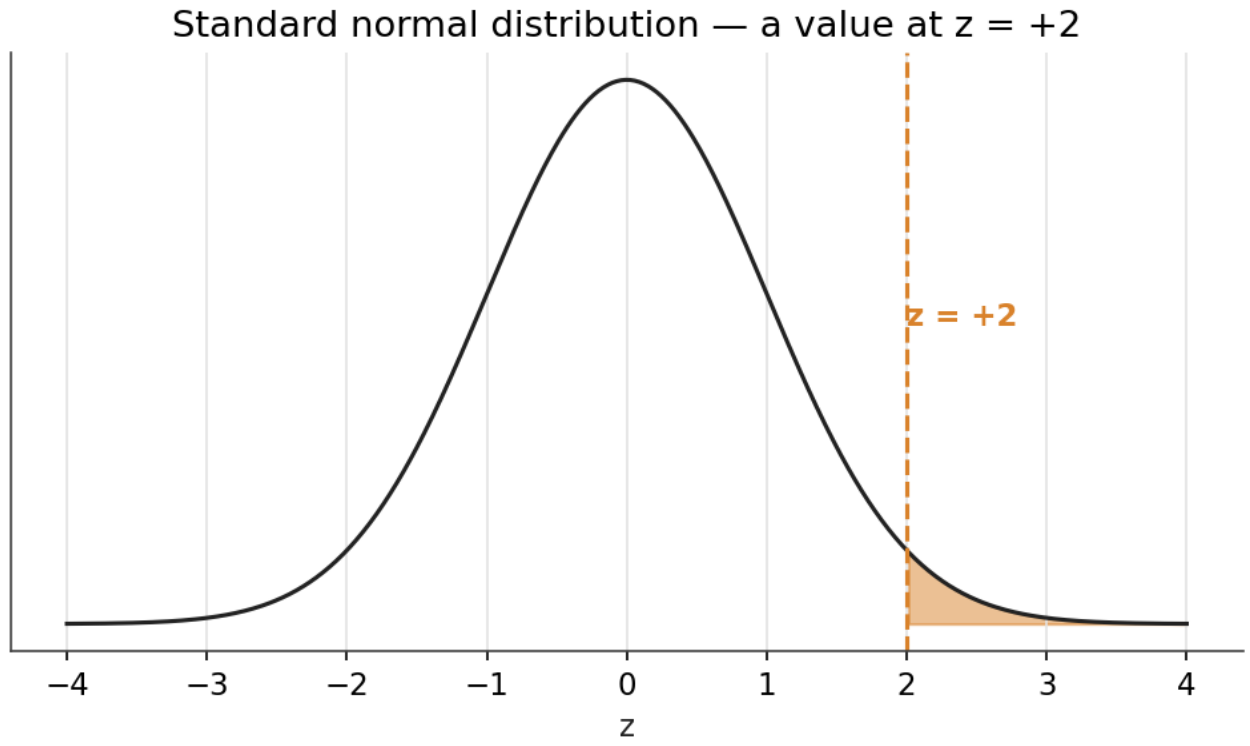


Figure 5.7. The standard Normal curve with z = +2 marked; the shaded tail holds 2.3% of the area.

MUST-KNOW NOTES (MD-LEVEL)

- z = (x − μ) ÷ σ for a population value; z = (x − x̄) ÷ s when the mean and SD are estimated from a sample.
- z = 0 at the mean; positive above, negative below.
- Reading the z-table: the table lists P(Z ≤ z). The value with 97.5% of the area below it is z = 1.96, so 95% of the area lies between −1.96 and +1.96. That 1.96 is the multiplier in 95% confidence intervals (§7.2).
- A z-score describes position, not normality or abnormality. Whether z = −2 matters clinically is a separate judgement.

5.7.2.2.2 Empirical Rule (68-95-99.7)

DEFINITION

The **empirical rule** is a rounded summary of the standard Normal table: about 68% of a Normal distribution lies within 1 SD of the mean, about 95% within 2 SD, and about 99.7% within 3 SD.

SIMPLIFIED EXPLANATION

The rule gives an instant sense of how unusual a distance from the mean is, without a table. It is a rounding. The exact areas are 68.27%, 95.45% and 99.73%, and exactly 95% lies within ±1.96 SD. Use ±2 for a quick mental estimate and ±1.96 whenever you calculate.

ILLUSTRATION / EXAMPLE

With adult height μ = 165 cm and σ = 7 cm, the rule says roughly 95% of adults are between 165 − 2 × 7 = 151 cm and 179 cm. The exact 95% range is 165 ± 1.96 × 7 = 151.3 to 178.7 cm. About 2.3% of adults (the area above z = 2) are taller than 179 cm.

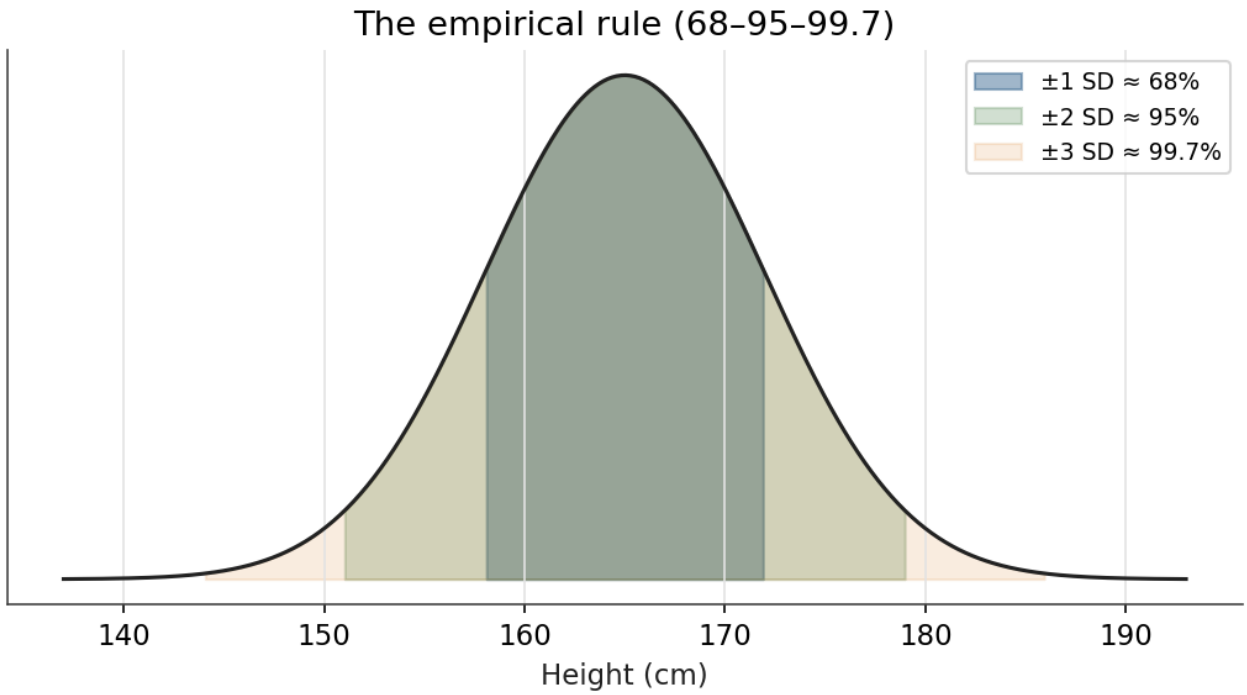


Figure 5.8. The empirical rule for height (mean 165 cm, SD 7 cm): ±1, ±2 and ±3 SD hold about 68%, 95% and 99.7% of adults.

MUST-KNOW NOTES (MD-LEVEL)

- ±1 SD ≈ 68.3%; ±2 SD ≈ 95.4%; ±3 SD ≈ 99.7%. Exactly 95% lies within ±1.96 SD.
- Values beyond ±2 SD are sometimes flagged as “unusual” in exploratory work. This is an informal cousin of the significance level α = 0.05 (the tolerated probability of a false-positive conclusion in a hypothesis test, §8.4), not the same thing.
- The rule applies only to a Normal distribution. For a skewed variable the shares within ±1, ±2 and ±3 SD can be quite different.

CHECKPOINT 5.7

1. A colleague says a z-score of −2 means "something is wrong". Explain precisely what z = −2 tells you and what it does not.
2. What percentage of a Normal variable lies more than 2 SD from the mean in either direction? Give the empirical-rule answer and the exact answer, and say which multiplier gives exactly 5%.
3. Adult height is Normal with mean 165 cm and SD 7 cm. What percentage of adults are taller than 179 cm?

(Answers in the Appendix at the end of the book.)

5.7.2.2.3 Reference Ranges, Centiles and Growth Z-Scores

DEFINITION

A **reference interval** (reference range) is the range that contains a stated central share, conventionally the central 95%, of values in a defined healthy reference population. Its limits are the 2.5th and 97.5th **centiles** (percentiles; recall §3.3). A **growth z-score** is a child's z-score calculated against the median and spread of a growth standard for the same age and sex: height-for-age (HAZ), weight-for-age (WAZ) and weight-for-height (WHZ).

SIMPLIFIED EXPLANATION

A laboratory's "normal range" answers the question: where do healthy people's values lie? The convention sets the limits at the 2.5th and 97.5th centiles of healthy people, and so, by construction, 5% of healthy people fall outside it and are flagged (Ozarda, 2016). If the analyte is Normal in healthy people, those centiles are $\mu \pm 1.96\sigma$. If it is skewed, $\mu \pm 1.96\sigma$ puts the limits in the wrong places, and the laboratory either reads the centiles directly from the data (the non-parametric method) or transforms the values (often by logarithm) until they are Normal, calculates $\mu \pm 1.96\sigma$ on that scale and transforms back (Ozarda, 2016).

Growth charts apply the same idea to children, with the reference changing month by month. The WHO Child Growth Standards tabulate, for every month of age and each sex, the median and the spread of healthy growth, and a child is located by a z-score against them. Stunting is HAZ below −2, underweight is WAZ below −2 and wasting is WHZ below −2, each measured from the median of the WHO Child Growth Standards (WHO Global Health Observatory, n.d.); a z-score below −3 marks the severe form (severe stunting, severe underweight, severe wasting; Boua et al., 2023, Table 1).

For height, the WHO table's distribution is symmetric at each age, so HAZ = (height − median) ÷ SD. Weight is right-skewed, so a single SD does not fit both sides. The WHO tables use the **LMS method** (Cole & Green, 1992), which summarises the distribution at each age by three numbers: M (the median), S (a coefficient of variation) and L (a power that removes the skewness, a Box–Cox power). The z-score is

$$z = [(y \div M)^{L} - 1] \div (L \times S),$$

which reduces to $(y - M) \div (M \times S)$, an ordinary z-score, when $L = 1$.

WORKED CALCULATION 1: A SKEWED LABORATORY VALUE

Suppose serum triglyceride in a healthy reference population is right-skewed, with its logarithm Normal: median 110 mg/dL and SD of ln(value) 0.45 (hypothetical figures).

1. On the original scale the mean is 121.7 mg/dL and the SD 57.7 mg/dL.
2. Naive interval: 121.7 ± 1.96 × 57.7 = **8.7 to 234.8 mg/dL**. The lower limit is far below any healthy value, and 4.6% (not 2.5%) of healthy people lie above the upper limit.
3. Log method: exp(ln 110 ± 1.96 × 0.45) = **45.5 to 265.7 mg/dL**, which contains exactly the central 95%.

WORKED CALCULATION 2: HEIGHT-FOR-AGE (WHO TABLE, L = 1)

A boy aged 30 months measures 84.0 cm. The WHO height-for-age table for boys at 30 months gives median 91.9327 cm and SD 3.4052 cm.

1. HAZ = (84.0 − 91.9327) ÷ 3.4052 = **−2.33**.
2. −2.33 is below −2 but above −3, so the boy is **stunted, not severely** (the −2 and −3 SD heights at 30 months are 85.1 and 81.7 cm).
3. As a centile: P(Z < −2.33) = 0.0099, the 1st centile. As a percentage of the median: 84.0 ÷ 91.93 = 91.4%.

WORKED CALCULATION 3: WEIGHT-FOR-AGE (WHO TABLE, L ≠ 1)

For boys aged 24 months the WHO weight-for-age table gives L = −0.0137, M = 12.1515 kg, S = 0.11426. A boy of that age weighs 9.0 kg.

1. (9.0 ÷ 12.1515)^(−0.0137) = 1.00412.
2. WAZ = (1.00412 − 1) ÷ (−0.0137 × 0.11426) = **−2.63**: underweight, not severely.
3. The skew shows in the table itself: the −2 z-score weight is 2.48 kg below the median (9.67 kg), but the +2 z-score weight is 3.13 kg above it (15.28 kg). A naive (9.0 − 12.15) ÷ (12.15 × 0.114) would give −2.27 and understate how light the child is.
4. Check: inverting the formula, y = M(1 + L × S × z)^(1/L), gives 9.67 kg at z = −2 and 8.63 kg at z = −3, matching the WHO table's printed 9.7 and 8.6 kg.

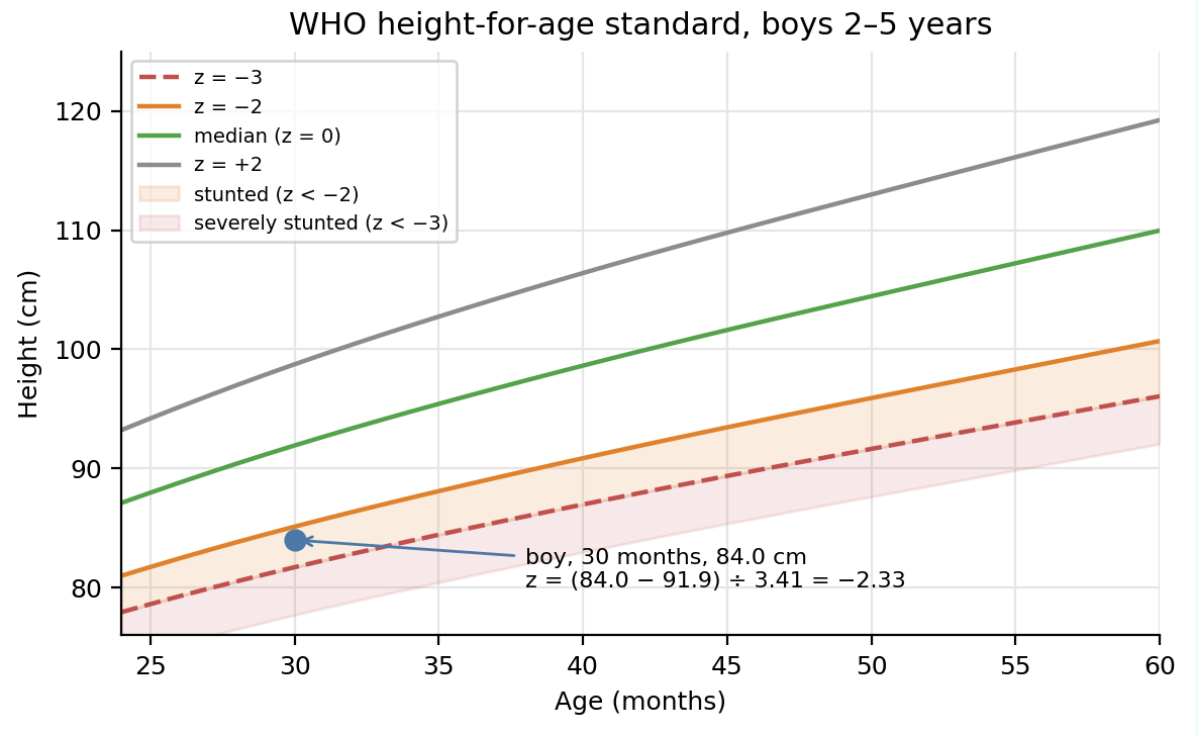


Figure 5.9. WHO height-for-age standard for boys, 24–60 months. The boy of Worked Calculation 2 (30 months, 84.0 cm) sits between the −2 and −3 lines: HAZ −2.33, stunted.

MUST-KNOW NOTES (MD-LEVEL)

- A reference interval is descriptive of health, not a diagnostic threshold. 5% of healthy people fall outside it by construction, and if a healthy person has 20 independent tests, the chance that at least one is flagged is $1 - 0.95^{20} = 64\%$. Clinical decision limits (thresholds tied to outcomes or treatment) are a different thing and should not be confused with reference limits (Ozarda, 2016).
- Use $\mu \pm 1.96\sigma$ only when the analyte is close to Normal in the reference group; otherwise use centiles or a transformation.
- Stunting, underweight and wasting: HAZ, WAZ, WHZ < −2; severe: < −3 (WHO Child Growth Standards).
- z-score, centile and percentage of median are three ways of saying where a child sits. The z-score is equally spaced at the extremes: z = −2 and z = −3 are 1 SD apart, but they correspond to the 2.3rd and the 0.13th centiles, which look almost identical on a centile scale. Percentage of median does not account for the spread at each age, so the same percentage means a different z-score at different ages.
- Growth z-scores in surveys are summarised like any other variable: prevalence of z < −2, and the mean z-score of the population. WHO's LMS tables mean you should use software built on them rather than hand calculation for large datasets.

5.7.2.2.4 Normal Approximations to Binomial and Poisson

DEFINITION

For large enough counts, a binomial variable X ~ Binomial(n, p) is approximately Normal with mean np and SD √[np(1 − p)], and a Poisson variable X ~ Poisson(λ) is approximately Normal with mean λ and SD √λ. The **continuity correction** adds or subtracts 0.5 at the boundary when a discrete count is approximated by a continuous curve.

SIMPLIFIED EXPLANATION

Before computers, binomial and Poisson tail probabilities with large n or λ were tedious sums, so they were read from the Normal curve instead. The approximation still matters because it is the reason the simple z-based confidence interval for a proportion (§7.2.2) and the chi-square test (§10.6) work. The approximation is good when the distribution is not piled up against 0 or n. A widely used criterion for the binomial is that both np and n(1 − p) are at least 5 (NIST/SEMATECH e-Handbook, n.d.); for the Poisson, a mean of at least 5 is quoted as adequate (NIST/SEMATECH e-Handbook, n.d.), and larger is better. The continuity correction treats the bar for count k as the interval k − 0.5 to k + 0.5.

DERIVATION

The mean and SD are those already derived for each distribution (§5.7.1.2 and §5.7.1.3). The Normal shape follows because a binomial count is a sum of n independent Bernoulli variables, and sums of many independent variables tend to a Normal shape (the Central Limit Theorem, §6.6). A Poisson(λ) count is the sum of λ independent Poisson(1) counts when λ is a whole number (independent Poisson counts add to a Poisson count; proof beyond this book), so the same argument applies as λ grows.

WORKED CALCULATION 1: BINOMIAL

In villages where 30% of households are CBHI members, a supervisor visits 50 households. What is P(10 or fewer members)?

1. np = 15 and n(1 − p) = 35, both ≥ 5. Mean 15, SD = √(50 × 0.3 × 0.7) = 3.24.
2. With continuity correction: z = (10.5 − 15) ÷ 3.24 = −1.39; P(Z < −1.39) = **0.082**.
3. Without it: z = (10 − 15) ÷ 3.24 = −1.54; P = 0.061.
4. Exact binomial sum: **0.079**. The corrected approximation is off by 0.004; the uncorrected one by 0.018.

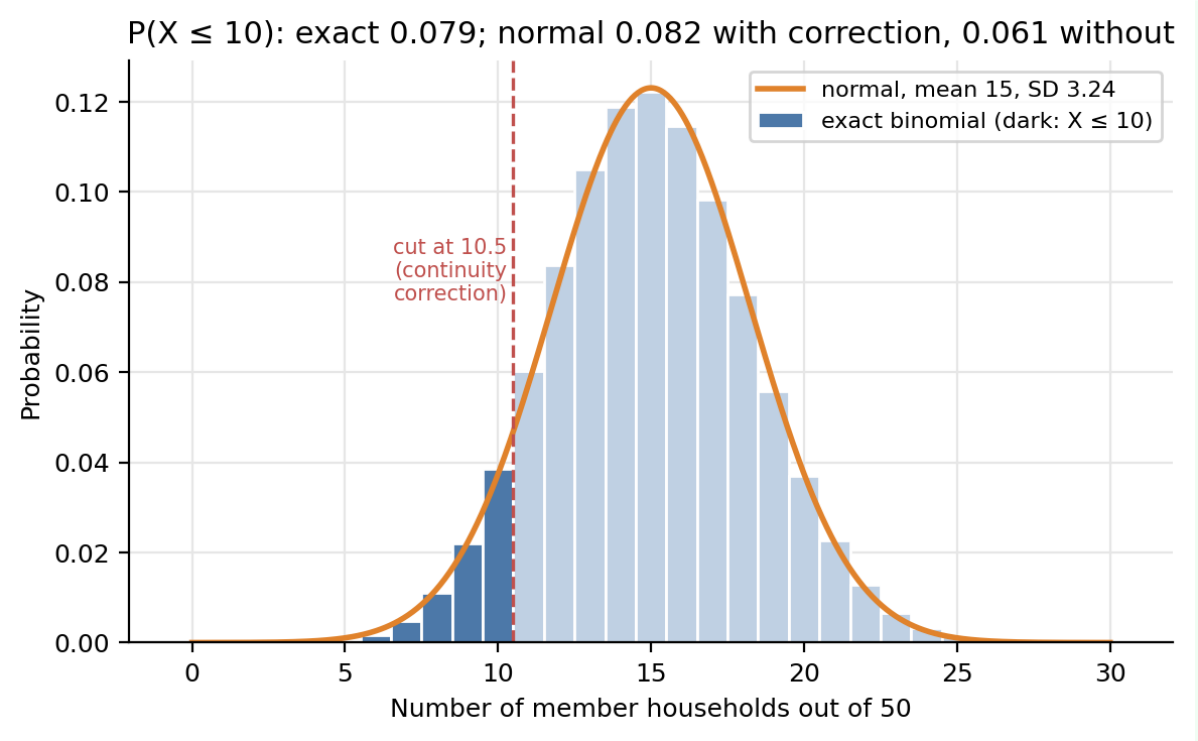


Figure 5.10. Binomial(50, 0.3) bars with the Normal curve (mean 15, SD 3.24). The bars for 0–10 end at 10.5, which is why the continuity correction cuts the curve there.

WORKED CALCULATION 2: POISSON

A district averages λ = 20 road-traffic deaths a month. What is P(30 or more next month)?

1. Mean 20, SD √20 = 4.47.
2. With continuity correction: z = (29.5 − 20) ÷ 4.47 = 2.12; P(Z > 2.12) = **0.017**.
3. Exact Poisson sum: **0.022**. The approximation is in the right region but understates the upper tail, because a Poisson with λ = 20 is still slightly right-skewed. For tail probabilities, use exact calculation when software is available.

MUST-KNOW NOTES (MD-LEVEL)

- Binomial ≈ Normal(np, √[np(1 − p)]); Poisson ≈ Normal(λ, √λ).
- Conditions (a convention): np and n(1 − p) ≥ 5 for the binomial; λ ≥ 5 for the Poisson, better if larger. With n = 20 and p = 0.05, np = 1, and the approximation fails.
- Use the continuity correction: P(X ≤ k) ≈ P(Z < (k + 0.5 − mean) ÷ SD); P(X ≥ k) ≈ P(Z > (k − 0.5 − mean) ÷ SD).
- Approximations are least accurate in the far tails, which is often where p-values (the probability, if nothing is going on, of a result at least this extreme; §8.5) live. Software gives exact binomial and Poisson probabilities (§7.2.4, §10.1.1).

CHECKPOINT 5.8

1. A boy aged 36 months measures 88.0 cm. The WHO height-for-age table gives median 96.0835 cm and SD 3.7069 cm at that age. Calculate his HAZ, classify him, and give his approximate centile.
2. A laboratory reports a "normal range" for serum ALT. Explain why a healthy person can fall outside it, and why $\mu \pm 1.96\sigma$ may be the wrong way to calculate the range.
3. Twenty per cent of children in a district are anaemic. In a random sample of 100 children, use the Normal approximation with continuity correction to find P(30 or more anaemic). Check the condition first.

(Answers in the Appendix at the end of the book.)

5.7.2.3 t-Distribution

DEFINITION

If Z is standard Normal and V is an independent sum of the squares of ν independent standard Normal variables (the chi-square variable of §5.7.2.4), then

$$T = Z \div \sqrt{(V \div \nu)}$$

has a **t-distribution with ν degrees of freedom**. It is symmetric and bell-shaped like the standard Normal but has heavier tails.

SIMPLIFIED EXPLANATION

The z-score $(\bar{x} - \mu) \div (\sigma \div \sqrt{n})$ uses the true σ. In practice σ is unknown and you replace it with the sample SD s. The denominator is now itself random: a small sample can give an s that happens to be too small, which makes the ratio too large. The t-distribution builds that extra uncertainty in, through the random V in its denominator, and pays for it with heavier tails. The more data s is based on, the less $V \div \nu$ varies around 1, and the closer t comes to the standard Normal.

DERIVATION

When the observations are drawn from a Normal population with mean μ and SD σ:

1. $Z = (\bar{x} - \mu) \div (\sigma \div \sqrt{n})$ is standard Normal. ($\sigma \div \sqrt{n}$ is the standard error of the mean, the SD of $\bar{x}$ over repeated samples, derived in §6.5 and §6.7.)
2. $V = (n - 1)s^2 \div \sigma^2$ is chi-square with $n - 1$ degrees of freedom, and it is independent of $\bar{x}$ (both facts are standard results; proof beyond this book).
3. Divide: $Z \div \sqrt{(V \div (n - 1))} = [(\bar{x} - \mu) \div (\sigma \div \sqrt{n})] \div (s \div \sigma) = (\bar{x} - \mu) \div (s \div \sqrt{n})$.

The unknown σ cancels. The familiar ratio $(\bar{x} - \mu) \div (s \div \sqrt{n})$ therefore has a t-distribution with $n - 1$ degrees of freedom (recall degrees of freedom, §3.6.3).

ILLUSTRATION / EXAMPLE

Ten patients' systolic BP has mean 128 mmHg and SD s = 12 mmHg. The standard error is 12 ÷ √10 = 3.79 mmHg. The 97.5% point of t with 9 degrees of freedom is 2.262, not 1.96, so a 95% interval for the mean is 128 ± 2.262 × 3.79 = 128 ± 8.6 mmHg, 15% wider than the ±7.4 that 1.96 would give (§7.2.1 builds this interval). The 97.5% points for other degrees of freedom:

df	4	9	29	99	1,000	∞ (Normal)
t (97.5%)	2.776	2.262	2.045	1.984	1.962	1.960

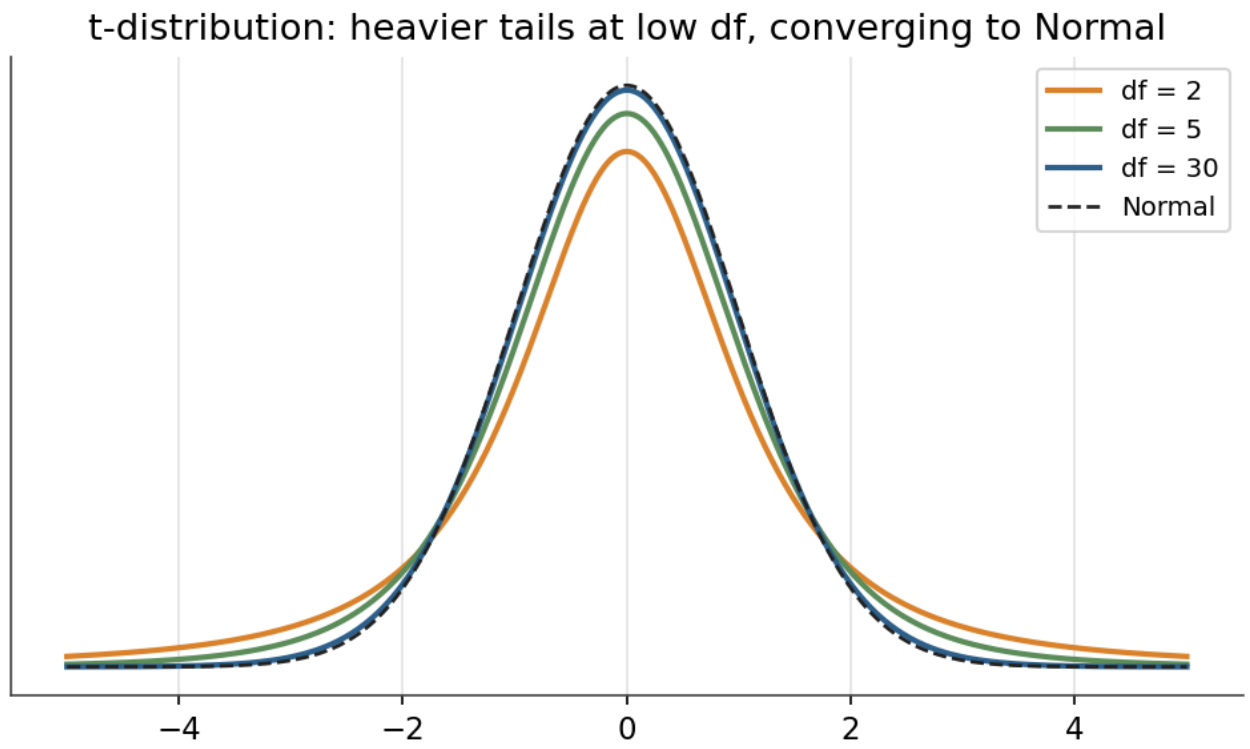


Figure 5.11. t-distributions with increasing degrees of freedom approach the standard Normal curve.

MUST-KNOW NOTES (MD-LEVEL)

- T = Z ÷ √(V ÷ ν); for one sample, (x̄ − μ) ÷ (s ÷ √n) ~ t with n − 1 df.
- Use t whenever σ is estimated from the data, at any sample size. The exact result assumes the population is Normal; for larger n the Central Limit Theorem (§6.6) makes the t-based interval approximately right even for non-Normal data. "Small sample → t, large sample → z" is a poor summary: with n = 30 the multiplier is still 2.045.
- Heavier tails mean wider intervals: with 9 df, 8.2% of the t-distribution lies beyond ±1.96, not 5%.
- The t-distribution underlies the t-tests of §9.2 and the tests of regression coefficients (Chapter 12).

5.7.2.4 Chi-Square Distribution

DEFINITION

If $Z_1, Z_2, \ldots, Z_\nu$ are independent standard Normal variables, then $V = Z_1^2 + Z_2^2 + \ldots + Z_\nu^2$ has a **chi-square (χ^2) distribution with ν degrees of freedom**. It is never negative and is right-skewed; its mean is ν.

SIMPLIFIED EXPLANATION

Squaring deviations made variance tractable (recall §3.6). The chi-square distribution describes a sum of squared standardised deviations. Squares cannot be negative, so the distribution starts at 0. Small sums are common, and now and then one large squared deviation pushes the sum far to the right. With more terms the sum becomes less skewed.

ILLUSTRATION / EXAMPLE

With $\nu = 1$, $V = Z^2$, and $P(Z^2 > 1.96^2) = P(|Z| > 1.96) = 0.05$. So the 95% point of χ^2 with 1 df is $1.96^2 =$ **3.841**, exactly the tabulated value. With $\nu = 5$ the 95% point is 11.07. The chi-square test of association (§10.6) compares observed counts O with the counts E expected when the null hypothesis (the assumption of no association that a test tries to discredit, §8.2) is true, by squaring each difference, dividing by the expected count and summing: $X^2 = \Sigma(O - E)^2 \div E$. Under the null hypothesis, X^2 is approximately χ^2 when expected counts are large enough.

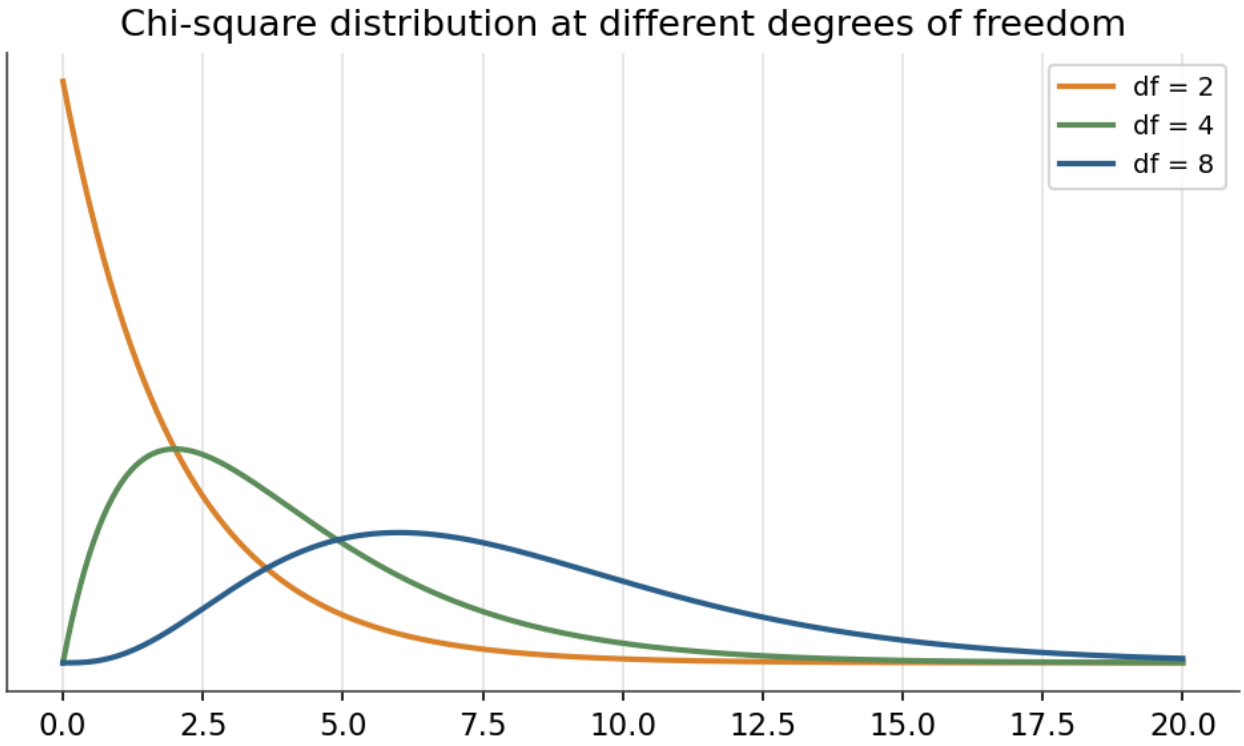


Figure 5.12. Chi-square densities: strongly right-skewed with few degrees of freedom, more symmetric with more.

MUST-KNOW NOTES (MD-LEVEL)

- χ^2_ν = sum of ν squared independent standard Normals; mean ν; never negative.

- χ^2_1 is the square of a z: its 95% point $3.841 = 1.96^2$.
- Uses: the chi-square tests for categorical data (§10.6) and confidence intervals for a variance, through $(n-1)s^2 \div \sigma^2 \sim \chi^2_{n-1}$ for Normal data.

5.7.2.5 F-Distribution

DEFINITION

If V_1 and V_2 are independent chi-square variables with ν_1 and ν_2 degrees of freedom, then $F = (V_1 \div \nu_1) \div (V_2 \div \nu_2)$ has an **F-distribution with (ν_1, ν_2) degrees of freedom**. It is never negative and right-skewed.

SIMPLIFIED EXPLANATION

Each $V \div \nu$ is a scaled estimate of variance. F compares two of them as a ratio. If the two variances are equal, F is near 1; a large F says the numerator variance is bigger. This is the comparison ANOVA (analysis of variance, the comparison of several group means, §9.5) makes: variation between group means against variation within groups.

ILLUSTRATION / EXAMPLE

Three treatment groups of 10 patients each: ANOVA's ratio has (2, 27) degrees of freedom, and its 95% point is 3.35. An observed F above 3.35 would be unusual if all three groups shared one mean. A useful identity links the family: t with ν df, squared, is F with (1, ν) df. The 95% point of F(1, 9) is $5.117 = 2.262^2$, the square of the t multiplier in §5.7.2.3.

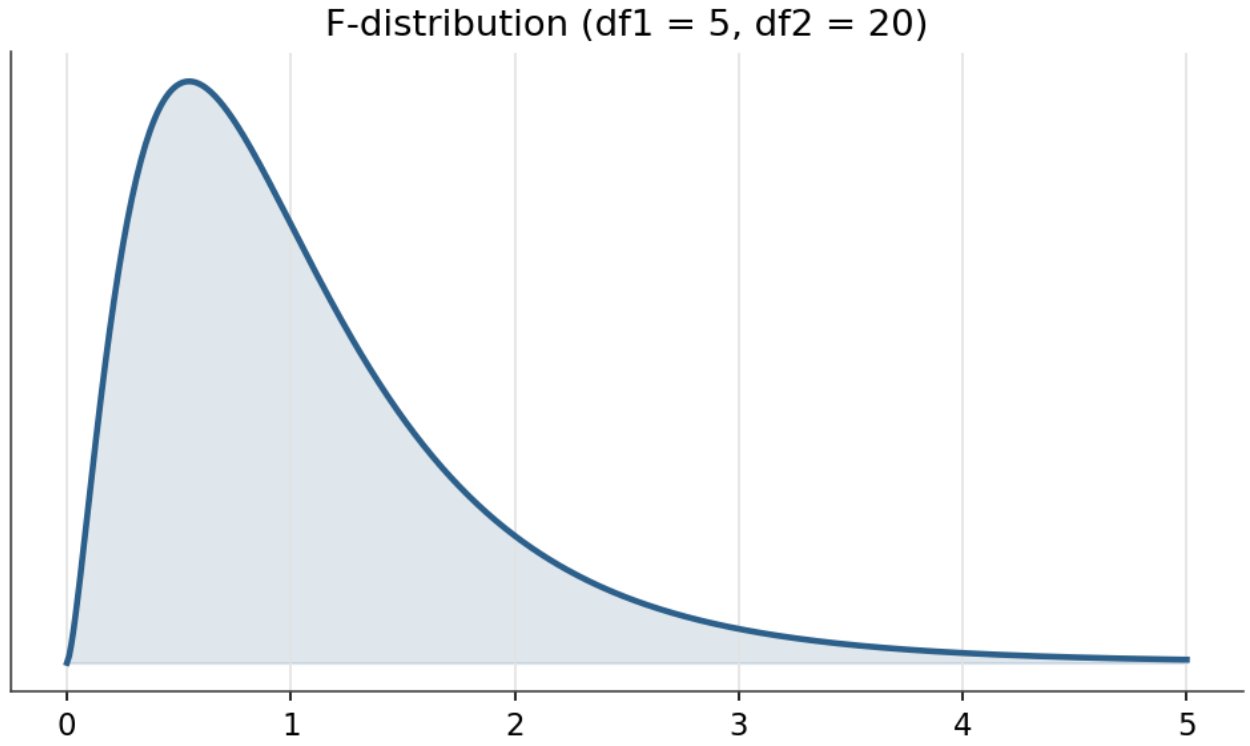


Figure 5.13. An F-distribution with (5, 20) degrees of freedom: never negative, right-skewed, fixed by two degrees of freedom.

MUST-KNOW NOTES (MD-LEVEL)

- $F = (V_1 \div \nu_1) \div (V_2 \div \nu_2)$; two degrees of freedom, numerator first.
- $t^2_\nu = F(1, \nu)$; $z^2 = \chi^2_1$. The four distributions of §5.7.2.2–§5.7.2.5 are one family built from standard Normal variables.
- Uses: ANOVA (§9.5), comparing two variances (§9.2.4), and the overall F-test of a regression model (Chapter 12).

CHECKPOINT 5.9

1. Explain why the t-distribution looks almost identical to the Normal for a large sample but noticeably different for a small one. Use the construction $T = Z \div \sqrt{(V \div \nu)}$ in your answer.
2. What are the t, chi-square and F distributions each built from, and what is each used for?
3. For a sample of 10, the 95% multiplier is 2.262 instead of 1.96. By what percentage does this widen a 95% interval for the mean? Show that 2.262^2 equals the 95% point of $F(1, 9)$, 5.117.

(Answers in the Appendix at the end of the book.)

Sampling

Recall §1.2: we study a sample because measuring the whole population is rarely feasible. This chapter asks the harder question: *how* do you choose a sample so that what it shows can be carried back to the population? It covers the probability methods (including the survey designs used in Indian programmes), the non-probability methods, the difference between chance error and bias, and then the theory that lets one sample speak for the population: the sampling distribution, the Law of Large Numbers, the Central Limit Theorem and the standard error. No amount of correct arithmetic afterwards can rescue a badly chosen sample.

6.1 Why We Sample

DEFINITION

A **census** measures every unit in the population. A **sample survey** measures a subset and uses it to estimate the population's values. Sampling is chosen over a census for cost, time and feasibility, and sometimes because the measurement itself destroys or uses up the unit.

SIMPLIFIED EXPLANATION

Why not always measure everyone? Because a census takes far more time, staff and money, and by the time it is finished the situation it measured may have changed. Some measurements also cannot be taken on every unit without destroying the population itself. A sample is not only cheaper. Because it needs far fewer field staff, those staff can be better trained and supervised, so errors in data collection can be smaller than in a census.

ILLUSTRATION / EXAMPLE

Testing the potency of every vaccine vial in a batch would use up the whole batch: sampling is the only option that leaves any vaccine to give. At the other extreme, a population census is so large that its quality is itself checked by a sample. A sub-sample is re-enumerated "by using better trained and more experienced staff than those employed for the main investigation" (Banda, 2003).

MUST-KNOW NOTES (MD-LEVEL)

- A well-designed sample can match, and sometimes beat, a census in accuracy, because a census carries large non-sampling errors (coverage, response and recording errors, §6.4) that a smaller, better-supervised survey can reduce.

- "Well-designed" is the crucial qualifier. The rest of this chapter is about what makes a design good.

6.2 Probability Sampling Methods

A **probability sample** is one in which every unit in the population has a known, non-zero chance of selection. That known chance is what lets the formal machinery of estimation (Chapter 7) attach a margin of error (how far the result may plausibly lie from the population value by chance alone) to a sample result. Most probability methods need a **sampling frame**: a list, map or register from which units are drawn, such as a village register, an anganwadi child list or a Census list of villages. The six sections below run from the simplest design (simple random sampling) to the large survey designs used by Indian programmes.

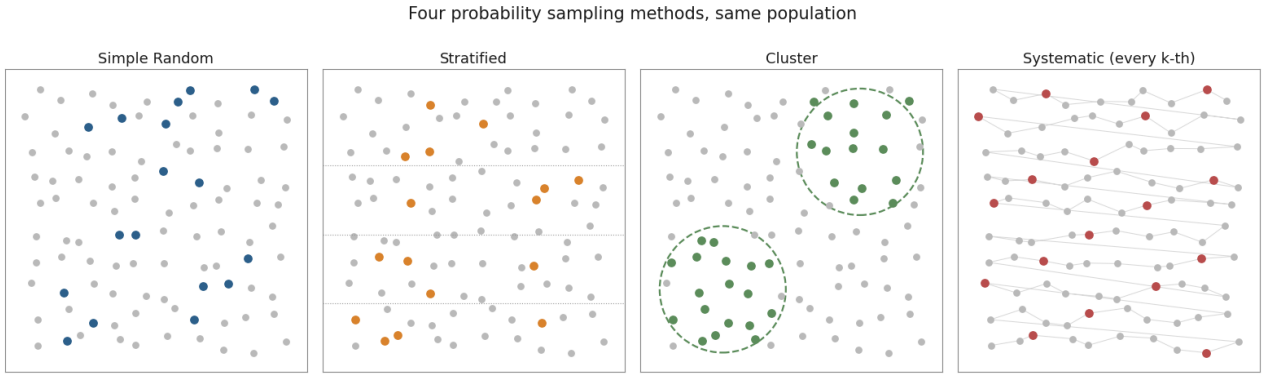


Figure 6.1. Four probability sampling methods applied to the same population.

6.2.1 Simple Random Sampling

DEFINITION

Simple random sampling (SRS) selects n units from N so that every possible sample of size n has the same chance of being chosen. This implies that every unit has the same chance of inclusion, n/N. The reverse is not true: other designs (systematic sampling, cluster sampling with equal-sized clusters) also give every unit the chance n/N but are not SRS, because many possible samples can never occur under them.

SIMPLIFIED EXPLANATION

SRS is the "equally likely outcomes" idea of classical probability (recall §5.1.1) applied to choosing people instead of rolling dice. No subgroup is favoured, and no combination of units is ruled out. When sampling is without replacement, the draws are not independent of each other: once a person is drawn, they cannot be drawn again, which changes the chances for everyone left. What defines SRS is the equal chance of every possible sample, not independence.

ILLUSTRATION / EXAMPLE

Number every child on an anganwadi register of 1,000, then use a random number generator to pick 100 distinct numbers. Every child, whatever their age or hamlet, had the chance 100/1,000 = 0.1 of inclusion, and every set of 100 children was equally likely to be the sample.

MUST-KNOW NOTES (MD-LEVEL)

- SRS needs a complete, current sampling frame of individual units. Building that frame is often the hardest part in the field.
- SRS is the reference design: the formulas for standard errors in §6.5–6.7 and Chapter 7 are derived for it. Other designs are judged by how far their precision departs from SRS of the same size (the design effect, §6.2.6).
- For large, scattered populations SRS is costly: randomly chosen households across a state means travelling to every corner of it. This is why cluster and multistage designs (§6.2.4–6.2.5) exist.

6.2.2 Systematic Sampling

DEFINITION

Systematic sampling arranges the population in a list and selects every kth unit after a random start between 1 and k, where the **sampling interval** k = N ÷ n.

SIMPLIFIED EXPLANATION

One random number sets the start; after that you count forward by k. This is much easier in the field than drawing n random numbers, and every unit still has the chance 1/k = n/N of selection. But only k different samples are possible (one per starting point), not all the combinations SRS allows. That is harmless when the list order is unrelated to what you measure, and dangerous when the list has a cycle that lines up with k.

ILLUSTRATION / EXAMPLE

To sample 100 patients from a clinic register of 1,000: k = 1,000 ÷ 100 = 10. A random start between 1 and 10 gives, say, 7. The sample is patients 7, 17, 27, ... 997, which is 100 patients. Only 10 distinct samples are possible, one for each start from 1 to 10.

MUST-KNOW NOTES (MD-LEVEL)

- k = N ÷ n; the start is random between 1 and k; every unit has inclusion chance 1/k.
- **Periodicity risk:** if the list repeats in a cycle that matches k (for example, a ward register where every 10th bed is the high-dependency bed), every selected unit falls at the same point in the cycle and the estimate is biased. SRS cannot fail this way.
- A list sorted by a relevant variable (village size, literacy) gives an *implicit stratification* that usually improves precision. NFHS-5 sorts its lists this way before selecting (§6.2.5).

6.2.3 Stratified Sampling

DEFINITION

Stratified sampling divides the population into non-overlapping subgroups (**strata**) on a relevant characteristic, then draws a separate random sample within every stratum.

SIMPLIFIED EXPLANATION

SRS can, by bad luck, under-represent a small subgroup. Stratification removes that risk by sampling each subgroup (urban/rural, age group, block) separately. You choose how many to take from each. With **proportional allocation**, each stratum's share of the sample equals its share of the population, and the combined sample can be averaged directly. With **disproportionate allocation**, a small stratum is deliberately over-sampled so that it has enough cases for its own estimate. The overall estimate must then be a weighted mean (recall §2.2.2), with each stratum weighted by its population share.

ILLUSTRATION / EXAMPLE

A district's under-five children are 80% rural and 20% urban. To get a usable urban estimate you sample 50 rural and 50 urban children (disproportionate allocation). Suppose 40% of rural and 20% of urban sampled children are stunted.

WORKED CALCULATION

1. Unweighted (wrong) district estimate: $(0.40 \times 50 + 0.20 \times 50) \div 100 = 30\%$.
2. Weighted estimate, using population shares: $0.80 \times 40\% + 0.20 \times 20\% = 32\% + 4\% = 36\%$.
3. The unweighted figure understates district stunting by 6 percentage points, because it gives urban children 50% of the weight when they are only 20% of the population.

MUST-KNOW NOTES (MD-LEVEL)

- Stratification improves precision when strata are internally similar (homogeneous) and different from each other. Stratifying on an irrelevant characteristic gives little benefit.
- With **proportional allocation**, stratified sampling gives an overall estimate that is (apart from negligible finite-population terms) never less precise than SRS of the same total size, and usually more precise. Disproportionate allocation can *lower* the precision of the overall estimate (the unequal weights add variance) while raising the precision of the small stratum's own estimate.
- Whenever sampling fractions differ between strata, combine them with population weights, as in step 2.

6.2.4 Cluster Sampling

DEFINITION

Cluster sampling divides the population into naturally occurring groups (**clusters**) such as villages, schools or anganwadi centres, selects some clusters at random, and measures units within

the selected clusters. In **one-stage** cluster sampling every unit in a selected cluster is measured. When units are sub-sampled within each selected cluster, the design becomes two-stage (§6.2.5).

SIMPLIFIED EXPLANATION

SRS of households across a district sends field teams to every corner. Cluster sampling sends them to a few villages instead. It also needs a frame only of clusters (a list of villages) plus a list of households inside the few selected villages, not a household list for the whole district.

The price is precision. Households in one village share a water source, a sub-centre, an ASHA and local prices. They resemble each other more than households chosen at random from the whole district. So 25 households from one village carry less independent information than 25 households scattered across the district. This is the opposite of stratification: stratification gains precision by making sure every different group is represented; clustering loses precision because each selected group is internally alike. The size of this loss is measured by the **design effect** (DEFF: the variance of an estimate under the cluster design divided by its variance under SRS of the same size), defined and computed in §6.2.6.

ILLUSTRATION / EXAMPLE

A district has 200 villages. Instead of drawing households from all 200, you randomly select 10 villages and survey every household in them. The field team travels to 10 places, not 200. If all 200 villages had the same number of households, say 100, each household's chance of inclusion would be 10/200 = 0.05, equal for all. Yet this is not SRS: every possible sample consists of whole villages, so a sample of 1,000 households spread thinly over 150 villages, which SRS could produce, can never occur.

MUST-KNOW NOTES (MD-LEVEL)

- Cluster sampling buys feasibility and cost at the price of precision. For the same total number of units, it is usually *less* precise than SRS, because units within a cluster tend to be alike.
- It fits when a list of individuals does not exist but a list of clusters does.
- **Unequal cluster sizes** create a problem: if villages are chosen with equal probability and a fixed number of households is taken from each, a household in a small village has a much higher chance of selection than one in a large village. Selecting villages with probability proportional to size (PPS, §6.2.6) fixes this.
- The CBHI household survey used throughout this book took 25 households in each of 20 villages (500 households). The villages are its clusters, and §6.2.6 computes what that clustering costs.

6.2.5 Multistage Sampling

DEFINITION

Multistage sampling selects units in two or more nested stages: first large units (primary sampling units, **PSUs**, such as villages), then smaller units within the selected ones (households), and possibly smaller units again (one person per household). Each stage uses its own random method.

SIMPLIFIED EXPLANATION

One-stage cluster sampling surveys every household in each selected village, which is heavy when villages are large. Multistage sampling adds a random step inside each cluster: take 22 households from each selected village instead of all of them. It keeps the travel savings of clustering while spreading the sample across more clusters, which recovers precision.

ILLUSTRATION / EXAMPLE

NFHS-5 (2019–21) is the standard Indian example (IIPS, 2021; World Bank Microdata Library, 2022):

1. **Stratification.** Each district is split into rural and urban strata, and rural strata are sub-stratified by village size and the SC/ST share of the population.
2. **Stage 1 (PSUs).** In rural areas, villages are selected with probability proportional to size after sorting by women's literacy (implicit stratification, §6.2.2). In urban areas, Census Enumeration Blocks are selected.
3. **Household listing.** Every selected PSU is mapped and its households listed afresh, creating the second-stage frame.
4. **Stage 2 (households).** "A fixed number of 22 households per cluster was selected with an equal probability systematic selection from a newly created list of households."

In all, 30,456 PSUs were selected across 707 districts, and 636,699 households were interviewed. The design combines the methods of §6.2.2–6.2.4 with PPS (selection in proportion to size, §6.2.6): stratification, PPS cluster selection, fresh listing, and systematic selection within clusters.

MUST-KNOW NOTES (MD-LEVEL)

- Multistage designs are the norm for large household surveys because no national list of households exists, while lists of villages and enumeration blocks do (from the Census).
- Precision is driven mainly by the *number of PSUs*, not only by the number of households; adding households within the same villages helps less and less (§6.2.6).
- Do not confuse **multistage** with **multiphase** sampling. Multistage samples *different units* at each stage (villages, then households). Multiphase samples the *same units* in successive phases: a large first-phase sample gets a cheap measurement (a screening question), and a sub-sample of it gets an expensive one (a blood test).

CHECKPOINT 6.1

1. A researcher must survey households across a state but has only a list of villages, not of households. Which design fits, and why would simple random sampling of households fail here?
2. A district is 70% rural and 30% urban. A survey samples 100 rural and 100 urban households and finds that 25% of rural and 10% of urban households use a PHC for deliveries. Give the unweighted and the correctly weighted district estimate.
3. Stratified and cluster sampling both begin by dividing the population into groups. State the structural difference, and its effect on precision.

(Answers in the Appendix at the end of the book.)

6.2.6 Survey Designs in Indian Programmes — PPS, 30×7 Cluster, LQAS, Rapid Survey and Kish Grid; Design Effect

DEFINITION

- **Probability proportional to size (PPS)** selects clusters so that each cluster's chance of selection is proportional to its size (its population or number of households), usually by systematic selection from a cumulative list of sizes.
- The **design effect (DEFF)** is the variance of an estimate under the actual design divided by its variance under SRS of the same total size: DEFF = Var(design) ÷ Var(SRS) (Bilukha, 2008). Here the variance of an estimate means its spread over repeated samples drawn by the same design (§6.5 develops this). Its square root, **DEFT**, is the factor by which the design inflates the standard error (the square root of that variance, §6.4 and §6.7).
- The **WHO EPI 30×7 cluster survey** is a two-stage survey of 210 children, 7 in each of 30 clusters chosen by PPS, designed to estimate immunisation coverage.
- **Lot quality assurance sampling (LQAS)** takes a small random sample (commonly 19) from each supervision area and classifies the area as reaching or not reaching a coverage target, using a pre-set decision rule.
- A **rapid survey** is a small, quick probability survey (such as the 30×7 design or LQAS) built for programme decisions within days or weeks rather than precise estimation.
- The **Kish grid** is a pre-printed selection table that picks one eligible person in a selected household at random, using only the number of eligible members and a table assigned to that household.

SIMPLIFIED EXPLANATION

These designs answer practical field questions. Villages differ in size, so how do you give every *person* an equal chance when you choose *villages* first? Use PPS. How much precision do you lose by clustering? Measure it with the DEFF. How do you get a coverage estimate in a fortnight with limited staff? Use the 30×7 design. Your block manager does not need a coverage estimate, only to know which sub-centre areas are failing: use LQAS. You have reached the household; whom do you interview? Use the Kish grid rather than whoever opens the door.

DERIVATION

Why PPS gives every person the same chance. Let village i have M_i persons, total M, and select k villages by PPS, then a fixed number m of persons in each selected village.

P(person selected) = P(village i selected) × P(person selected | village i selected) = $(k \times M_i/M) \times (m/M_i) = k \times m / M$

M_i cancels, so every person in the district has the same chance $k \times m / M$, whatever the size of their village. Such a design is called **self-weighting**: the sample can be averaged without weights.

What the DEFF depends on. For a cluster sample with m units per cluster, the DEFF is approximately

$$DEFF = 1 + (m - 1) \times \rho$$

where ρ is the **intracluster correlation coefficient** (ICC: the share of the outcome's total variance that lies between clusters rather than within them; developed fully in §15.2.1). The two extremes show why the formula has this shape (a full proof is given with the ICC in §15.2.1):

- ρ = 0: units in a cluster are no more alike than random units, so DEFF = 1 and clustering costs nothing.
- ρ = 1: every unit in a cluster is identical, so the m units carry the information of one. DEFF = 1 + (m − 1) = m, and n units give the information of only n ÷ m clusters.

Dividing the sample size by the DEFF gives the **effective sample size**, n ÷ DEFF: the size of an SRS with the same precision.

ILLUSTRATION / EXAMPLE

1. PPS on a cumulative list. A block has 10 villages (data/pps_villages.csv) and you must select 3 by PPS.

Village	Population	Cumulative	Numbers assigned	Selected?
PV01	1,450	1,450	1–1,450	
PV02	820	2,270	1,451–2,270	Yes (1,873)
PV03	2,310	4,580	2,271–4,580	
PV04	640	5,220	4,581–5,220	
PV05	1,780	7,000	5,221–7,000	
PV06	3,050	10,050	7,001–10,050	Yes (7,206.3)
PV07	990	11,040	10,051–11,040	
PV08	1,210	12,250	11,041–12,250	
PV09	2,600	14,850	12,251–14,850	Yes (12,539.7)
PV10	1,150	16,000	14,851–16,000	

WORKED CALCULATION (PPS)

1. Total population M = 16,000. Sampling interval = 16,000 ÷ 3 = 5,333.3.
2. Random start between 1 and 5,333: suppose the random number drawn is 1,873.
3. Selection numbers: 1,873; 1,873 + 5,333.3 = 7,206.3; 1,873 + 2 × 5,333.3 = 12,539.7.
4. Find each number's range in the cumulative column: 1,873 falls in 1,451–2,270 (PV02); 7,206.3 in 7,001–10,050 (PV06); 12,539.7 in 12,251–14,850 (PV09).
5. Selection chances differ by size: PV06 had 3 × 3,050 ÷ 16,000 = 0.572; PV04 had 3 × 640 ÷ 16,000 = 0.120.
6. Now take m = 20 persons in each selected village. Every person's chance is 3 × 20 ÷ 16,000 = 0.00375, the same in every village (the Derivation above).
7. Compare choosing 3 villages by SRS (chance 3/10 each) and then 20 persons: a person in PV04 would have chance 0.3 × 20/640 = 0.0094, and a person in PV06 0.3 × 20/3,050 = 0.0020. Small-village residents would be 4.8 times over-represented, and the sample would need weights.

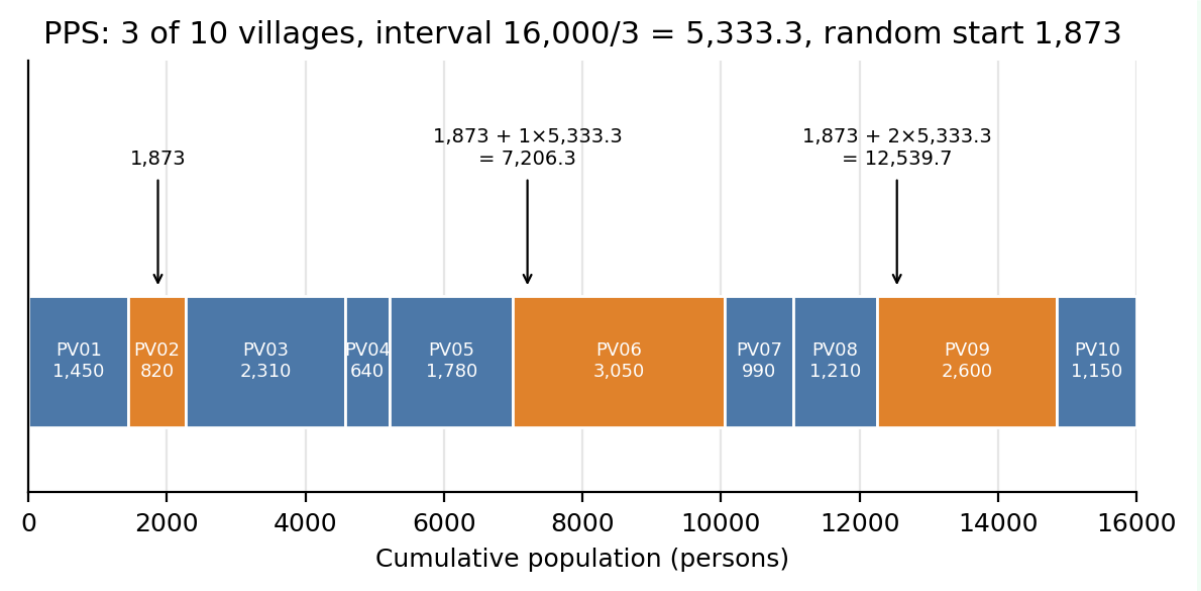


Figure 6.2. PPS selection of 3 of 10 villages (data/pps_villages.csv). Each village occupies a stretch of the cumulative line equal to its population, so large villages are more likely to contain a selection number. Orange: selected villages.

WORKED CALCULATION (DESIGN EFFECT IN THE CBHI SURVEY)

The CBHI survey has 500 households in 20 villages, m = 25 per village, and the village ICC for membership is ρ = 0.06.

1. DEFF = 1 + (25 − 1) × 0.06 = 1 + 1.44 = 2.44.
2. Effective sample size = 500 ÷ 2.44 ≈ 205 households. The 500 clustered households estimate membership about as precisely as 205 households chosen by SRS.
3. DEFT = √2.44 = 1.56: standard errors are about 1.56 times their SRS values. The SRS standard error of the 30% membership prevalence is 0.021; allowing for villages it is 0.032 (the survey analysis in §15.2.2 gives an observed DEFF of 2.41, close to the formula's 2.44).
4. If ρ had been 1, DEFF would be 25 and the 500 households would be worth only 500 ÷ 25 = 20 households, one per village.

2. The WHO EPI 30×7 survey. The Expanded Programme on Immunization has promoted a 30-cluster design with 7 children per cluster since 1978 (Bilukha, 2008; Litvinov et al., 1986). Clusters are selected by PPS; within each, 7 eligible children (for example, aged 12–23 months) are enrolled. Surveys adapted from the WHO 30-cluster method state its working parameters as 95% confidence, a precision of ±10 percentage points and an assumed DEFF of 2 (Mugali et al., 2017). Why 210 children deliver about ±10 points is checked in §6.7, once the standard error is available. Two known weaknesses:

- The traditional EPI method of picking households inside a cluster is not a random selection, and field teams short of time may drift towards conveniently or centrally located houses, biasing the estimate (Bilukha, 2008).
- A design that assumes DEFF = 2 is too small when the outcome is strongly clustered. In emergency surveys, design effects of 5 and higher have been observed for vaccination coverage, and field manuals now advise calculating the sample size for each survey (Bilukha, 2008; sample-size formulas: §7.3).

3. LQAS. A district divides its area into supervision areas (sub-centre areas, for example) and interviews 19 randomly selected children aged 12–23 months in each. The programme sets an upper threshold (the target, say 80% full immunisation) and a lower threshold 30 points below it (50%), as in the Gavi LQAS toolkit, which notes that the sample size most commonly used is n = 19 (Gavi Zero-Dose Learning Hub, n.d.). A decision rule d is chosen from the binomial distribution (recall §5.7.1.2): the area "passes" if at least d of the 19 are immunised.

WORKED CALCULATION (LQAS DECISION RULE, N = 19, THRESHOLDS 80% AND 50%)

1. α (risk of failing an area whose true coverage is 80%: a false alarm, the Type I error of §8.3) = $P(X \le d - 1)$ for $X \sim \text{Binomial}(19, 0.8)$.
2. β (risk of passing an area whose true coverage is only 50%: a missed problem, the Type II error of §8.3) = $P(X \ge d)$ for $X \sim \text{Binomial}(19, 0.5)$.
3. For d = 12: α = 0.023, β = 0.180. For d = 13: α = 0.068, β = 0.084. For d = 14: α = 0.163, β = 0.032.
4. d = 13 is the only rule that keeps both errors below 10%. Rule: pass the area if 13 or more of 19 children are fully immunised; otherwise flag it for action.

LQAS classifies; it does not estimate each area's coverage precisely. Across areas, the samples can be combined (weighted by population) into a district estimate. A global review found 805 LQAS surveys between 1984 and 2004, 266 of them on immunisation coverage (Robertson & Valadez, 2006).

4. The Kish grid. In India's National NCD Monitoring Survey, "one adult (18-69 years) per household was selected using the World Health Organization-KISH grid" (Mathur et al., 2021). The interviewer lists the eligible members in a fixed order (in Kish's form, males by decreasing age, then females), then reads the member's serial number off the selection table printed on that household's cover sheet (Kish, 1949, as described by Nemeth, n.d.). Using the grid instead of "whoever answers the door" prevents a sample full of people who are usually at home. One consequence must be handled in analysis: with one person per household, a person in a household with 4 eligible adults has one-quarter the within-household chance of a person living alone, so a person's chance of selection is inversely proportional to the number of eligible adults in the household (Nemeth, n.d.), and estimates need weights proportional to that number.

MUST-KNOW NOTES (MD-LEVEL)

- **PPS + fixed take per cluster = self-weighting.** This is why the 30×7 survey and NFHS take a fixed number of units in PPS-selected clusters. A cluster larger than the sampling interval can be hit twice; it is then given two "takes" (in the example no village exceeds 5,333.3).
- **DEFF = Var(design) ÷ Var(SRS) ≈ 1 + (m − 1)ρ.** Effective n = n ÷ DEFF. DEFF grows with the cluster size m even when ρ is small: at ρ = 0.06, 25 per cluster gives 2.44 but 7 per cluster gives 1.36. For a fixed total, more clusters with fewer units each is more precise, at a higher travel cost.
- Stratification with proportional allocation usually gives DEFF below 1; clustering usually above 1; unequal weights (disproportionate allocation, Kish-grid selection) push it up.
- **30×7:** 210 children, 30 PPS clusters, ±10 points, DEFF assumed 2. Know its logic and its two weaknesses (non-random household selection within clusters; DEFF 2 often too low).

- **LQAS:** n = 19 per supervision area, decision rule d from the binomial so that both classification errors stay below about 10%. It answers "which areas are failing?", not "what is each area's coverage?".
- **Kish grid:** random selection of one respondent per household; requires weighting by the number of eligible members.

CHECKPOINT 6.2

1. In the PPS example, what was the chance of selection for village PV01 (population 1,450), and what is the overall chance that a given resident of PV01 is interviewed if 20 persons are taken in each selected village?
2. A cluster survey takes 10 households in each of 30 villages (300 households). The ICC for the outcome is 0.05. Compute the DEFF and the effective sample size.
3. Explain why an LQAS area with 12 of 19 children immunised is "flagged" under the rule in this section, and what error the programme risks by using d = 12 instead of d = 13.

(Answers in the Appendix at the end of the book.)

6.3 Non-Probability Sampling Methods

In **non-probability sampling** the chance that a given unit is selected is unknown and cannot be calculated. Selection rests on convenience, judgement, quotas or referral. The trade is feasibility for the ability to attach a valid margin of error or to generalise formally to a population. That trade is sometimes necessary (no frame exists, the population is hidden, the study is qualitative), but its limits must be stated in the report. The four standard forms follow.

6.3.1 Convenience Sampling

DEFINITION

Convenience sampling includes whichever units are easiest to reach, with no randomisation.

SIMPLIFIED EXPLANATION

You take whoever is in front of you. It needs no frame and no planning. That same ease is why it generalises worst: the people who are easy to reach usually differ from the people who are not, in ways connected to the question.

ILLUSTRATION / EXAMPLE

Surveying whichever patients are in the OPD on a Tuesday morning about satisfaction with the hospital is fast. But Tuesday-morning attenders may differ from evening or weekend attenders, and all of them differ from people who avoid the hospital, which is exactly the group a satisfaction survey most needs to hear from.

MUST-KNOW NOTES (MD-LEVEL)

- Fastest and cheapest, and carries the highest risk of selection bias of the methods in this chapter.
- Do not present a convenience sample as representative of a wider population; state explicitly who was likely to be missed.
- Legitimate for pilot studies and for generating hypotheses, not for confirmatory claims about a population.

6.3.2 Purposive Sampling

DEFINITION

Purposive (judgemental) sampling deliberately selects units that the researcher judges best able to inform the study's aim.

SIMPLIFIED EXPLANATION

Selection is intentional and reasoned rather than random or merely convenient. It is the natural choice when depth of insight from particular people matters more than representativeness, as in most qualitative research.

ILLUSTRATION / EXAMPLE

In a qualitative study of barriers to CBHI enrolment, a researcher interviews village health committee members and scheme coordinators, because their inside knowledge of administrative barriers is what the question needs. A random sample of villagers might know little about how the scheme is run.

MUST-KNOW NOTES (MD-LEVEL)

- The sample is only as good as the selection criteria; poorly reasoned purposive selection can be as biased as convenience sampling.
- It supports no formal statistical generalisation, but it is often the right choice for qualitative designs.

6.3.3 Quota Sampling

DEFINITION

Quota sampling sets target numbers (quotas) for subgroups, as stratified sampling does, but fills each quota by non-random, convenience selection.

SIMPLIFIED EXPLANATION

It is stratified sampling's structure without stratified sampling's randomness. Subgroup representation is guaranteed; who is chosen inside each subgroup is left to the interviewer.

ILLUSTRATION / EXAMPLE

A field team must interview 50 men and 50 women in a locality. They approach whoever they meet until each quota of 50 is filled. The sex split is guaranteed; the men and women who happen to be at home, or on the main road, are the ones interviewed.

MUST-KNOW NOTES (MD-LEVEL)

- The single difference from stratified sampling: no random selection within the subgroups.
- Faster and cheaper than stratified sampling, because no frame is needed within subgroups, but it carries convenience sampling's selection bias inside each quota.

6.3.4 Snowball Sampling

DEFINITION

Snowball sampling recruits a few initial participants and asks each to refer others from their own networks, so the sample grows through chains of referral.

SIMPLIFIED EXPLANATION

It is built for populations that no frame lists, because membership is hidden, stigmatised or simply not recorded. Existing participants become the way to find new ones.

ILLUSTRATION / EXAMPLE

Studying people who inject drugs, or unregistered traditional healers: no list exists and direct approaches fail. A researcher finds a few initial contacts, who introduce others they know, and the sample grows along these chains.

MUST-KNOW NOTES (MD-LEVEL)

- Suited to hidden or stigmatised populations where other methods cannot find enough participants.
- Strong bias towards well-connected people: isolated members, or those whose networks do not touch the initial contacts, are rarely reached however many chains are followed.

6.4 Sampling Error vs Bias

DEFINITION

- **Sampling error** is the difference between a sample statistic and the population parameter that arises by chance, because only part of the population was measured. For one sample it is $\bar{x} - \mu$, a number you never know, since μ is unknown.
- The **standard error (SE)** is the standard deviation of the sampling error over all possible samples of the same design: the typical size of the sampling error (defined formally in §6.7). You cannot know the sampling error of your sample, but you can estimate its typical size.

- **Non-sampling error** is every error not caused by sampling: an incomplete or outdated frame (coverage error), non-response, measurement and interviewer error, and data-entry or processing error (Banda, 2003). It occurs in a census too.
- **Bias**, in the design sense used in this section, is a systematic error from how units were selected (**selection bias**) or measured (**measurement bias**) that pushes results in one direction. It is the systematic part of non-sampling error, and it does not shrink as the sample grows.

SIMPLIFIED EXPLANATION

Sampling error is the honest static of measuring a part instead of the whole (recall §1.3: a statistic differs from its parameter). It is random in direction, and it shrinks as the sample grows. Bias is different in kind: it is baked into how the sample was chosen or measured, it points in one direction, and a bigger sample of the same flawed kind only gives a more precisely wrong answer. The total error of a survey is sampling error plus non-sampling error, and "if not properly controlled nonsampling error can be more damaging than sampling error for large-scale household surveys" (Banda, 2003).

A second meaning of "bias" appears in §6.5 and Chapter 7: **estimator bias**, the gap between the average of an estimator over repeated samples and the parameter. The divide-by-n variance is biased in this sense (recall §3.6.2), but its bias shrinks to zero as n grows. When this book says bias "does not shrink with n", it means design bias; say which sense you mean.

ILLUSTRATION / EXAMPLE

A genuinely random sample of 50 children's weights has a mean that differs a little from the true mean, by chance. That is sampling error, and 500 children instead of 50 shrinks it roughly threefold (§6.7 shows why: $\sqrt{10} \approx 3.2$). If the "random" sample was drawn only from a well-funded urban clinic, missing poorer rural children who tend to weigh less, the mean will be too high whatever the sample size. That is selection bias.

	Sampling error	Standard error	Non-sampling error
What it is	$\bar{x} - \mu$ for this sample	SD of $\bar{x} - \mu$ over repeated samples	Frame, non-response, measurement, processing errors
Known for your sample?	No	Estimated ($s/\sqrt{n}$ for a mean)	Rarely; assessed by re-survey or checks
Direction	Random	Not applicable (a spread)	Often systematic (bias)
Larger n	Shrinks on average	Shrinks as $1/\sqrt{n}$	Does not shrink; may grow
Occurs in a census?	No	No	Yes

MUST-KNOW NOTES (MD-LEVEL)

- A larger sample reduces sampling error and does nothing to fix design bias. Sample size and design quality are separate levers; one cannot make up for the other. A convenience sample of 10,000 can be less trustworthy than a random sample of 100.
- Non-sampling error "may increase with increases in sample size" (Banda, 2003): more interviewers and more forms mean more chances to err. This is the reason a well-run sample can beat a census (§6.1).

- Probability sampling controls selection bias; adequate sample size controls sampling error; training, piloting, supervision and re-checks control non-sampling error.

CHECKPOINT 6.3

1. State the single structural difference between quota sampling and stratified sampling, and explain the specific bias that snowball sampling risks.
2. The true mean weight of under-five children in a district is 11.8 kg. A team weighs 2,000 children attending a well-funded urban clinic and gets a mean of 12.6 kg; its standard error is 0.04 kg. How large is the error of this estimate, how much of it could sampling error plausibly explain, and would weighing 8,000 clinic children fix it?
3. Name three sources of non-sampling error in a household survey and say which of them would also affect a census.

(Answers in the Appendix at the end of the book.)

6.5 Sampling Distribution of a Statistic

DEFINITION

The **sampling distribution** of a statistic is the probability distribution of its values over all possible samples of a given size and design from the same population. The **Law of Large Numbers (LLN)**, first met for proportions in §5.1.2, states that, for independent observations from a population with mean μ, the sample mean $\bar{x}$ gets arbitrarily close to μ, with probability approaching 1, as n increases.

SIMPLIFIED EXPLANATION

Imagine drawing not one sample of 30 children but every possible sample of 30, and computing each sample's mean. Those means form a distribution of their own: not of weights, but of *means*. It has a centre, a spread and a shape, and they differ from the population's. The Derivation below proves two facts about it: it is centred on μ, and its spread shrinks as n grows. The LLN follows directly: a distribution centred on μ whose spread shrinks to zero must pile up at μ.

DERIVATION

To reason about a randomly drawn value we need two definitions for a random variable X (recall §5.6). Its **expected value** $E(X) = \Sigma\, x \times P(X = x)$ is its long-run average; for one unit drawn at random from a population, $E(X) = \mu$. Its **variance** $\mathrm{Var}(X) = E[(X - \mu)^2]$, the long-run average squared deviation; for a randomly drawn unit this is the population variance σ^2 (recall §3.6.1). Three rules follow.

1. **Constants scale.** $E(cX) = c \times E(X)$, and $\mathrm{Var}(cX) = c^2 \times \mathrm{Var}(X)$, because multiplying every value by c multiplies every deviation by c and every squared deviation by c^2.
2. **Means add.** $E(X + Y) = E(X) + E(Y)$: the long-run average of a sum is the sum of the long-run averages.

3. **Variances add for independent variables.** $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ when X and Y are independent. Expand: $\mathrm{Var}(X + Y) = E[((X - \mu_X) + (Y - \mu_Y))^2] = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \times E[(X - \mu_X)(Y - \mu_Y)]$. For independent X and Y, the multiplication rule (recall §5.2.2, §5.4) gives $P(X = x \text{ and } Y = y) = P(X = x) \times P(Y = y)$, so the last expectation factorises: $E[(X - \mu_X)(Y - \mu_Y)] = E(X - \mu_X) \times E(Y - \mu_Y) = 0 \times 0 = 0$. The cross term vanishes. (Without independence it does not, which is why clustered data, §6.2.4, have larger variances.)

Now let $X_1, \ldots, X_n$ be n independent draws from a population with mean μ and variance σ^2, and $\bar{x} = (X_1 + \ldots + X_n)/n$.

$E(\bar{x}) = (1/n) \times [E(X_1) + \ldots + E(X_n)] = (1/n) \times n\mu = \mu$

$\mathrm{Var}(\bar{x}) = (1/n)^2 \times [\mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)] = (1/n^2) \times n\sigma^2 = \sigma^2/n$

So the sample mean is centred exactly on μ (it is **unbiased** in the estimator sense of §6.4), and its variance is σ^2/n, which goes to 0 as n grows. That is the LLN. (A formal proof of the LLN uses Chebyshev's inequality and is beyond this book.)

ILLUSTRATION / EXAMPLE

A tiny population of four children weighs 10, 11, 12 and 15 kg. Its mean is $\mu = 12$ kg and its variance is $\sigma^2 = [(-2)^2 + (-1)^2 + 0^2 + 3^2] \div 4 = 14 \div 4 = 3.5$ kg^2. Draw samples of n = 2 *with replacement* (so the draws are independent).

WORKED CALCULATION

1. There are $4 \times 4 = 16$ equally likely ordered samples. Their means take the values 10 (1 sample), 10.5 (2), 11 (3), 11.5 (2), 12 (1), 12.5 (2), 13 (2), 13.5 (2) and 15 (1).
2. Average of the 16 sample means = 12.0 kg = μ. The sampling distribution is centred on μ.
3. Variance of the 16 sample means = 1.75 kg^2, which equals $\sigma^2/n = 3.5 \div 2 = 1.75$. The formula holds exactly.
4. Without replacement there are 6 possible pairs, with means 10.5, 11, 12.5, 11.5, 13 and 13.5. Their average is still 12.0, but their variance is 1.167 kg^2, smaller than 1.75, because the draws are no longer independent.

MUST-KNOW NOTES (MD-LEVEL)

- A sampling distribution describes variation *across repeated samples*, not within one sample. Do not confuse its spread with the SD of one sample's raw data.
- $E(\bar{x}) = \mu$ and $\mathrm{Var}(\bar{x}) = \sigma^2/n$ need only independence and a finite variance, not normality. Normality of the *shape* is the separate contribution of the CLT (§6.6).
- **Finite population correction.** When sampling without replacement, $\mathrm{Var}(\bar{x}) = (\sigma^2/n) \times (N - n)/(N - 1)$. In step 4: $1.75 \times (4 - 2)/(4 - 1) = 1.167$. When the sampling fraction n/N is small, the correction is close to 1 and is ignored; §7.3 uses it when the sampling fraction is large.
- **LLN in practice:** the relative frequency of an event is the mean of 0/1 values, so the LLN is why empirical probability (recall §5.1.2) settles on the true probability as trials increase. The LLN says $\bar{x}$ approaches μ; it says nothing about the shape of $\bar{x}$'s distribution. That is the CLT's job.

- In real research you draw one sample. The sampling distribution is a thought experiment, but its properties are what justify every confidence interval and test built from that single sample.

6.6 Central Limit Theorem

DEFINITION

The **Central Limit Theorem (CLT)** states that for independent observations from a population with mean μ and finite variance σ^2, the sampling distribution of $\bar{x}$ approaches a Normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ as n increases, whatever the shape of the population distribution. Equivalently, $(\bar{x} - \mu) \div (\sigma/\sqrt{n})$ approaches the standard Normal distribution (recall §5.7.2.2.1).

SIMPLIFIED EXPLANATION

§6.5 fixed the centre (μ) and spread ($\sigma/\sqrt{n}$) of the sampling distribution of $\bar{x}$. The CLT adds its shape: approximately Normal, even when individual values are skewed or bimodal. This is why the Normal distribution matters so much in the rest of the book. It is not that most raw variables are Normal (many are not); it is that *means* of almost anything become approximately Normal. The binomial's approach to a Normal shape for large n (recall §5.7.1.2) is one case of it, since a binomial count is a sum of independent 0/1 values.

ILLUSTRATION / EXAMPLE

Hospital length of stay is right-skewed (recall §3.9): most stays are short, a few are very long. Draw repeated samples of 40 patients and compute each sample's mean stay. The histogram of those means is roughly bell-shaped and centred on the population mean, although the individual stays are heavily skewed.

MUST-KNOW NOTES (MD-LEVEL)

- The CLT is about the sampling distribution of the *mean* (or a sum). It does **not** say that raw data become Normal in large samples.
- **How large is large enough? A convention, not a law:** $n \geq 30$ is widely used as the point where the Normal approximation for a mean is adequate (Kwak & Kim, 2017). The exception: strongly skewed or heavy-tailed populations need larger samples, and for a proportion the check is on the expected counts, np and n(1 − p) (recall §5.7.2.2.4), not on n alone.
- Independence matters. For clustered samples the CLT applies to cluster-level summaries, and the effective sample size (§6.2.6), not the raw count, governs how good the approximation is.

6.6.1 Implications of the Central Limit Theorem

DEFINITION

A statistical method is **robust to non-normality** when its stated error rates (for example, 95% coverage of a confidence interval) stay approximately correct even though the data are not Normal.

The CLT is what makes methods built on means robust in this sense when samples are large enough.

SIMPLIFIED EXPLANATION

Many methods in later chapters (the t-test, ANOVA, confidence intervals for means) are derived assuming Normal data. What they actually need is a Normal sampling distribution of the mean, and the CLT supplies that for moderately large n even when individual values are skewed.

ILLUSTRATION / EXAMPLE

Haemoglobin may be mildly skewed in a population. A 95% confidence interval for the *mean* haemoglobin, built on the Normal or t-distribution (recall §5.7.2.3; the interval itself is §7.2.1), is still valid for a sample of reasonable size, because the sampling distribution of the mean is approximately Normal even though individual readings are not.

MUST-KNOW NOTES (MD-LEVEL)

- **Proportions and sums:** a sample proportion is the mean of 0/1 values, so the CLT covers proportions too. This is the basis of the Normal-approximation confidence interval for a proportion (§7.2.2).
- **The limit of robustness:** with small samples from strongly skewed data, the Normal approximation can mislead. Non-parametric methods (Chapter 10) and exact or bootstrap intervals (§7.2.4) are the fallback.

6.7 Standard Error

DEFINITION

The **standard error (SE)** of a statistic is the standard deviation of its sampling distribution. For the mean of n independent observations, $SE(\bar{x}) = \sigma/\sqrt{n}$, which is estimated from one sample by $s/\sqrt{n}$. For a proportion, $SE(\hat{p}) = \sqrt{[p(1-p)/n]}$, estimated by $\sqrt{[\hat{p}(1-\hat{p})/n]}$.

SIMPLIFIED EXPLANATION

Recall §6.5: $Var(\bar{x}) = \sigma^2/n$. The SE is its square root, $\sigma/\sqrt{n}$: the typical distance between a sample mean and the population mean. You never know σ, so you plug in the sample SD, s. That is why a single sample is enough to estimate how much its own mean would wobble across repeated samples. For a proportion, recall that a 0/1 (Bernoulli) variable has variance $p(1-p)$ (§5.7.1.1); putting $\sigma^2 = p(1-p)$ into σ^2/n gives the SE of a proportion.

ILLUSTRATION / EXAMPLE

For the 30 children seen at the under-five clinic today (recall §1.4), the mean weight is 11.81 kg and s = 1.75 kg.

WORKED CALCULATION

1. Estimated SE = $s/\sqrt{n} = 1.75 \div \sqrt{30} = 1.75 \div 5.477 = 0.32$ kg.
2. Meaning: over repeated samples of 30 from the same population, sample means would typically lie about 0.32 kg from the true mean.
3. Two independent sample means differ from each other by more: the SE of their difference is $\sqrt{2} \times 0.32 \approx 0.45$ kg (rule 3 of §6.5: the variances add).
4. *Checking the EPI 30×7 design (recall §6.2.6).* Take the worst case $p = 0.5$, which makes $p(1 - p)$ largest. With $n = 210$ children at DEFF = 2, the effective sample size is $210 \div 2 = 105$. SE = $\sqrt{(0.5 \times 0.5 \div 105)} = 0.049$. A 95% margin is $1.96 \times 0.049 = 0.096$, about ±10 percentage points, as the design promises. The 1.96 is the z value that cuts off the central 95% of the Normal curve (recall §5.7.2.2.1); §7.2 builds confidence intervals on it.

MUST-KNOW NOTES (MD-LEVEL)

- **SE is not SD.** The SD describes the spread of individual values; the SE describes the spread of a statistic over repeated samples. $SE = SD/\sqrt{n}$ is always smaller for $n > 1$, so reporting "mean ± SE" as a description of the sample makes the data look far less variable than they are. In one audit of four anaesthesia journals, 198 of 860 articles (23%) used the SE inappropriately to describe sample variability (Nagele, 2003). Describe data with the SD; describe the precision of an estimate with the SE or a confidence interval.
- **SE shrinks with √n, not n.** Doubling n multiplies the SE by $1/\sqrt{2} \approx 0.71$; halving the SE needs four times the sample.
- For cluster samples the SRS formula understates the SE: multiply it by DEFT = √DEFF (§6.2.6). In the CBHI survey, 0.021 becomes 0.032 (0.0205 × 1.56 before rounding).
- The SE is the building block of confidence intervals (§7.2) and of most test statistics (Chapters 8–10).

CHECKPOINT 6.4

1. A colleague doubles her sample size and expects her standard error to halve. Is she right? What sample-size multiple would halve it?
2. If the clinic survey weighed 120 children instead of 30 and the SD stayed at 1.75 kg, what would the estimated SE of the mean be?
3. Using the rule "variances of independent variables add", show that $\text{Var}(\bar{x}) = \sigma^2/n$. Then state what the Law of Large Numbers adds, and what the Central Limit Theorem adds beyond it.

(Answers in the Appendix at the end of the book.)

Estimation

Chapter 6 built the machinery that describes how a statistic varies from sample to sample: the sampling distribution, the Central Limit Theorem and the standard error. This chapter puts it to work on the one sample you actually have. It covers the single best guess (point estimation, §7.1), the range of plausible values around it (confidence intervals, §7.2), how large a sample must be for a chosen precision (§7.3), and the rates, standardised rates and life tables through which Indian health statistics are reported (§7.4–7.5).

7.1 Point Estimation

DEFINITION

Point estimation uses a single sample statistic as the best guess for an unknown population parameter (recall §1.3): the sample mean $\bar{x}$ for the population mean μ, the sample proportion $\hat{p}$ for the population proportion p.

SIMPLIFIED EXPLANATION

Calculate the statistic from your sample and offer it as your estimate. It is direct and easy to report. A single number, however, says nothing about how far from the truth it might be.

ILLUSTRATION / EXAMPLE

Today's 30 children at the under-five clinic have a mean weight of 11.81 kg. The point estimate of the mean weight of all children who attend the clinic is 11.81 kg. Nothing in that one number tells you whether the true mean could be 11.5 kg or 13 kg.

MUST-KNOW NOTES (MD-LEVEL)

- A point estimate is one number computed from the sample. It carries no statement of its precision. Confidence intervals (§7.2) add that.
- Not every way of turning a sample into a number is equally trustworthy. §7.1.1 gives the criteria.

7.1.1 Properties of a Good Estimator

DEFINITION

An **estimator** is the rule (formula) that turns a sample into an estimate. A good estimator is judged on three properties. It is **unbiased** if its average over all possible samples equals the parameter. It is more **efficient** than a rival if its sampling variance is smaller. It is **consistent** if it converges on the parameter as n grows.

SIMPLIFIED EXPLANATION

Think of the estimator as a marksman firing once per sample. Unbiased means the shots centre on the target. Efficient means they cluster tightly. Consistent means that, as the sample grows, the shots close in on the target.

ILLUSTRATION / EXAMPLE

Recall Sample Variance and Bessel's Correction (§3.6.2). For today's 30 weights, dividing the sum of squared deviations by n − 1 gives $s^2 = 3.07\ \text{kg}^2$; dividing by n gives $3.07 \times 29/30 = 2.97\ \text{kg}^2$. Averaged over all possible samples, the divide-by-n version falls short of σ^2 by the factor $(n-1)/n$, because the expected sum of squares is $(n-1)\sigma^2$ (derived in §3.6.2). The n − 1 divisor removes that shortfall: it is Bessel's correction, applied to make s^2 unbiased.

MUST-KNOW NOTES (MD-LEVEL)

- **Unbiasedness.** E(estimator) = parameter. $\bar{x}$ is unbiased for μ (recall §6.5) and s^2 is unbiased for σ^2.
- **Efficiency.** For normally distributed data both the sample mean and the sample median are unbiased for μ, but the mean has the smaller sampling variance (the median's is about $1.57 = \pi/2$ times larger in large samples, because their asymptotic relative efficiency is $0.637 = 2/\pi$; proof beyond this book; Shao, n.d.), so the mean is the more efficient estimator.
- **Consistency.** The divide-by-n variance is biased, but its bias, σ^2/n, shrinks to zero as n grows, and so does its sampling variance, so it is consistent. A small-sample bias that vanishes with n can be acceptable.
- **Two meanings of "bias".** *Estimator bias*, as here, is a property of a formula and often shrinks as n grows. *Design bias* (selection or measurement bias, recall §6.4) is a flaw in how the data were collected, and no increase in n removes it. A survey of 5,000 households that misses every migrant household is precise and still wrong.

CHECKPOINT 7.1

1. Explain why Bessel's correction (dividing by n − 1) is an application of the unbiasedness property.
2. For today's 30 weights, $s^2 = 3.07\ \text{kg}^2$. What is the divide-by-n variance, and in which direction is it biased?
3. A point estimate alone is called incomplete. What is missing from it?

(Answers in the Appendix at the end of the book.)

7.2 Confidence Intervals

DEFINITION

A **confidence interval (CI)** is a range calculated from a sample by a procedure that, over repeated samples, captures the true parameter a stated proportion of the time: the **confidence level**, conventionally 95%. Its general form is: point estimate ± (critical value × standard error).

SIMPLIFIED EXPLANATION

The standard error (recall §6.7) says how far a statistic typically lands from the parameter. A CI turns that into a range: go a set number of standard errors either side of the estimate. The confidence level describes the procedure, not the particular interval you computed.

DERIVATION

Where the formula and the number 1.96 come from, for a mean with known σ:

1. By the CLT (recall §6.6), $\bar{x}$ is approximately Normal with mean μ and standard error $\sigma/\sqrt{n}$. So $Z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$ approximately follows the standard Normal distribution. Its distribution involves no unknown parameter; a quantity like this is called a **pivot**.
2. The value that cuts off 2.5% in each tail of the standard Normal is 1.96, because $P(Z \le 1.96) = 0.975$. Hence $P(-1.96 \le Z \le 1.96) = 0.95$.
3. Substitute Z: $P(-1.96 \le (\bar{x} - \mu)/(\sigma/\sqrt{n}) \le 1.96) = 0.95$.
4. Multiply through by $\sigma/\sqrt{n}$ and rearrange so that μ stands alone in the middle: $P(\bar{x} - 1.96\sigma/\sqrt{n} \le \mu \le \bar{x} + 1.96\sigma/\sqrt{n}) = 0.95$.

The random quantity in step 4 is the interval, whose ends depend on $\bar{x}$. μ is fixed. The same rearrangement with other cut-offs gives 1.645 for 90% and 2.576 for 99%.

Link to the ±2 SD rule. The empirical rule (recall §5.7.2.2.2) says about 95% of a Normal distribution lies within ±2 SD. The exact figure for ±2 SD is 95.45%. The exact multiplier for 95% is 1.96, which rounds to 2. "Estimate ± 2 SE" is therefore the empirical rule applied to the sampling distribution, and 1.96 is its precise version.

ILLUSTRATION / EXAMPLE

Figure 7.1 draws 20 samples of 30 weights from a population whose mean is known to be 11.81 kg, and builds a 95% interval from each. Nineteen of the 20 intervals contain 11.81 kg; one misses. In real research you see only one interval and cannot tell which kind it is.

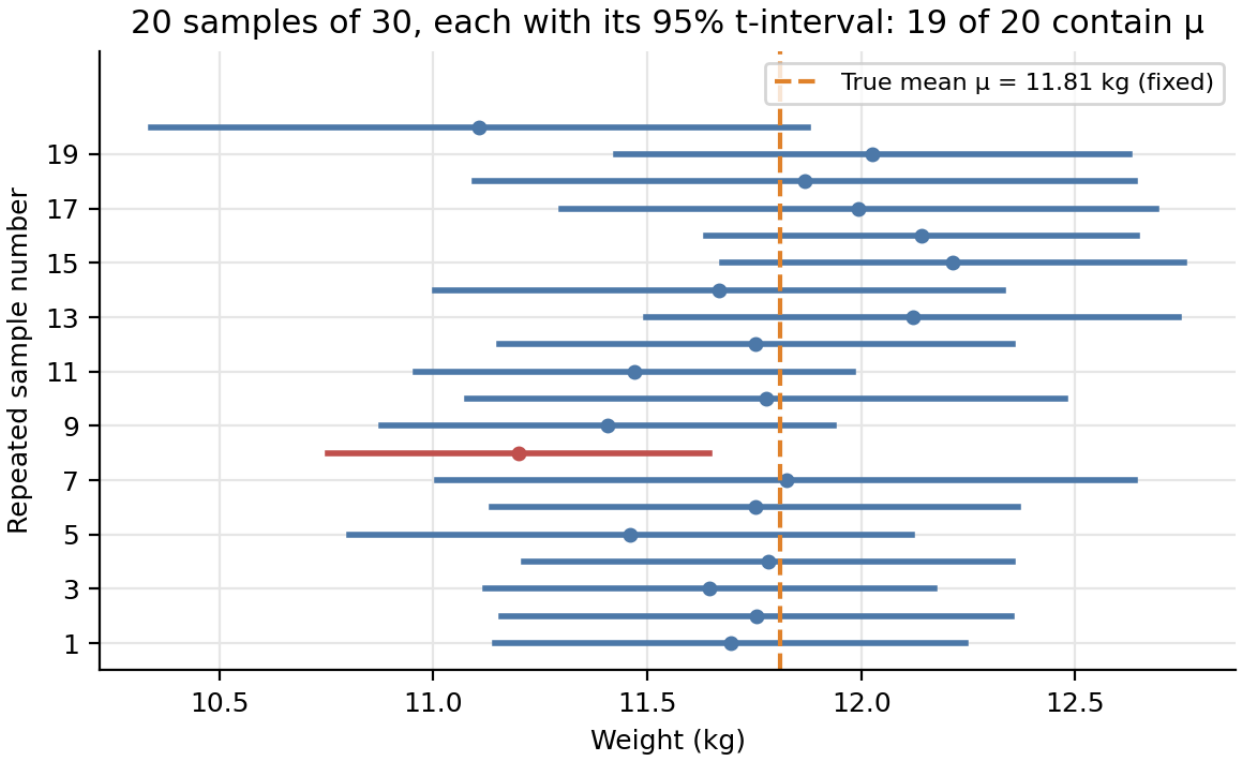


Figure 7.1. Twenty samples of 30 drawn from a population with mean 11.81 kg and SD 1.75 kg. Each line is one sample's 95% t-interval; the red interval misses the true mean. The mean is fixed; the intervals vary.

MUST-KNOW NOTES (MD-LEVEL)

- **Correct interpretation.** A 95% CI does *not* mean "there is a 95% probability that μ lies in this interval". Once computed, the interval either contains μ or it does not. The 95% is the long-run success rate of the procedure (Figure 7.1).
- **Width.** The interval widens with a higher confidence level (1.96 → 2.576 for 99%), with more variable data (larger s) and with a smaller sample (SE ∝ 1/√n: SE is proportional to 1/√n). §7.3 uses this to plan sample size.
- **Which critical value.** Use the t-distribution (recall §5.7.2.3) whenever σ is estimated by s, whatever the sample size. As n grows, t approaches z (t_{29} = 2.045; t_{100} = 1.984), so z is a large-sample approximation, not a separate rule.

7.2.1 CI for a Mean

DEFINITION

A **confidence interval for a mean** is $\bar{x} \pm t(n-1) \times s/\sqrt{n}$, where t(n−1) is the t-value that leaves 2.5% in each tail with n − 1 degrees of freedom.

SIMPLIFIED EXPLANATION

When σ is unknown, which is nearly always, replacing it by s adds uncertainty. The pivot becomes $(\bar{x} - \mu)/(s/\sqrt{n})$, which follows a t-distribution with $n - 1$ degrees of freedom rather than the standard Normal. Everything else in the derivation of §7.2 is unchanged, with 1.96 replaced by the t-value.

WORKED CALCULATION

Today's 30 children: $\bar{x} = 11.81$ kg, $s = 1.75$ kg.

1. $SE = s/\sqrt{n} = 1.75/\sqrt{30} = 0.32$ kg.
2. With 29 degrees of freedom, $t_{29} = 2.045$.
3. Margin $= 2.045 \times 0.32 = 0.65$ kg.
4. 95% CI $= 11.81 \pm 0.65 =$ **11.16 to 12.47 kg**.
5. For comparison, the z-approximation gives $11.81 \pm 1.96 \times 0.32 = 11.81 \pm 0.63 = 11.19$ to 12.44 kg, about 0.03 kg narrower at each end. Here, where σ is estimated from 30 children, the t-interval is the correct one.

IN R (OPTIONAL)

```
cl ← read.csv("data/clinic_children.csv")
w  ← cl$weight_kg[cl$seen_today == 1]      # today's 30 children
t.test(w)$conf.int                          # t-based 95% CI (29 df)
# z-based, for comparison
mean(w) + c(-1, 1) * qnorm(0.975) * sd(w) / sqrt(length(w))
```

```
[1] 11.15935 12.46732
attr(,"conf.level")
[1] 0.95
[1] 11.18661 12.44005
```

The first line is the t-interval of step 4 (11.16 to 12.47 kg); `t.test` uses s and 29 degrees of freedom automatically. The second is the z-approximation of step 5, slightly narrower because $1.96 < 2.045$.

MUST-KNOW NOTES (MD-LEVEL)

- Formula: $\bar{x} \pm t(n-1) \times s/\sqrt{n}$. The t-interval assumes the observations are independent and, for small n, roughly Normal. For larger n the CLT protects it against moderate skew (recall §6.6.1).
- A larger sample shrinks $s/\sqrt{n}$ and, through the degrees of freedom, the t-value too.

7.2.2 CI for a Proportion

DEFINITION

The **Wald confidence interval for a proportion** is $\hat{p} \pm 1.96 \times \sqrt{[\hat{p}(1 - \hat{p})/n]}$, where $\hat{p} = x/n$ is the sample proportion.

SIMPLIFIED EXPLANATION

A proportion is the mean of 0/1 values (recall §5.7.1.1, Bernoulli), so the CLT applies to it too. The variance of a single 0/1 value is $p(1 - p)$, so the standard error of $\hat{p}$ is $\sqrt{[p(1 - p)/n]}$, estimated by putting $\hat{p}$ in place of p. The pivot argument of §7.2 then gives the interval.

WORKED CALCULATION

Sixty of 200 surveyed households are CBHI members.

1. $\hat{p} = 60/200 = 0.30$.
2. SE $= \sqrt{(0.30 \times 0.70/200)} = \sqrt{0.00105} = 0.032$.
3. Margin $= 1.96 \times 0.032 = 0.063$.
4. 95% CI $= 0.30 \pm 0.063 =$ **23.6% to 36.4%**.

IN R (OPTIONAL)

The full CBHI survey has 150 members among 500 households.

```
cb ← read.csv("data/cbhi_survey.csv")
x ← sum(cb$cbhi_member); n ← nrow(cb)
p ← x / n
c(p = p, se = sqrt(p * (1 - p) / n))
p + c(-1, 1) * qnorm(0.975) * sqrt(p * (1 - p) / n)   # Wald 95% CI
```

```
        p        se
0.3000000 0.0204939
[1] 0.2598327 0.3401673
```

The same 30% with n = 500 instead of 200 gives SE 0.0205 (the 0.021 of §6.2.6) and a CI of 26.0% to 34.0%, narrower by the factor $\sqrt{(200/500)} = 0.63$. This calculation treats the 500 households as a simple random sample. They were in fact drawn in 20 village clusters, so the true SE is larger: 0.032 (recall §6.2.6: clustering inflates the variance by the factor DEFF; §7.3 uses it).

MUST-KNOW NOTES (MD-LEVEL)

- Formula: $\hat{p} \pm 1.96 \times \sqrt{[\hat{p}(1 - \hat{p})/n]}$.
- The Wald interval performs poorly when n is small or $\hat{p}$ is near 0 or 1. It can even run below 0% (§7.2.4). One rule of thumb asks for $n\hat{p}$ and $n(1 - \hat{p})$ both at least 5 before using it. That is a convention, not a guarantee: simple Wald intervals can cover the true value less often than the stated 95% (Newcombe, 1998). The Wilson and exact (Clopper–Pearson) intervals of §7.2.4 avoid these problems.
- Every prevalence, coverage or success rate in a report should carry its CI.

7.2.3 CI for a Difference Between Groups

DEFINITION

A **confidence interval for a difference** between two independent groups is (estimate$_1$ − estimate$_2$) ± 1.96 × SE_diff. For two means, SE_diff = $\sqrt{(s_1^2/n_1 + s_2^2/n_2)}$. For two proportions, SE_diff = $\sqrt{[p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2]}$.

SIMPLIFIED EXPLANATION

Research questions usually compare groups. Each group's estimate has its own sampling error, and the difference inherits both, so the difference is less precise than either estimate alone. The rule is that **variances add, standard errors do not**: square each group's SE, add, then take the square root.

DERIVATION

Write the deviations of the two sample means from their targets as $a = \bar{x}_1 - \mu_1$ and $b = \bar{x}_2 - \mu_2$. The deviation of the difference is a − b, and its variance is the average of $(a-b)^2$:

$E[(a-b)^2] = E[a^2] + E[b^2] - 2E[ab]$.

The samples are independent (recall §5.4), so $E[ab] = E[a] \times E[b] = 0 \times 0 = 0$. Hence $Var(\bar{x}_1 - \bar{x}_2) = SE_1^2 + SE_2^2$, and SE_diff = $\sqrt{(SE_1^2 + SE_2^2)} = \sqrt{(s_1^2/n_1 + s_2^2/n_2)}$.

WORKED CALCULATION

Two means (hypothetical). Urban children: $n_1 = 40$, mean 12 kg, SD 2.0 kg. Rural children: $n_2 = 50$, mean 11 kg, SD 1.8 kg.

1. $SE_1 = 2.0/\sqrt{40} = 0.316$; $SE_2 = 1.8/\sqrt{50} = 0.255$.
2. SE_diff = $\sqrt{(0.316^2 + 0.255^2)} = \sqrt{(0.100 + 0.065)} = \sqrt{0.165} = 0.406$ kg. This is larger than either SE but smaller than their sum, 0.571.
3. 95% CI = 1 ± 1.96 × 0.406 = 1 ± 0.80 = **0.20 to 1.80 kg**.

Had the interval been, say, −0.2 to 2.2 kg, a true difference of zero would have remained compatible with the data.

Two proportions (CBHI survey). CBHI membership is 69/155 = 44.5% among households where the mother has secondary or higher education, and 31/168 = 18.5% where she has none.

1. Difference = 0.445 − 0.185 = 0.261.
2. SE_diff = $\sqrt{(0.445 \times 0.555/155 + 0.185 \times 0.815/168)} = \sqrt{(0.00159 + 0.00090)} = 0.050$.
3. 95% CI = 0.261 ± 1.96 × 0.050 = **16.3 to 35.8 percentage points**.

IN R (OPTIONAL)

```
cb ← read.csv("data/cbhi_survey.csv")
tab ← table(cb$maternal_education, cb$cbhi_member)
tab
prop.test(x = tab[c("secondary+", "none"), "1"],
          n = rowSums(tab)[c("secondary+", "none")], correct = FALSE)
```

```
              0   1
  none      137  31
  primary   127  50
  secondary+  86  69
...
95 percent confidence interval:
 0.1628515 0.3584235
sample estimates:
   prop 1    prop 2
0.4451613 0.1845238
```

The table gives the counts used above (members in column `1`). `prop 1` and `prop 2` are the two proportions, and the CI (0.163 to 0.358) is the hand calculation. The trimmed lines report a test statistic and p-value, which §8.5 explains. As in §7.2.2, this SE ignores clustering by village; §15.2 (analysis of clustered data) shows the correction.

MUST-KNOW NOTES (MD-LEVEL)

- **Null values.** For a difference, "no effect" is 0. For a ratio (odds ratio, risk ratio, rate ratio), "no effect" is 1, so check whether the CI includes 1.
- **Reading an interval that excludes the null.** If a 95% CI for a difference excludes 0, the data are incompatible with a zero difference at the 5% level. This is not the probability that the result is "due to chance". §8.1 (the logic of hypothesis testing) formalises the link.
- **Paired data.** The formula assumes independent groups. For before–after measurements on the same people, analyse the within-person differences as one sample (§9.2.3, the paired t-test).
- For small samples of means, replace 1.96 by a t-value; §9.2.2 (the independent-samples t-test) gives the Welch degrees of freedom.

CHECKPOINT 7.2

1. A colleague says, "There is a 95% probability that the true mean lies in this interval." Correct the statement.
2. For today's 30 children ($\bar{x}$ = 11.81 kg, SE = 0.32 kg, t_{29} for 99% = 2.756), compute the 99% CI. Is it wider or narrower than the 95% CI, and why?
3. Two independent groups of 50 each have SD 2 kg and means differing by 0.8 kg. Compute SE_diff and the 95% CI for the difference. Does it include 0?

(Answers in the Appendix at the end of the book.)

7.2.4 Small-Sample and Alternative Intervals: Exact, Wilson, Median and Bootstrap

DEFINITION

When the Normal approximation behind the Wald interval fails, other intervals are used. For a proportion: the **Wilson (score) interval** and the **Clopper–Pearson (exact) interval**, and, when no events are seen, the **rule of three**. For a median: an **order-statistic interval**. For almost any statistic: the **percentile bootstrap interval**.

SIMPLIFIED EXPLANATION

The Wald interval places a symmetric band around $\hat{p}$ using an SE estimated from $\hat{p}$ itself. Near 0 or 1 that SE is too small and the band can cross 0. The alternatives avoid the shortcut. Wilson inverts the test, asking which values of p would make the observed $\hat{p}$ unsurprising. Clopper–Pearson uses exact binomial probabilities. The bootstrap builds the sampling distribution by resampling the data you have.

DERIVATION

Wilson. Keep the unknown p inside the SE and ask for which p the observed $\hat{p}$ lies within 1.96 SEs: $(\hat{p} - p)^2 \le 1.96^2\, p(1-p)/n$. Solving this quadratic in p gives

$$\text{centre} = (x + z^2/2)/(n + z^2),\ \text{half-width} = [z/(n+z^2)] \times \sqrt{[x(n-x)/n + z^2/4]},$$

with $z = 1.96$. The centre is pulled from $\hat{p}$ towards 0.5, and the limits always stay inside 0 to 1.

Clopper–Pearson. The lower limit is the p at which seeing x or more events has probability 0.025; the upper limit is the p at which seeing x or fewer has probability 0.025. Both use the binomial distribution (recall §5.7.1.2) directly, so no approximation is involved. The interval is conservative: its coverage is at least 95%, and often more.

Rule of three. If 0 events occur in n independent trials, the one-sided 95% upper limit is the p for which $(1-p)^n = 0.05$. Then $n \times \ln(1-p) = \ln 0.05 = -3.00$, and for small p, $\ln(1-p) \approx -p$, so $p \approx 3/n$.

Median. Each observation falls below the population median with probability 0.5, so the number of observations below it follows Binomial(n, 0.5). Choose the largest rank r for which $P(X \le r-1) \le 0.025$. The interval from the r-th to the $(n+1-r)$-th ordered value then covers the median with probability at least 95%. No distributional assumption about the data is needed.

Bootstrap. The sampling distribution (recall §6.5) is what you would see if you could draw many samples from the population. The bootstrap uses the sample as a stand-in for the population: draw n values *with replacement* from the n observations, compute the statistic, repeat 1,000 or more times, and read off the 2.5th and 97.5th centiles of the results (Efron, 1979).

WORKED CALCULATION

Three adverse events following immunisation (AEFI) among 40 infants. An AEFI is any untoward medical event after a vaccination. At one PHC session, 3 of 40 infants had a minor AEFI, and the infants' outcomes are independent.

1. $\hat{p} = 3/40 = 0.075$; $n\hat{p} = 3$, below the rule-of-thumb 5.

2. Wald: SE = √(0.075 × 0.925/40) = 0.042; 0.075 ± 1.96 × 0.042 = −0.7% to 15.7%. The lower limit is impossible.
3. Wilson: centre = (3 + 1.92)/43.84 = 0.112; half-width = (1.96/43.84) × √(3 × 37/40 + 0.96) = 0.0447 × 1.933 = 0.086. CI = **2.6% to 19.9%**.
4. Clopper–Pearson: **1.6% to 20.4%**, a little wider, as expected of an exact method.

No serious AEFI among 150 infants. The rule of three gives an upper 95% limit of 3/150 = **2%**. The exact one-sided limit, 1 − 0.05^(1/150), is 1.98%. "None observed" is compatible with a true risk as high as 2 in 100 (Eypasch et al., 1995, after Hanley & Lippman-Hand, 1983).

95% CI for the median weight of today's 30 children (median 11.7 kg).

1. For Binomial(30, 0.5), P(X ≤ 9) = 0.021 ≤ 0.025, while P(X ≤ 10) = 0.049. So r = 10 and n + 1 − r = 21.
2. The 10th and 21st of the sorted weights are 11.0 kg and 12.6 kg. CI = **11.0 to 12.6 kg**.
3. Actual coverage = 1 − 2 × 0.021 = 95.7%.

Bootstrap. 2,000 resamples of the 30 weights give a percentile interval for the median of **11.0 to 12.6 kg** (Figure 7.2), the same as the order-statistic interval. The bootstrap SE of the mean from the same approach is 0.31 kg, close to s/√n = 0.32 kg: a check that the resampling reproduces the formula where a formula exists.

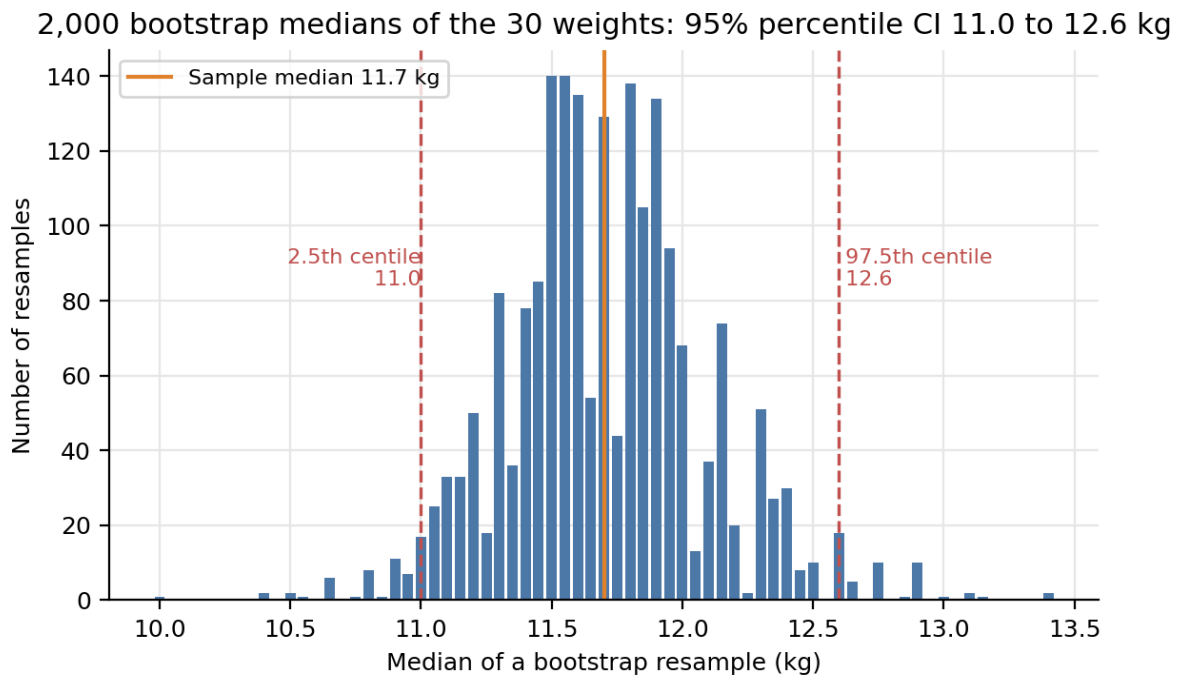


Figure 7.2. Medians of 2,000 bootstrap resamples of today's 30 weights. The 2.5th and 97.5th centiles (11.0 and 12.6 kg) form the percentile 95% CI. The spikes reflect that a median of 30 values can take only a few distinct values.

IN R (OPTIONAL)

```
binom.test(3, 40)$conf.int                  # Clopper-Pearson (exact)
prop.test(3, 40, correct = FALSE)$conf.int  # Wilson score
library(boot)
cl ← read.csv("data/clinic_children.csv")
w  ← cl$weight_kg[cl$seen_today == 1]
set.seed(2026)
b ← boot(w, function(d, i) median(d[i]), R = 2000)
b
boot.ci(b, type = "perc")
```

```
[1] 0.01574218 0.20386475
...
[1] 0.02583603 0.19864234
...
Bootstrap Statistics :
    original   bias    std. error
t1*     11.7 0.005625   0.3855275
...
Intervals :
Level     Percentile
95%   (11.0, 12.6 )
```

`binom.test` returns the exact interval (1.6% to 20.4%). `prop.test` without continuity correction returns the Wilson interval (2.6% to 19.9%). In the bootstrap output, `original` is the sample median (11.7 kg) and `std. error` (0.39 kg) is the SD of the 2,000 resampled medians, the bootstrap estimate of the median's SE. The percentile interval matches steps 1–2 above. `set.seed` fixes the random resampling so the result can be reproduced.

MUST-KNOW NOTES (MD-LEVEL)

- **Which interval for a proportion.** Newcombe (1998) compared seven methods and recommended the score (Wilson) and tail-area (exact) methods over the simple Wald interval. Report which method you used.
- **Rule of three.** Upper 95% limit ≈ 3/n when 0 of n have the event. Useful for safety claims ("no complications in 60 procedures" still allows up to 5%).
- **Median CI** uses only the ranks, so it suits skewed data such as hospital stay. Its coverage is at least, not exactly, 95%.
- **Bootstrap** needs no formula for the SE, which makes it useful for medians, ratios and model-based quantities (§15.5.2 uses it to validate prediction models). It cannot rescue a biased or tiny sample: with n = 10 the resamples can only recombine 10 values.

CHECKPOINT 7.3

1. Two of 25 children in an anganwadi test positive for a condition. Compute the Wald 95% CI and state what is wrong with it. Which intervals would you report instead?
2. A surgeon reports "no wound infections in 60 caesarean sections". Using the rule of three, what infection risk is still compatible with this record?

3. For n = 50, $P(X \le 17) = 0.016$ and $P(X \le 18) = 0.032$ for $X \sim \text{Binomial}(50, 0.5)$. Which ordered values form the 95% CI for the median?

(Answers in the Appendix at the end of the book.)

7.3 Margin of Error and Sample Size Estimation

DEFINITION

The **margin of error (d)** is the "±" part of a CI: critical value × standard error. **Sample size estimation** reverses the CI calculation: fix the margin you can accept, the confidence level and a guess at the variability, and solve for n. For a proportion, $n = Z^2p(1 - p)/d^2$. For a mean, $n = Z^2\sigma^2/d^2$.

SIMPLIFIED EXPLANATION

Before a survey you decide how precise the answer must be, for example "membership prevalence within ±5 percentage points". The CI formula links that precision to n, so you can solve for n. The simple formula assumes a simple random sample with everyone responding. Real surveys are clustered and some households do not respond, so the basic n is then inflated in steps (Arya et al., 2012).

DERIVATION

Proportion, absolute precision. The margin is $d = Z\sqrt{[p(1 - p)/n]}$. Square both sides: $d^2 = Z^2p(1 - p)/n$. Solve for n:

$$n = Z^2\, p(1 - p)/d^2$$

With Z = 1.96 rounded to 2, $Z^2 = 4$, and with $q = 1 - p$ and L the allowable error, this becomes the shortcut $\mathbf{n = 4pq/L^2}$.

Why p = 0.5 is the safe guess. $p(1 - p) = p - p^2 = 0.25 - (p - 0.5)^2$. The squared term is never negative, so $p(1 - p)$ is at most 0.25, reached at p = 0.5. With no prior estimate, p = 0.5 gives the largest, safest n.

Relative precision. If the margin is set as a fraction ε of p (for example ±20% of p), then $d = \varepsilon p$ and $n = Z^2p(1 - p)/(\varepsilon p)^2 = \mathbf{Z^2q/(\varepsilon^2 p)}$. Rare conditions need relative precision: ±5 points around a prevalence of 3% would be useless.

Mean. $d = Z\sigma/\sqrt{n}$ gives $\mathbf{n = Z^2\sigma^2/d^2}$, with σ from a pilot study or earlier work.

Halving the margin. $n \propto 1/d^2$, so halving d multiplies n by four.

Finite population correction (FPC). Sampling without replacement from a small population of size N reduces the variance of the estimate by the factor $(N - n)/(N - 1) \approx 1 - n/N$ (recall the finite population correction checked in §6.5). Setting the reduced variance equal to that of the uncorrected n_0 gives $1/n - 1/N = 1/n_0$, so $\mathbf{n = n_0/(1 + n_0/N)}$. It matters only when n_0 is more than a few per cent of N.

Design effect. A cluster sample of n gives the precision of a simple random sample of n/DEFF (recall §6.2.6, where $DEFF = 1 + (m - 1)\rho$ was previewed, m being the households per cluster and ρ the intra-cluster correlation). To keep the precision, **multiply n by DEFF**.

Non-response. If only a fraction r of those approached respond, approach **n/r**.

WORKED CALCULATION

Planning a repeat CBHI survey in the district: expected membership p = 0.30, margin ±5 percentage points, 95% confidence, village clusters of 25 households with DEFF 2.44 (m = 25, ρ = 0.06), and 90% expected response.

1. Simple random sample: $n_0 = 1.96^2 \times 0.30 \times 0.70/0.05^2 = 3.8416 \times 0.21/0.0025 =$ **322.7 → 323 households**.
2. Shortcut: $4pq/L^2 = 4 \times 0.21/0.0025 = 336$. The small excess comes from rounding 1.96 up to 2.
3. Without a prior estimate (p = 0.5): $1.96^2 \times 0.25/0.0025 = 384.2 \rightarrow 385$ (or 400 by the shortcut).
4. Design effect: 322.7 × 2.44 = **787.4**.
5. Non-response: 787.4/0.90 = **874.9 → 875 households**, that is, 35 villages of 25.
6. Relative precision instead: ±20% of 30% means d = 0.06, and $n_0 = 3.8416 \times 0.70/(0.04 \times 0.30) = 224.1 \rightarrow 225$ before the design effect.
7. FPC, only if the frame were small: if the survey covered a block of just N = 2,500 households, n_0 would fall to $322.7/(1 + 322.7/2{,}500) = 285.8 \rightarrow 286$ before the design effect.

Check against the survey actually done. The CBHI survey interviewed 500 households in 20 villages. Its effective sample size is 500/2.44 = 205, so its margin is $1.96 \times \sqrt{(0.21/205)} = \pm 6.3$ points, not the ±4.0 points a simple random sample of 500 would give. That matches the clustered SE of 0.032 (recall §6.2.6; 1.96 × 0.032 = 0.063). This is why step 5 needs 875 households, not 323, for ±5 points.

A mean. To estimate the mean weight of clinic children within ±0.5 kg, taking σ = 1.75 kg from today's 30: $n = 1.96^2 \times 1.75^2/0.5^2 = 3.8416 \times 3.0625/0.25 = 47.1 \rightarrow$ **48 children**. For ±0.25 kg, $n = 188.2 \rightarrow 189$, four times as many.

IN R (OPTIONAL)

```
z ← qnorm(0.975); p ← 0.30; d ← 0.05
n0 ← z^2 * p * (1 - p) / d^2          # simple random sample
n1 ← n0 * 2.44                          # x design effect
n2 ← n1 / 0.90                          # / expected response
round(c(n0 = n0, with_deff = n1, with_nonresponse = n2), 1)
ceiling(n2 / 25)                        # villages of 25 households
```

```
             n0        with_deff with_nonresponse
          322.7            787.3            874.8
[1] 35
```

The three numbers are steps 1, 4 and 5. They differ from the hand values in the first decimal only because R uses z = 1.959964 rather than 1.96. Always round the final n *up*: 874.8 households means 875, and 35 villages.

MUST-KNOW NOTES (MD-LEVEL)

- **Choosing d.** There is no universal rule. One convention is ±5 points when the expected prevalence lies between 10% and 90%, and a margin of about one-quarter to one-fifth of p when p is small (Pourhoseingholi et al., 2013). State and justify your choice.
- **Order of adjustments.** n_0 → FPC (if the frame is small) → × DEFF → ÷ response rate → round up to whole clusters.
- **What the inflations do not do.** Dividing by the response rate keeps the number of respondents up. It does not remove non-response bias if non-responders differ from responders (recall §6.4).
- **Precision, not power.** These formulas give a sample big enough to *estimate* one quantity precisely. Samples for *comparing* groups need power calculations (power: the chance that a study detects a difference that really exists; §15.6, where n depends on the smallest difference worth detecting).
- **A planning guess, not a fact.** The n is only as good as the guesses of p, σ and DEFF. When a guess is uncertain, compute n for a range of values and use the largest affordable.

CHECKPOINT 7.4

1. You want to halve the margin of error of a planned prevalence survey. Roughly how much larger must the sample be, and why?
2. Expected anaemia prevalence among adolescent girls is 10%, with a margin of ±3 percentage points at 95% confidence. Compute n_0. Then adjust for a cluster design with DEFF = 2 and 85% expected response.
3. For the same 10%, what n_0 is needed for a relative precision of ±20%? Why is it larger than in question 2?

(Answers in the Appendix at the end of the book.)

7.4 Rates, Ratios and Standardisation

Most Indian health statistics, from the SRS infant mortality rate to a district's crude death rate, are rates or ratios, and comparing them across places is fair only after adjusting for age structure. This section defines the building blocks and the core indicators (§7.4.1), shows the two ways of removing the effect of age, direct (§7.4.2) and indirect (§7.4.3), and attaches confidence intervals to rates, rate ratios and SMRs (§7.4.4). It builds on the one-line definitions of prevalence, risk and rate (recall §1.5) and uses the weighted mean (recall §2.2.2), the Poisson distribution (recall §5.7.1.3) and the CIs of §7.2.

7.4.1 Ratios, Proportions, Rates and Health Indicators

DEFINITION

- A **ratio** is one quantity divided by another; the numerator need not be part of the denominator (e.g. doctors per 10,000 population).
- A **proportion** is a ratio whose numerator is part of its denominator (CDC, 2012a).

- A **rate** is the frequency of an event in a defined population over a specified period (CDC, 2012a). Its denominator is **person-time**: the total time at risk, summed across people. For a whole population over one year, person-time is estimated by the **mid-year population** × 1 year (CDC, 2012b).
- **Crude** rates use the whole population. **Specific** rates restrict numerator and denominator to one age group, sex or cause. **Proportional (proportionate) mortality** is deaths from one cause divided by all deaths, a proportion, not a rate (CDC, 2012b).
- **Incidence** counts new cases in a period; **prevalence** counts all existing cases at a point or during a period (CDC, 2012c).

SIMPLIFIED EXPLANATION

A count means little until you know how many people, over how long, produced it. A rate adds the person-time, so districts of different sizes can be compared. The indicators you meet in SRS reports and NFHS fact sheets are all numerator/denominator pairs, and many exam questions on them turn on getting the denominator right: live births, total births or the mid-year population.

Prevalence, incidence and duration. Prevalence depends on both incidence and the duration of illness (CDC, 2012c). A simple steady-state argument shows how.

DERIVATION

In a stable population of size N with C current cases, new cases arise at rate I among the N − C people without the disease, and cases end (by cure or death) after an average duration D, so about C/D end each year. If prevalence is steady, inflow equals outflow:

$I \times (N - C) = C/D \Rightarrow C/(N - C) = I \times D.$

$C/(N - C)$ is $P/(1 - P)$, where $P = C/N$. When P is small, $1 - P \approx 1$, so $\mathbf{P \approx I \times D}$. Example: incidence 2 per 1,000 per year and mean duration 5 years give $P \approx 0.002 \times 5 = 1\%$ (the exact steady-state value is 0.99%).

ILLUSTRATION / EXAMPLE

The core indicators, with the definitions used by WHO and CDC:

Indicator	Numerator	Denominator	Per
Crude birth rate (CBR)	Live births in the year	Mid-year population	1,000
Crude death rate (CDR)	Deaths in the year	Mid-year population	1,000
Infant mortality rate (IMR)	Deaths under 1 year of age	Live births in the same year	1,000 live births
Neonatal mortality rate (NMR)	Deaths in the first 28 completed days	Live births	1,000 live births
Perinatal mortality rate (PMR)	Stillbirths (≥ 28 weeks) + early neonatal deaths (days 0–6)	Total births (live births + stillbirths)	1,000 births
Maternal mortality ratio (MMR)	Maternal deaths (pregnancy or within 42 days of its end, from related causes)	Live births	100,000 live births

Indicator	Numerator	Denominator	Per
Total fertility rate (TFR)	Sum of age-specific fertility rates, 15–49 (× 5 for 5-year groups)	(children per woman)	—
Case fatality rate (CFR)	Deaths from the disease among cases	Cases of the disease	100

Sources: WHO Indicator Metadata Registry for IMR, NMR, PMR, MMR and TFR; CDC (2012b, 2012d) for CBR, CDR and CFR; the SRS enumerator manual defines a stillbirth as a fetal death at 28 or more weeks of gestation (quoted in Dandona et al., 2023).

WORKED CALCULATION

A hypothetical vital-events register for District A (mid-year population 100,000; the population and deaths are those of District A in §7.4.2, the two-district standardisation example):

1. Live births 2,000 → CBR = 2,000/100,000 × 1,000 = **20 per 1,000**.
2. Deaths 789 → CDR = 789/100,000 × 1,000 = **7.89 per 1,000**.
3. Infant deaths 60 → IMR = 60/2,000 × 1,000 = **30 per 1,000 live births**. Of these, 40 were neonatal deaths → NMR = **20**, and post-neonatal mortality = (60 − 40)/2,000 × 1,000 = 10.
4. Stillbirths 30 and early neonatal deaths 28 → PMR = (30 + 28)/(2,000 + 30) × 1,000 = **28.6 per 1,000 births**. Note the denominator includes stillbirths.
5. Maternal deaths 2 → MMR = 2/2,000 × 100,000 = **100 per 100,000 live births**.
6. Age-specific fertility rates per 1,000 women aged 15–19, 20–24, … 45–49: 15, 140, 120, 65, 25, 8, 2. TFR = 5 × (15 + 140 + 120 + 65 + 25 + 8 + 2)/1,000 = 5 × 0.375 = **1.9 children per woman** (1.875).
7. A dengue outbreak: 250 cases, 4 deaths among them → CFR = 4/250 × 100 = **1.6%**.

MUST-KNOW NOTES (MD-LEVEL)

- **IMR is not a true rate.** Its denominator is live births, not person-time, and WHO defines it as a life-table probability of dying before age 1, expressed per 1,000 live births (WHO IMR metadata; §7.5 shows this probability).
- **MMR ratio vs rate.** The maternal mortality *ratio* divides by live births (a proxy for pregnancies) (WHO MMR metadata). A maternal mortality *rate* divides by women of reproductive age. Check which one a report uses.
- **CFR is a proportion.** WHO notes that "case fatality rate" is strictly a misnomer because there is no time component, and uses "case fatality ratio" instead (WHO, 2020). It measures severity, not risk to the population. During an ongoing outbreak it is biased by unresolved cases.
- **Denominators to memorise:** births (IMR, NMR, MMR); total births including stillbirths (PMR, stillbirth rate); mid-year population (CBR, CDR, specific rates).
- **Data sources** (recall §1.4.4): SRS gives state-level vital rates; NFHS gives survey-based estimates with CIs. Stillbirths appear under-reported in SRS relative to NFHS (Dandona et al., 2023).

CHECKPOINT 7.5

1. A block with a mid-year population of 60,000 registers 1,500 live births, 42 infant deaths (27 of them neonatal), 25 stillbirths, 18 early neonatal deaths and 3 maternal deaths. Compute the CBR, IMR, NMR, PMR and MMR.
2. Why is the case fatality rate not a rate, and why can it mislead early in an outbreak?
3. Diabetes has a prevalence of 10% in a town, and people live with it for an average of 20 years. Assuming a steady state, what is the approximate annual incidence among people without diabetes?

(Answers in the Appendix at the end of the book.)

7.4.2 Direct Standardisation

DEFINITION

A **directly age-standardised rate (DSR)** is the death rate a study population would have if it had the age structure of a chosen **standard population**. It is a weighted mean of the study population's age-specific rates, weighted by the standard population's age distribution (recall §2.2.2):

$$\text{DSR} = \Sigma(r_i \times S_i)/\Sigma S_i,$$

where r_i is the study population's rate in age group i and S_i the standard population in that group.

SIMPLIFIED EXPLANATION

Death rates rise steeply with age, so an older district can have a higher crude death rate even if every age group in it is healthier. A crude rate is itself a weighted mean of age-specific rates, with the district's own age structure as weights. Direct standardisation swaps in the same weights for both districts, so any remaining difference reflects the age-specific rates, not the age mix (CDC, 2012b).

WORKED CALCULATION

Two districts of 100,000 people each (`data/standardisation_districts.csv`); standard population 100,000.

Steps 1–2: the two districts.

Age	Pop A	Deaths A	Rate A /1000	Pop B	Deaths B	Rate B /1000
0–4	9,000	90	10.0	16,000	208	13.0
5–14	17,000	17	1.0	26,000	39	1.5
15–29	26,000	39	1.5	28,000	56	2.0
30–44	22,000	55	2.5	16,000	51	3.19
45–59	16,000	128	8.0	9,000	90	10.0
60+	10,000	460	46.0	5,000	280	56.0
Total	100,000	789	**7.89** (crude)	100,000	724	**7.24** (crude)

Steps 3–4: expected deaths in the standard population.

Age	Standard pop	Expected at A's rates	Expected at B's rates
0–4	10,000	100.0	130.0
5–14	20,000	20.0	30.0
15–29	27,000	40.5	54.0
30–44	20,000	50.0	63.75
45–59	14,000	112.0	140.0
60+	9,000	414.0	504.0
Total	100,000	**736.5**	**921.75**

1. Crude rates: A = 789/100,000 × 1,000 = 7.89; B = 7.24 per 1,000. A looks worse (crude rate ratio 1.09).
2. Age-specific rates: B is higher in *every* age group.
3. Expected deaths in the standard population, at each district's own rates: multiply each rate by the standard population and divide by 1,000 (e.g. 0–4 in A: 10.0 × 10,000/1,000 = 100).
4. DSR_A = 736.5/100,000 × 1,000 = **7.36 per 1,000**; DSR_B = 921.75/100,000 × 1,000 = **9.22 per 1,000**.
5. Standardised rate ratio A/B = 7.36/9.22 = **0.80**. After age adjustment, A's mortality is 20% *lower* than B's; the crude ranking was reversed.
6. Why: 10% of A's people are 60 or older against 5% of B's, and deaths at 60+ make up 58% of A's deaths but only 39% of B's (Figure 7.3).

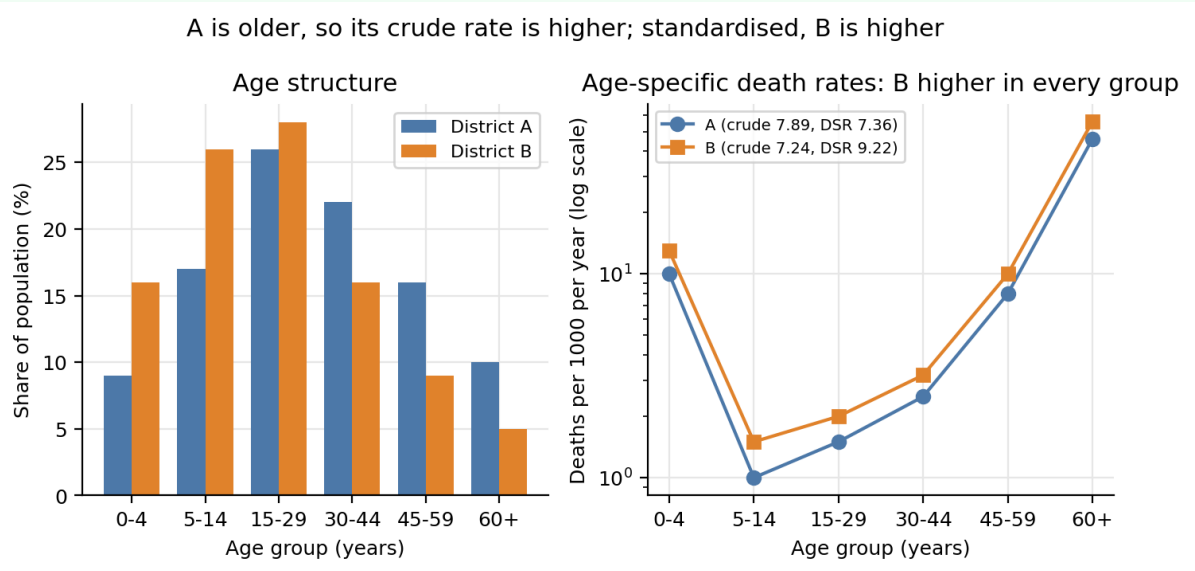


Figure 7.3. Left: District A has twice B's share of people aged 60+. Right: B's age-specific death rates are higher in every age group (log scale). The crude rates favour B; the directly standardised rates favour A.

IN R (OPTIONAL)

```
library(epitools)
sd ← read.csv("data/standardisation_districts.csv")
round(1000 * ageadjust.direct(count = sd$deathsA, pop = sd$popA,
                              stdpop = sd$std_pop), 2)
round(1000 * ageadjust.direct(count = sd$deathsB, pop = sd$popB,
                              stdpop = sd$std_pop), 2)
```

```
crude.rate   adj.rate       lci       uci
      7.89       7.36      6.86      7.90
crude.rate   adj.rate       lci       uci
      7.24       9.22      8.51      9.97
```

`crude.rate` and `adj.rate` are steps 1 and 4 (multiplied by 1,000 to read per 1,000). `lci` and `uci` are 95% confidence limits for each DSR (§7.4.4). The two intervals do not overlap, so the reversal is not a small-number artefact.

MUST-KNOW NOTES (MD-LEVEL)

- **The DSR is a comparison device.** Its value depends on the standard chosen. Only DSRs computed with the *same* standard can be compared.
- **Requirements.** Direct standardisation needs the age-specific rates of each study population. With few deaths per age group those rates are unstable; indirect standardisation (§7.4.3) is then preferred.
- **Confounding by age.** The crude-versus-standardised reversal is an example of confounding (§11.6: a third variable, here age, linked to both the "exposure", district, and the outcome, death). Standardisation is one way to control it.

7.4.3 Indirect Standardisation and the SMR

DEFINITION

Indirect standardisation applies a standard population's age-specific rates to the study population's own age structure, giving the **expected deaths (E)**. The **standardised mortality ratio** is

SMR = observed deaths (O)/expected deaths (E).

SMR = 1 means mortality as expected from the standard rates; SMR = 1.23 means 23% more deaths than expected.

SIMPLIFIED EXPLANATION

Indirect standardisation asks: "If this district died at the standard rates, how many deaths would we expect, given its own age mix?" It needs only the district's total deaths and its population by age, not its age-specific death rates, so it works when those rates are unreliable or unknown, for example in a small PHC catchment or an occupational cohort.

DERIVATION

$E = \Sigma(R_i \times n_i)$, where R_i is the standard rate and n_i the study population in age group i. $O = \Sigma(r_i \times n_i)$. So $SMR = \Sigma(r_i n_i)/\Sigma(R_i n_i)$: the study population's own rates relative to the standard rates, both weighted by the study population's age structure. Multiplying the standard population's crude rate by the SMR gives an **indirectly standardised rate**.

WORKED CALCULATION

Standard rates per 1,000 (last column of `data/standardisation_districts.csv`): 9.0, 1.2, 1.8, 2.9, 8.5, 48.0.

1. $E_A = (9.0 \times 9{,}000 + 1.2 \times 17{,}000 + 1.8 \times 26{,}000 + 2.9 \times 22{,}000 + 8.5 \times 16{,}000 + 48.0 \times 10{,}000)/1{,}000 = 81 + 20.4 + 46.8 + 63.8 + 136 + 480 =$ **828.0**.
2. $SMR_A = 789/828.0 =$ **0.95**: A has 5% fewer deaths than expected.
3. $E_B = 144 + 31.2 + 50.4 + 46.4 + 76.5 + 240 =$ **588.5**.
4. $SMR_B = 724/588.5 =$ **1.23**: B has 23% more deaths than expected.
5. Indirectly standardised rates: the standard population's crude rate is 7.72 per 1,000, so $ISR_A = 7.72 \times 0.953 = 7.35$ and $ISR_B = 7.72 \times 1.230 = 9.49$ per 1,000. Direct and indirect methods agree that B has the higher mortality.

MUST-KNOW NOTES (MD-LEVEL)

- **Direct vs indirect.** Direct: needs study age-specific rates, uses standard *population* weights, gives a DSR. Indirect: needs study age structure and total deaths, uses standard *rates*, gives an SMR.
- **Each SMR compares one population with the standard.** The derivation shows that each SMR weights by its own population's age structure, so two SMRs are averages over different weights. Comparing them is fair only if the ratio r_i/R_i is similar across age groups. Here it ranges from 0.83 to 1.11 in A and from 1.10 to 1.44 in B, so the DSRs of §7.4.2 are the better basis for ranking A against B.
- The same logic gives standardised incidence ratios (SIR) for cancer registries and standardised ratios for hospital mortality.

7.4.4 Confidence Intervals for Rates, Rate Ratios and SMRs

DEFINITION

When a count d of events arises from independent events in a large population, d follows a Poisson distribution (recall §5.7.1.3) with variance equal to its mean, so $SE(d) \approx \sqrt{d}$. Hence:

- Rate $r = d/PT$ (PT = person-time): $SE(r) = \sqrt{d}/PT$, and 95% CI = $r \pm 1.96\sqrt{d}/PT$ (or the exact Poisson interval when d is small).
- Rate ratio $RR = r_1/r_2$: $SE(\ln RR) = \sqrt{(1/d_1 + 1/d_2)}$, and 95% CI = $RR \times \exp(\pm 1.96 \times SE)$.
- SMR = O/E: exact Poisson limits for O divided by E, or approximately $SMR \times \exp(\pm 1.96/\sqrt{O})$.

SIMPLIFIED EXPLANATION

The precision of a rate depends on the number of events, not on the size of the population. 789 deaths among 100,000 people give the same relative precision as 789 deaths among 10 million. Ratios are handled on the log scale, where their sampling distribution is closer to Normal and the interval comes back asymmetric.

DERIVATION

SE of a log count. For a Poisson count, Var(d) = d. A small change in d changes ln d by about (1/d) × the change, so Var(ln d) ≈ $(1/d)^2 \times \text{Var}(d) = (1/d)^2 \times d = 1/d$.

SE of a log rate ratio. $\ln RR = \ln d_1 - \ln d_2 - \ln PT_1 + \ln PT_2$. The person-times are fixed, and d_1 and d_2 are independent, so variances add (as in §7.2.3): $\text{Var}(\ln RR) = 1/d_1 + 1/d_2$.

SMR. E is computed from standard rates and treated as fixed, so all the uncertainty is in O: Var(ln SMR) ≈ 1/O.

WORKED CALCULATION

1. *Crude death rate of District A.* d = 789, PT = 100,000 person-years. SE = √789/100,000 = 0.00028 = 0.28 per 1,000. 95% CI = 7.89 ± 1.96 × 0.28 = **7.34 to 8.44 per 1,000**.
2. *Crude rate ratio A/B.* RR = 7.89/7.24 = 1.09. SE(ln RR) = √(1/789 + 1/724) = 0.051. 95% CI = 1.09 × exp(± 1.96 × 0.051) = **0.99 to 1.21**. The crude comparison is compatible with no difference.
3. *SMRs.* Exact 95% CI: SMR_A = 0.95 (**0.89 to 1.02**); SMR_B = 1.23 (**1.14 to 1.32**). The approximation SMR × exp(± 1.96/√O) gives 0.89 to 1.02 and 1.14 to 1.32 with these large counts. B's excess mortality is clearly beyond what sampling variation explains; A's deficit is not.

IN R (OPTIONAL)

```
sd ← read.csv("data/standardisation_districts.csv")
E_B ← sum(sd$std_rate_per1000 * sd$popB / 1000)
poisson.test(sum(sd$deathsB), E_B)       # SMR for B, exact CI
poisson.test(c(sum(sd$deathsA), sum(sd$deathsB)),
             c(sum(sd$popA), sum(sd$popB)))
```

```
    Exact Poisson test
data:  sum(sd$deathsB) time base: E_B
...
95 percent confidence interval:
 1.142254 1.323219
sample estimates:
event rate
  1.230246
...
    Comparison of Poisson rates
...
95 percent confidence interval:
 0.9839624 1.2071123
sample estimates:
rate ratio
  1.089779
```

Giving `poisson.test` the observed count and E as the "time base" makes the estimated "event rate" the SMR (1.23) and returns its exact CI (1.14 to 1.32), step 3. With two counts and two populations it returns the crude rate ratio (1.09) with an exact CI of 0.98 to 1.21, slightly different in the second decimal from the log-scale approximation of step 2 (0.99 to 1.21). The p-values are explained in §8.5.

MUST-KNOW NOTES (MD-LEVEL)

- **Does the Poisson model apply?** It assumes events occur independently. That is reasonable for all-cause deaths in a district. It is not for cases in an infectious outbreak, where one case causes others; the counts then vary more than Poisson allows and these CIs are too narrow (§13.6 treats this overdispersion).
- Relative precision depends on d: SE(ln rate) = 1/√d. 100 events give about ±20%, 400 events about ±10%.
- When d is small, the Normal approximation ±1.96√d is poor and can even give a negative lower limit; use exact Poisson limits. There is no fixed cut-off; with a few tens of events or fewer, prefer the exact limits.

CHECKPOINT 7.6

1. Two districts have age-specific death rates of 2 and 20 per 1,000 (District X) and 3 and 24 per 1,000 (District Y) in the under-60 and 60+ groups. Using a standard population of 60,000 under-60 and 40,000 aged 60+, compute both DSRs.
2. A PHC catchment records 30 deaths where 20 were expected at state rates. Compute the SMR and its approximate 95% CI. What do you conclude?
3. Why is indirect rather than direct standardisation used for a small PHC catchment?

(Answers in the Appendix at the end of the book.)

7.5 The Life Table

DEFINITION

A **life table** follows a hypothetical group of 100,000 births (the radix) through the age-specific death rates of one period, recording at each age how many survive and how many years they will live. Its columns are:

- ${}_nM_x$: death rate between ages x and x + n (deaths ÷ mid-year population);
- ${}_nq_x$: probability that a person alive at exact age x dies before x + n;
- l_x: number alive at exact age x;
- ${}_nd_x = l_x \times {}_nq_x$: deaths between x and x + n;
- ${}_nL_x$: person-years lived between x and x + n;
- T_x: total person-years lived beyond age x;
- $e_x = T_x/l_x$: **life expectancy** at age x, the average number of further years lived by those who reach x.

SIMPLIFIED EXPLANATION

A life table turns a set of age-specific death rates into survival. Imagine 100,000 babies who experience, at each age, the death rates observed this year. Count how many die in each age band, how many years everyone lives in total, and divide by the number starting out. The result, e_0, summarises a whole mortality pattern in one number that does not depend on the population's age structure. This is a **current (period) life table**. A **cohort life table** instead follows people actually born in one year until all have died. An **abridged** table uses age bands, as here; a **complete** table uses single years.

DERIVATION

From rate to probability. In a band of width n, let a_x be the average years lived in the band by those who die in it. Person-years lived in the band are ${}_nL_x = n(l_x - {}_nd_x) + a_x \times {}_nd_x$, and the death rate is ${}_nM_x = {}_nd_x/{}_nL_x$. Substitute ${}_nd_x = l_x \times q$ and ${}_nL_x = l_x[n - (n - a_x)q]$, then solve for q:

${}_nq_x = n \times {}_nM_x/[1 + (n - a_x) \times {}_nM_x]$.

Survivors. Surviving each band is conditional on reaching it (recall §5.3), so $l_{x+n} = l_x \times (1 - {}_nq_x)$.

Open last band (60+). Everyone dies, so q = 1 and d = l. With a constant death rate M, person-years = l/M, so $e_{60} = 1/M_{60}$.

Life expectancy. T_x = sum of ${}_nL_x$ from age x upwards; $e_x = T_x/l_x$.

WORKED CALCULATION

`data/lifetable_district.csv` gives mid-year population and deaths in seven age bands for a hypothetical Chhattisgarh district. For this example, a = 0.1 year for infants who die (reflecting deaths concentrated in the first weeks of life), 1.6 years for ages 1–4, and n/2 for other bands; these values are assumptions of the example.

Age x	n	$_nM_x$	$_nq_x$	l_x	$_nd_x$	$_nL_x$	T_x	e_x
0	1	0.0310	0.0302	100,000	3,016	97,286	6,774,296	67.7
1	4	0.0024	0.0095	96,984	926	385,715	6,677,010	68.8
5	10	0.0007	0.0070	96,058	670	957,234	6,291,295	65.5
15	15	0.0015	0.0222	95,388	2,122	1,414,908	5,334,061	55.9
30	15	0.0030	0.0440	93,266	4,105	1,368,205	3,919,154	42.0
45	15	0.0090	0.1265	89,161	11,276	1,252,853	2,550,948	28.6
60+	—	0.0600	1.0000	77,886	77,886	1,298,095	1,298,095	16.7

1. Age 0: M = 93/3,000 = 0.0310. $q_0 = 0.0310/(1 + 0.9 \times 0.0310) = 0.0302$. So $1{,}000 \times q_0$ = **30.2 infant deaths per 1,000 live births**, the life-table form of the IMR.
2. $d_0 = 100{,}000 \times 0.0302 = 3{,}016$; $l_1 = 96{,}984$.
3. $L_0 = 1 \times (100{,}000 - 3{,}016) + 0.1 \times 3{,}016 = 97{,}286$ person-years.
4. Ages 1–4: M = 30/12,500 = 0.0024; $q = 4 \times 0.0024/(1 + 2.4 \times 0.0024) = 0.0095$; d = 926; $L = 4 \times (96{,}984 - 926) + 1.6 \times 926 = 385{,}715$.
5. Continue down the table. At 60+: q = 1, $L_{60} = 77{,}886/0.06 = 1{,}298{,}095$, and $e_{60} = 1/0.06$ = **16.7 years**.
6. T_0 = sum of all L = 6,774,296. $e_0 = 6{,}774{,}296/100{,}000$ = **67.7 years**.
7. $e_1 = 6{,}677{,}010/96{,}984$ = **68.8 years**, more than e_0, because a baby who survives the high-risk first year has left that risk behind.
8. Probability of surviving from birth to 60 = $l_{60}/l_0 = 77{,}886/100{,}000$ = **0.779**.

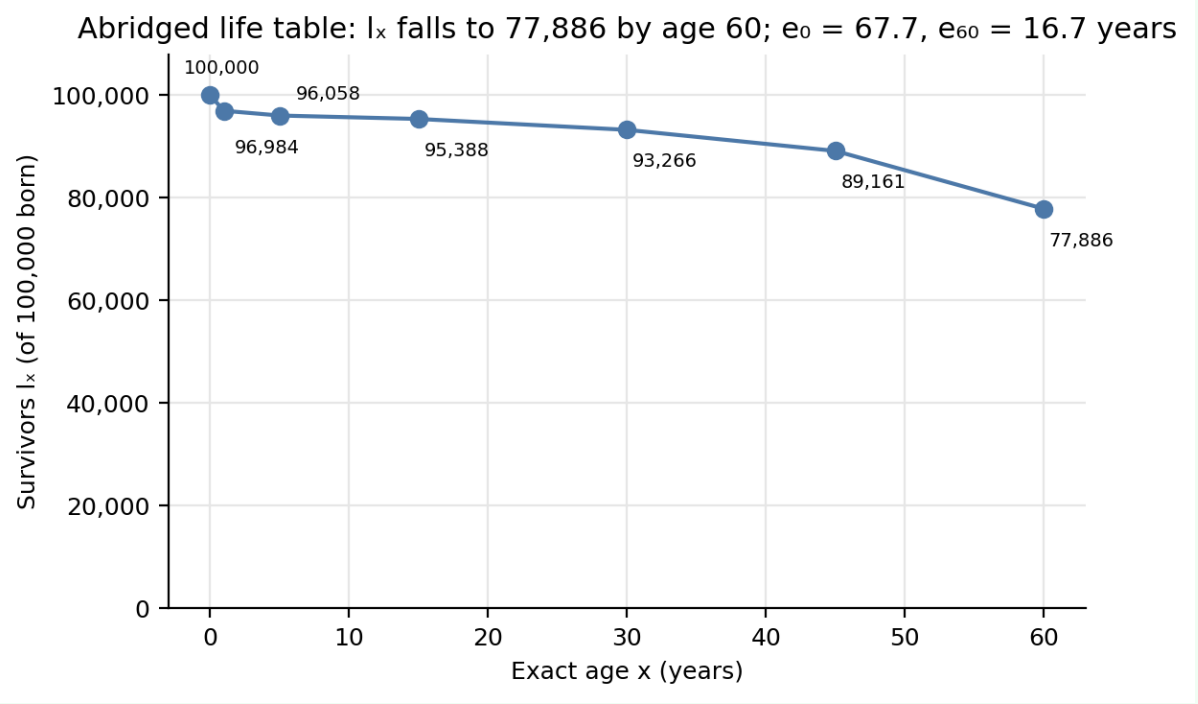


Figure 7.4. Survivors l_x out of 100,000 births. The steep drop in the first year and after age 45 shows where deaths concentrate. The curve stops at 60 because the last band is open.

IN R (OPTIONAL)

```
lt ← read.csv("data/lifetable_district.csv")
n  ← c(1, 4, 10, 15, 15, 15, NA)          # interval widths
a  ← c(0.1, 1.6, 5, 7.5, 7.5, 7.5, NA)    # years lived if dying
M  ← lt$deaths / lt$midyear_pop            # nMx
q  ← n * M / (1 + (n - a) * M); q[7] ← 1   # nqx
l  ← 100000 * cumprod(c(1, 1 - q[-7]))     # lx
d  ← l * q                                 # ndx
L  ← n * (l - d) + a * d; L[7] ← l[7] / M[7]   # nLx
Tx ← rev(cumsum(rev(L)))                   # Tx
e  ← Tx / l                                # ex
data.frame(age = lt$age_start, M = round(M, 4), q = round(q, 4),
           l = round(l), d = round(d), L = round(L), T = round(Tx),
           e = round(e, 1))
```

```
  age      M      q      l     d       L       T    e
1   0 0.0310 0.0302 100000  3016   97286 6774296 67.7
2   1 0.0024 0.0095  96984   926  385715 6677010 68.8
3   5 0.0007 0.0070  96058   670  957234 6291295 65.5
4  15 0.0015 0.0222  95388  2122 1414908 5334061 55.9
5  30 0.0030 0.0440  93266  4105 1368205 3919154 42.0
6  45 0.0090 0.1265  89161 11276 1252853 2550948 28.6
7  60 0.0600 1.0000  77886 77886 1298095 1298095 16.7
```

Each line of code is one column of the table: the rate-to-probability conversion, the cumulative product that builds l_x, the person-years, the reverse cumulative sum that builds T_x, and e_x. The printed rows reproduce the hand-worked table.

MUST-KNOW NOTES (MD-LEVEL)

- **e_0 is not "the age at which people die".** It is the mean of the life-table distribution of ages at death. High infant mortality pulls e_0 down, which is why e_1 can exceed e_0.
- **Period tables are snapshots.** $e_0 = 67.7$ is what a baby would live if this year's death rates held for life. Falling mortality means real cohorts usually live longer.
- **The IMR as a probability.** WHO defines the IMR as the life-table probability of dying before age 1 (WHO IMR metadata). Here $1{,}000 \times q_0 = 30.2$, close to, but not the same as, the death rate at age 0 ($1{,}000 \times M_0 = 31.0$).
- **Choices of a_x matter for infants.** When infant deaths cluster early in life, a_0 is well below 0.5, and the value chosen changes q_0 and e_0. State the values you assume.
- **Link to survival analysis.** The life table's l_x/l_0 is a survival curve built from grouped population data. §14.4 (Kaplan–Meier, the survival curve estimated from individuals' follow-up times) builds the same curve from a cohort in which people are followed for different lengths of time.

CHECKPOINT 7.7

1. In a 5-year band, the death rate is 0.02 per person-year and those who die live on average 2.5 years in the band. Compute ${}_5q_x$.
2. Using the table above, what is the probability that a 15-year-old survives to age 45?

3. If the open 60+ band had a death rate of 0.05 per person-year, what would e_{60} be? Why can e_1 exceed e_0?

(Answers in the Appendix at the end of the book.)

Hypothesis Testing – The Framework

Chapter 7 asked "what is our best estimate, and how precise is it?" For the urban and rural children of §7.2.3 the answer was a difference in mean weight of 1 kg, 95% CI 0.20 to 1.80 kg. This chapter asks a related question of the same data: is the evidence strong enough to reject a stated claim, such as "the true difference is zero"? It builds the logic and vocabulary shared by every test (§8.1–8.7) and ends with a table that matches each study design and outcome to its test and model (§8.8). The specific tests follow in Chapters 9 and 10.

8.1 Logic of Hypothesis Testing

DEFINITION

The **logic of hypothesis testing** assumes that a specific claim about the population is true, then asks how unusual the observed data would be under that assumption. If the data would be very unusual, the claim is rejected; otherwise it is retained.

SIMPLIFIED EXPLANATION

The reasoning resembles proof by contradiction, or a court's "innocent until proven guilty". You assume a sceptical claim, for example "urban and rural children do not differ in mean weight". You then ask: if that were true, how often would chance alone produce a difference as large as the one observed? If rarely, the claim becomes hard to believe. If often, you lack evidence against it. Lacking evidence against a claim is not the same as proving it.

ILLUSTRATION / EXAMPLE

Assume a new drug has no effect on recovery. If a difference in recovery as large as the one observed would arise by chance less than 5% of the time under that assumption, you have grounds to reject "no effect". If such a difference would be common even with no effect, you do not reject "no effect", although the drug may still have a smaller effect that this study could not detect.

MUST-KNOW NOTES (MD-LEVEL)

- The assumed claim is the null hypothesis (§8.2). How unusual the data are under it is measured by the p-value (§8.5), and the cut-off for "unusual enough" is the significance level (§8.4).
- "Failing to reject" the claim is not "proving" it. Absence of evidence is not evidence of absence.
- Every test in Chapters 9 and 10 (t-test, chi-square, ANOVA and the rest) applies this same reasoning to a different type of data.

8.2 Null and Alternative Hypothesis

DEFINITION

The **null hypothesis (H_0)** is the claim assumed true for the test. It fixes a value of the parameter, so that the sampling distribution of the statistic can be worked out. Usually that value is "no effect": a difference of 0 or a ratio of 1. The **alternative hypothesis (H_1 or H_a)** covers the other possible values: the effect, difference or association the researcher is usually looking for.

SIMPLIFIED EXPLANATION

H_0 has to name a value because the test needs a sampling distribution to compare the data with, and a sampling distribution needs a centre (recall §6.5). "The difference is 0" gives one; "there is some difference" does not. That is why the sceptical, precise claim is the one tested, and H_1 is supported only indirectly, by evidence against H_0.

ILLUSTRATION / EXAMPLE

For the urban–rural weights of §7.2.3: H_0: μ_urban − μ_rural = 0; H_1: μ_urban − μ_rural ≠ 0.

H_0 need not say "no effect". To ask whether a district's mean birthweight differs from a reference value of 3.0 kg, the hypotheses are H_0: μ = 3.0 kg and H_1: μ ≠ 3.0 kg (the one-sample tests of §9.1 and §9.2.1). In a non-inferiority trial (a trial designed to show that a new treatment is not worse than the standard by more than a stated margin M, §15.9), the null is "the new treatment is worse by M or more", and H_1 is "it is worse by less than M, or better" (ICH, 1998). Greenland et al. (2016) call the general case the *test hypothesis*; the "no effect" null is its most common form.

MUST-KNOW NOTES (MD-LEVEL)

- H_0 is usually "no difference / no association", but it can be any stated value: a reference mean, a target coverage, a non-inferiority margin.
- The test only ever assesses evidence against H_0. It never proves H_1, and failing to reject H_0 does not prove H_0 (recall §8.1).
- H_0 and H_1 are mutually exclusive and together cover every possible value of the parameter, so rejecting one supports the other.
- Whether H_1 specifies a direction ("the drug improves recovery") or not ("the drug changes recovery") decides whether the test is one-tailed or two-tailed (§8.6).

8.3 Type I and Type II Error

DEFINITION

A **Type I error** is rejecting H_0 when it is true: concluding there is an effect when there is none. A **Type II error** is failing to reject H_0 when it is false: missing an effect that exists.

SIMPLIFIED EXPLANATION

A test is a decision made from one sample, so it can go wrong in two ways. Cross the two true states (H_0 true, H_0 false) with the two decisions (reject, do not reject) and you get a 2 × 2 table:

Decision	H_0 true	H_0 false
Reject H_0	Type I error (probability α)	Correct (probability 1 − β, the power)
Do not reject H_0	Correct (probability 1 − α)	Type II error (probability β)

Figure 8.1 draws the two columns as two sampling distributions of the urban–rural difference, both with SE = 0.406 kg (recall §7.2.3). If H_0 is true, the observed difference centres on 0. If the true difference is 1 kg, it centres on 1. A two-sided test at α = 0.05 rejects H_0 when the observed difference lies beyond ±1.96 × 0.406 = ±0.80 kg. The red tails are the Type I error: the chance, under H_0, of landing beyond the cut-offs. The grey area is the Type II error: the chance, when the truth is 1 kg, of landing inside them.

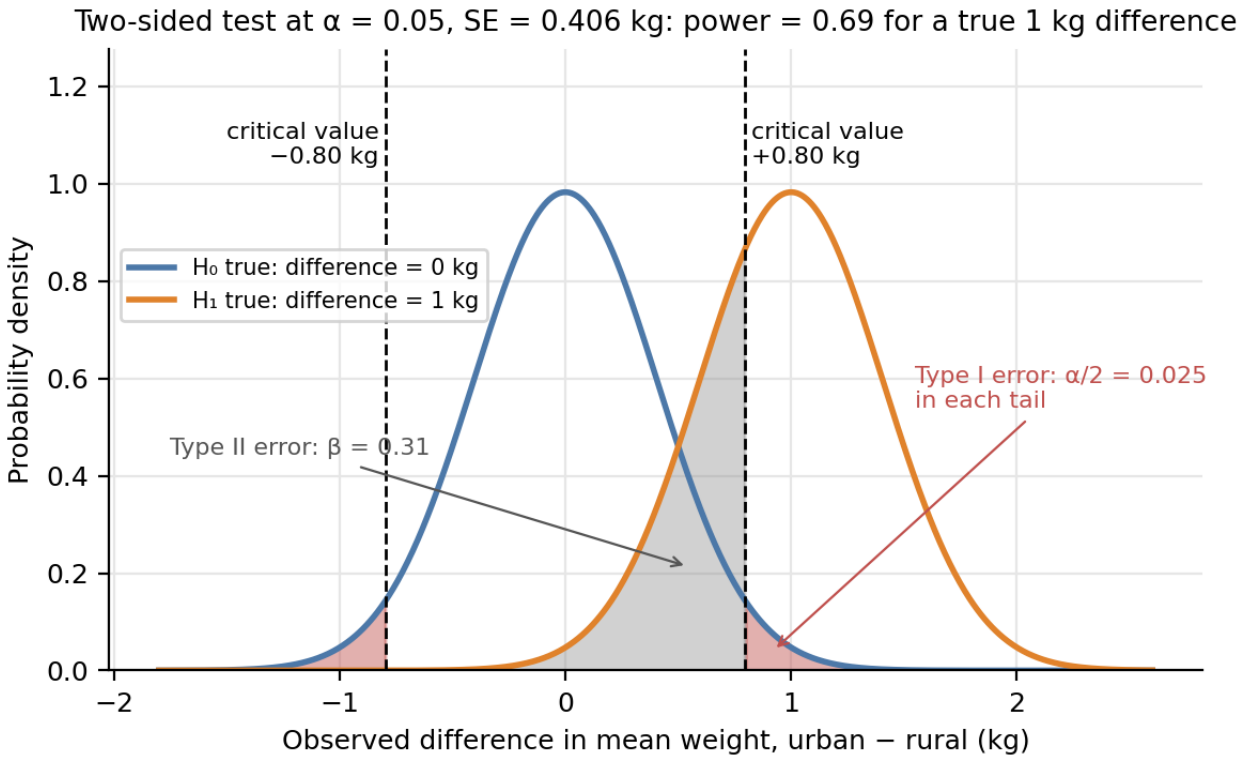


Figure 8.1. Type I and Type II error for the urban–rural comparison of §7.2.3 (SE 0.406 kg). Two-sided test at α = 0.05: α/2 = 0.025 in each tail of the H_0 curve; β = 0.31 under a true 1 kg difference, so power = 0.69.

ILLUSTRATION / EXAMPLE

Testing a new drug against H_0 "no effect": a Type I error means approving a useless drug; a Type II error means abandoning a useful one. The two errors often differ in cost, which is why the choice of α (§8.4) is a judgement, not a law. Move the cut-offs in Figure 8.1 outwards (a smaller α) and the red tails shrink while the grey area grows.

MUST-KNOW NOTES (MD-LEVEL)

- α = P(reject H_0 | H_0 true); β = P(do not reject H_0 | H_0 false). 1 − β is the power (§8.4).

- For a fixed sample size and design, lowering α raises β. To lower both you must shrink the standard error. For a given design and variability the main lever is sample size (SE falls with √n, recall §6.7). A paired or matched design, or a more precise measurement, also lowers the SE and so reduces both errors.
- β is not one number: it depends on the true effect. In Figure 8.1, β = 0.31 for a true 1 kg difference; for a true 0.5 kg difference it is larger (§8.4).

8.4 Significance Level and Power

DEFINITION

The **significance level (α)** is the probability of a Type I error that the researcher accepts, fixed before the data are seen. **Power (1 – β)** is the probability that the test rejects H_0 when a specified true effect exists. An **effect size** is the magnitude of the effect or association being studied. A **raw** effect size is on the outcome's own scale: a difference in means (kg), a risk difference, a risk ratio. A **standardised** effect size divides a difference by a standard deviation, which makes it unit-free; for two means it is Cohen's d = $(\bar{x}_1 - \bar{x}_2)/s_p$, where s_p is the pooled SD, $s_p = \sqrt{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)}$.

SIMPLIFIED EXPLANATION

α is chosen by the researcher: "I accept rejecting a true H_0 this often, and no more." Power is the other side of the Type II error rate: the chance that the study will find an effect of a stated size if it is there. Power always refers to a stated effect size, so the effect size must be defined first. A confidence interval is not an effect size. It is an interval estimate *of* an effect size, showing how precisely the effect has been measured.

DERIVATION

Power of a two-sided z-test for a true difference Δ, when the estimate has standard error SE:

1. Reject H_0 when the observed difference D lies beyond ±1.96 × SE.
2. If the truth is Δ, then D is approximately normal with mean Δ and SD SE (recall §6.6), so (D − Δ)/SE is a standard normal Z.
3. P(D > 1.96 SE) = P(Z > 1.96 − Δ/SE) = Φ(Δ/SE − 1.96). Here Φ(x) denotes the probability that a standard normal Z (recall §5.7.2.2.1) falls below x, read from normal tables or from any statistical software.
4. The opposite tail adds Φ(−Δ/SE − 1.96), which is negligible once Δ/SE exceeds about 1.
5. So **power ≈ Φ(Δ/SE − 1.96)**. Power rises as Δ grows or SE shrinks.
6. Reversing it: for equal groups of n with SD σ, SE = σ√(2/n). Setting Δ/SE = 1.96 + 0.84 (0.84 is the z-value with 80% below it) gives $\mathbf{n = 2(1.96 + 0.84)^2\sigma^2/\Delta^2}$ per group for 80% power. §15.6 (sample size for comparative studies) develops this.

WORKED CALCULATION

Urban–rural weights (recall §7.2.3: $n_1 = 40$, SD 2.0 kg; $n_2 = 50$, SD 1.8 kg; SE_diff = 0.406 kg).

1. Raw effect size: 1 kg.
2. Pooled SD: $s_p = \sqrt{[(39 \times 2.0^2 + 49 \times 1.8^2)/88]} = \sqrt{[(156.0 + 158.8)/88]} = \sqrt{3.577} = 1.891$ kg. Standardised effect size: $d = 1/1.891 = 0.53$.
3. Power for a true 1 kg difference at $\alpha = 0.05$: $\Delta/SE = 1/0.406 = 2.463$; power $= \Phi(2.463 - 1.96) = \Phi(0.503) =$ **0.69**. So $\beta = 0.31$ (Figure 8.1).
4. For a true 0.5 kg difference: $\Phi(0.5/0.406 - 1.96) = \Phi(-0.73) =$ **0.23**. The same study would miss a 0.5 kg difference more than three times in four.
5. Doubling both groups (80 and 100): $SE = \sqrt{(4/80 + 3.24/100)} = 0.287$ kg, power for 1 kg = **0.94**.
6. Tightening α to 0.01 (cut-off 2.576) with the original groups: power for 1 kg = **0.46**.
7. Planning: equal groups with $\sigma = 2$ kg and $\Delta = 1$ kg need $n = 2 \times 2.80^2 \times 4/1 = 62.7 \rightarrow$ **63 per group** for 80% power. The slightly more exact calculation with the t-distribution, which software does, gives 63.8 → 64.

IN R (OPTIONAL)

```
se ← sqrt(2.0^2/40 + 1.8^2/50)     # SE of urban-rural difference
pnorm(1/se - qnorm(0.975))         # power, true difference 1 kg
power.t.test(delta = 1, sd = 2, power = 0.80)
```

```
[1] 0.6926441

     Two-sample t test power calculation

              n = 63.76576
          delta = 1
             sd = 2
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

The first line is step 3: power 0.69 for a true 1 kg difference. `power.t.test` solves the planning question of step 7 using the t-distribution: 63.8, rounded up to 64 children per group, one more than the z-formula's 63 because t has heavier tails (recall §5.7.2.3).

COMMON MISREADING

"The result was not significant, so we calculated the power from our observed effect, and it was low." This *observed* (post-hoc) power adds nothing. Put the observed z into the power formula and you get $\Phi(|z| - 1.96)$: a one-to-one function of the p-value. At $p = 0.05$, $|z| = 1.96$ and observed power is exactly 0.50; at $p = 0.20$ it is 0.25. Any result with $p > 0.05$ has observed power below about 50%, by arithmetic alone (Hoenig & Heisey, 2001; Greenland et al., 2016). Power is a planning quantity,

computed before the study for an effect size chosen as clinically important. After the study, report the effect estimate with its confidence interval.

MUST-KNOW NOTES (MD-LEVEL)

- **α = 0.05 is a convention.** Fisher (1925) called P = 0.05 "convenient to take … as a limit"; it has been the usual default since (Cohen, 1992). Where many tests are run, stricter thresholds are used (§8.5).
- **80% power is a convention.** Cohen (1992) proposed 0.80 "for general use"; ICH E9 notes that the Type II error is "conventionally" set at 10% to 20%, that is, power of 80% to 90% (ICH, 1998). Cohen also labelled d = 0.2, 0.5 and 0.8 small, medium and large, but called these subjective conventions; a clinically important difference should be judged on the raw scale.
- **Power depends on four things:** sample size, α, the true effect size and the variability of the data (together, Δ/SE and α). A larger n, a larger α, a larger effect or a smaller SD each raise power.
- **A non-significant result from a small study is ambiguous.** There may be no effect, or the study may have been too small to detect one. The confidence interval shows which: a wide interval that includes clinically important effects means "inconclusive", not "no effect".

CHECKPOINT 8.1

1. A small study reports "not statistically significant", and adds that its observed power was 30%. Explain why this does not show there is no real effect, and why the observed power adds no information. What should the authors report instead?
2. Explain the trade-off between Type I and Type II error using a clinical example of your choosing. What, besides a larger sample, could reduce both?
3. For the urban–rural comparison (SE_diff = 0.406 kg, pooled SD 1.891 kg), compute the standardised effect size d and the power for a true difference of 0.5 kg. How many children per group (σ = 2 kg, z-formula) would give 80% power for 0.5 kg?

(Answers in the Appendix at the end of the book.)

8.5 p-value – Correct Interpretation

DEFINITION

The **p-value** is the probability, calculated assuming H_0 and every other assumption of the statistical model are true, of obtaining a test statistic at least as extreme as the one observed.

SIMPLIFIED EXPLANATION

The p-value answers the question of §8.1: "how unusual would these data be if H_0 were true?" A small p-value means data like these would be unusual under H_0 and the model. A large p-value means they would not be unusual; it does not show that H_0 is true. The p-value is compared with the pre-chosen α: if $p < \alpha$, reject H_0.

DERIVATION

For a statistic that is approximately normal under H_0:

1. z = (estimate − value under H_0)/SE. This counts how many standard errors the estimate lies from the H_0 value.
2. Two-sided $p = P(|Z| \geq |z|) = 2 \times [1 - \Phi(|z|)]$.
3. $p < 0.05$ exactly when $|z| > 1.96$, that is, when |estimate − H_0 value| > 1.96 × SE.
4. That is the same as saying the H_0 value lies outside estimate ± 1.96 × SE, the 95% CI (recall §7.2). So a 95% CI excludes the H_0 value exactly when the two-sided $p < 0.05$, provided both use the same SE.

WORKED CALCULATION

Urban–rural weights (recall §7.2.3), H_0: difference = 0.

1. $z = (1 - 0)/0.406 =$ **2.46**.
2. Two-sided $p = 2 \times [1 - \Phi(2.46)] = 2 \times 0.0069 =$ **0.014**.
3. $0.014 < 0.05$, so reject H_0. Consistently, the 95% CI 0.20 to 1.80 kg excludes 0.
4. Reading: if urban and rural children truly had the same mean weight (and the model held), a difference of 1 kg or more in either direction would arise in about 1.4% of such pairs of samples.

IN R (OPTIONAL)

Chapter 7's `prop.test` output for the CBHI survey trimmed its test lines (recall §7.2.3). Here they are.

```
cb ← read.csv("data/cbhi_survey.csv")
tab ← table(cb$maternal_education, cb$cbhi_member)
prop.test(x = tab[c("secondary+", "none"), "1"],
          n = rowSums(tab)[c("secondary+", "none")], correct = FALSE)
2 * pnorm(-1/0.406)             # urban-rural: two-sided p, z = 2.46
```

```
    2-sample test for equality of proportions without continuity
    correction
...
X-squared = 25.622, df = 1, p-value = 4.153e-07
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1628515 0.3584235
...
[1] 0.01377591
```

H_0 is equal membership in the two education groups (observed: 44.5% vs 18.5%). `X-squared` is z^2, so $z = \sqrt{25.622} = 5.06$, and $p = 4.2 \times 10^{-7}$ (§9.1.2, the two-proportion z-test, shows why z^2 equals this chi-square). The CI (0.163 to 0.358) excludes 0, as the tiny p requires; the test's SE uses the pooled proportion under H_0 (0.0515) while the CI uses the unpooled SE (0.050), so for proportions the CI–test agreement is close rather than exact. The last line is step 2 of the urban–rural calculation. The

§7.2.3 caveat applies to the CBHI p-value too: households are clustered in villages, so this SE is too small and the p-value too small (§15.2, clustered data).

COMMON MISREADING

"p = 0.014 means there is a 1.4% probability that H_0 is true" or "a 1.4% probability that the result is due to chance". Both are wrong. The p-value is calculated *assuming* H_0 and chance alone; it cannot also be the probability of those assumptions (ASA, 2016, principle 2; Greenland et al., 2016).

MUST-KNOW NOTES (MD-LEVEL)

- **What the p-value is not.** It is not the probability that H_0 is true, nor that the data arose by chance alone. It does not measure the size or importance of an effect. A result should not be judged only by whether p crosses a threshold (ASA, 2016, principles 2, 3 and 5).
- **It tests the whole model.** A small p can arise because H_0 is false, or because another assumption failed: non-random sampling, ignored clustering, or selection of the analysis for its small p (Greenland et al., 2016).
- **Significance is not importance.** With a large n, a trivial difference gives a tiny p, because the SE shrinks with √n (recall §6.7). Always report the effect estimate with its confidence interval; the p-value alone says nothing about size or direction.
- **p-values are continuous.** A smaller p does indicate greater incompatibility of the data with H_0 under the model, but p = 0.04 and p = 0.06 are nearly the same evidence, not two different realities. Report the exact p (e.g. p = 0.014), not "p < 0.05"; an inequality is reasonable only for very small values such as p < 0.001 (Greenland et al., 2016).
- **Multiplicity.** Each test at α = 0.05 has a 5% Type I error rate. Run k independent tests on true null hypotheses and the chance of at least one false positive, the **family-wise error rate**, is $1 - (1 - \alpha)^k$. (Each test avoids a false positive with probability 0.95; by the multiplication rule for independent events, recall §5.2.2, all k do so with probability 0.95^k; subtract from 1 by the complement rule, §5.2.3.) For k = 5 it is $1 - 0.95^5 = 0.226$; for k = 10, 0.401; for k = 20, 0.642. Twenty subgroup tests of a useless drug will "find" something about two times in three. The simplest remedy, Bonferroni, tests each at α/k (0.01 for k = 5); §9.5.3 (post-hoc tests) treats Bonferroni, Holm and others. ICH E9 asks trialists to consider adjusting significance levels "to control the overall type I error" whenever more than one test is used (ICH, 1998).

8.6 One-Tailed vs Two-Tailed Tests

DEFINITION

A **two-tailed test** looks for a difference from H_0 in either direction (H_1: $\mu_1 - \mu_2 \neq 0$). A **one-tailed test** looks in one pre-specified direction only (H_1: $\mu_1 - \mu_2 > 0$, or < 0).

SIMPLIFIED EXPLANATION

The direction in H_1 (recall §8.2) decides where the rejection region sits. A two-tailed test splits α between the two tails: α/2 = 0.025 in each at α = 0.05. A one-tailed test puts all of α in one tail. That makes it easier to reject H_0 in the chosen direction, and makes the test blind to an effect in the other direction.

WORKED CALCULATION

Urban–rural weights, z = 2.46 (§8.5).

1. Two-tailed: critical z = ±1.96, critical difference ±1.96 × 0.406 = ±0.80 kg; p = 0.014.
2. One-tailed (H_1: urban heavier): critical z = 1.645, critical difference 0.67 kg; p = P(Z ≥ 2.46) = **0.007**, half the two-tailed value.
3. An observed difference of 0.75 kg would be "significant" one-tailed (> 0.67) but not two-tailed (< 0.80).

ILLUSTRATION / EXAMPLE

A school deworming drive might change mean haemoglobin. A two-tailed test asks "is Hb different, in either direction?" A one-tailed test of "does Hb rise?" could not detect a fall, however large, because that possibility was excluded when H_1 was written.

MUST-KNOW NOTES (MD-LEVEL)

- Choose one- or two-tailed *before* seeing the data, on scientific grounds. Switching to one-tailed after seeing the direction halves the p-value (step 2) and is a form of **p-hacking**: changing the analysis until p falls below α.
- Two-tailed tests are the usual default. When a one-sided hypothesis is the real question, as in a non-inferiority trial (§8.2), a one-sided test is appropriate, but say when and why it is used (Greenland et al., 2016).
- For a symmetric null distribution (z or t), the two-tailed p is exactly twice the one-tailed p in the observed direction.

CHECKPOINT 8.2

1. Two studies report p = 0.001 and p = 0.049. A colleague says the first is "much more significant", so its effect must be larger. What is right and what is wrong in this?
2. A trial tests a new drug in 8 independent subgroups, each at α = 0.05, and the drug truly does nothing. What is the probability of at least one "significant" subgroup? What per-test threshold would Bonferroni use?
3. A district sample of 100 births has mean birthweight 2.9 kg. Test H_0: μ = 3.0 kg with σ = 0.5 kg: compute z and the two-sided p. The team, seeing the result lies below 3.0 kg, switches to a one-tailed test. What p would they report, and why is the switch wrong?

(Answers in the Appendix at the end of the book.)

8.7 General Steps in Hypothesis Testing

DEFINITION

Every hypothesis test follows the same seven steps:

1. State H_0 and H_1, including the direction (one- or two-tailed).
2. Choose α.
3. Choose the test that fits the design and the data (§8.8), and check its assumptions.
4. Calculate the test statistic from the data.
5. Find the p-value.
6. Decide: reject H_0 if $p < \alpha$.
7. State the conclusion in the terms of the research question, with the effect estimate and its confidence interval.

SIMPLIFIED EXPLANATION

Each specific test in Chapters 9 and 10 changes only steps 3–5: a different statistic and a different reference distribution. Steps 1–3 must be fixed before looking at the data (recall the p-hacking warning in §8.6).

WORKED CALCULATION

Does not being exclusively breastfed (not EBF) raise the risk of diarrhoea? A cohort of 200 infants, followed up at two PHCs: 33 of 80 not-EBF infants and 14 of 120 EBF infants developed diarrhoea.

1. H_0: the risk of diarrhoea is the same in not-EBF and EBF infants ($p_1 - p_2 = 0$). H_1: the risks differ (two-tailed).
2. $\alpha = 0.05$.
3. Binary outcome, two independent groups: compare two proportions (§8.8; the test itself is §9.1.2). The normal approximation needs enough events and non-events in each group (recall §5.7.2.2.4): the smallest expected count, $80 \times 0.235 = 18.8$, is well above the conventional minimum of 5.
4. Risks: $p_1 = 33/80 = 0.412$; $p_2 = 14/120 = 0.117$; difference 0.296. Pooled risk under $H_0 = 47/200 = 0.235$. $SE_0 = \sqrt{[0.235 \times 0.765 \times (1/80 + 1/120)]} = \sqrt{(0.1798 \times 0.02083)} = 0.0612$. $z = 0.296/0.0612 = \mathbf{4.83}$.
5. Two-sided $p = 2 \times [1 - \Phi(4.83)] = 1.3 \times 10^{-6}$.
6. $p < 0.05$: reject H_0.
7. Conclusion: in this cohort the risk of diarrhoea was 29.6 percentage points higher in infants who were not exclusively breastfed (41.2% vs 11.7%; 95% CI 17.4 to 41.8 points; $p < 0.001$). This is an association. Whether a difference between the two PHCs explains part of it is checked in §11.7 (stratified analysis).

IN R (OPTIONAL)

```
e ← read.csv("data/ebf_diarrhoea_cohort.csv")
tab ← table(not_ebf = e$not_ebf, diarrhoea = e$diarrhoea)
tab
prop.test(x = tab[c("1", "0"), "1"], n = rowSums(tab)[c("1", "0")],
          correct = FALSE)
```

```
       diarrhoea
not_ebf   0   1
      0 106  14
      1  47  33
...
X-squared = 23.367, df = 1, p-value = 1.338e-06
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1736206 0.4180461
sample estimates:
   prop 1    prop 2
0.4125000 0.1166667
```

The table gives the counts in step 4 (row `1` = not EBF, column `1` = diarrhoea). `X-squared` = 23.37 = 4.83^2, the z of step 4; the p-value is step 5; the CI and the two risks are the numbers reported in step 7.

MUST-KNOW NOTES (MD-LEVEL)

- Textbooks number the steps differently; the content is the same.
- Deciding steps 1–3 after seeing the results (picking whichever test or direction gives $p < 0.05$) is research malpractice.
- Step 7 is where most reports fall short: "$p < 0.05$" is not a conclusion. Give the effect size, its CI, and the exact p.

8.8 Choosing the Right Statistical Test

DEFINITION

Choosing the right test means matching step 3 to five features of the study: the type of outcome, what is being compared (one group, two groups, more than two, or an association between two variables), whether observations are independent, paired/matched or clustered, whether the distributional assumptions of a parametric test hold, and the study design that produced the data.

SIMPLIFIED EXPLANATION

Recall §1.5: the design decides which measures the data can give (prevalence from a cross-sectional survey, risk and rate from a cohort, the odds ratio of exposure from a case-control study). The outcome type

decides the family of methods. The pairing or clustering decides the version within the family. Answer five questions in order:

1. What type is the outcome: quantitative (roughly normal, or skewed), ordinal, binary, categorical, a count or rate, or a time to an event (recall §1.1.2)?
2. What is compared: one group with a fixed value, two groups, more than two groups, or two variables measured on the same people?
3. Are the observations independent, paired or matched (the same person twice, or matched case–control pairs), or clustered (children within villages, repeated visits)?
4. For a quantitative outcome, are the parametric assumptions reasonable (recall the Central Limit Theorem, §6.6), or is a rank-based method safer?
5. Which design produced the data, and therefore which effect measure is valid?

ILLUSTRATION / EXAMPLE

Mean birthweight of babies of smokers and non-smokers, birthweight roughly normal: quantitative outcome, two independent groups, parametric assumptions reasonable → the independent-samples t-test (§9.2.2). If birthweight were heavily skewed with outliers → the Mann-Whitney U test (§10.2, a test on ranks) instead, or a transformation (§9.4). If the same data came from a cluster survey of 30 villages → the same comparison, but with an SE corrected for clustering (§6.2.6, §15.2).

The tables below are the map of the rest of the book. The tests and models in them are defined in the sections cited; for now read each name as a label. The last column of each table gives the regression model that does the same job while adjusting for other variables.

Table 8.1. Quantitative and ordinal outcomes.

Comparison	Independent observations: measure · test	Paired or matched	Regression model (adjusting)
One group vs a fixed value	Mean (SD), CI §7.2.1 · one-sample t §9.2.1 (z §9.1 if σ known)	–	–
Two groups, roughly normal	Difference in means, CI §7.2.3 · independent t §9.2.2 (Welch if SDs differ)	Mean of within-pair differences · paired t §9.2.3	Linear regression with a group term §12.2, §12.4.3
Two groups, skewed or ordinal	Medians (IQR) §2.3, §3.4 · Mann-Whitney §10.2	Median difference · Wilcoxon signed-rank §10.3 (needs roughly symmetric differences); sign test §10.1.1 for strictly ordinal scores or skewed differences	Transform and use linear regression §9.4; ordinal logistic §13.2.7
Three or more groups, roughly normal	Group means · one-way ANOVA §9.5.2, then post-hoc tests §9.5.3	Repeated-measures ANOVA §9.5.5	Linear regression with a categorical predictor §12.4.3; mixed model §15.2.4
Three or more groups, skewed or ordinal	Group medians · Kruskal-Wallis §10.4	Friedman §10.5	Ordinal logistic §13.2.7
Two quantitative variables	Scatter §4.4, r §11.2 (Spearman §11.3 if skewed or ordinal) · test of r §11.2	–	Simple and multiple linear regression §12.2, §12.4

Table 8.2. Binary and categorical outcomes.

Design and comparison	Measure	Test	Regression model (adjusting)
One group vs a target (e.g. coverage 80%)	Proportion, CI §7.2.2, §7.2.4	One-proportion z §9.1.1; exact binomial §10.1.1	–
Cross-sectional, two or more groups	Prevalence, prevalence difference or ratio §10.6.4	Two-proportion z §9.1.2; chi-square §10.6.1; Fisher's exact §10.7 when counts are small	Logistic (odds ratio) §13.2; log-binomial or modified Poisson (prevalence ratio) §13.7
Cohort or RCT, two groups	Risk, risk difference, risk ratio, NNT §10.6.4	As above	Logistic §13.2; modified Poisson (risk ratio) §13.7
Unmatched case-control	Odds ratio of exposure §10.6.4	Chi-square §10.6.1; Fisher §10.7	Logistic §13.2.5
Matched case-control, or the same people twice	Matched-pair odds ratio b/c §10.8	McNemar §10.8	Conditional logistic §13.2.9
Ordered exposure groups	Risks across groups	Chi-square for trend §10.6.5	Logistic with a trend term §12.4.5, §13.2
Stratified by a confounder	Stratum-specific and pooled OR or RR	Mantel–Haenszel §11.7	Multivariable logistic §13.2.5
Outcome with 3+ unordered categories	Proportions per category	Chi-square r × c §10.6.1	Multinomial logistic §13.2.8
Observed vs expected distribution	Proportions	Chi-square goodness of fit §10.6.2	–

Table 8.3. Counts, rates, time to event, and measurement studies.

Outcome and design	Measure	Test	Regression model (adjusting)
Events per person-time (cohort, surveillance)	Rate, rate ratio, SMR §7.4, CI §7.4.4	Test and CI of the rate ratio §7.4.4	Poisson with offset §13.4–13.5; negative binomial if over-dispersed §13.6
Time to event with censoring (cohort, RCT)	Kaplan–Meier survival, median survival §14.4	Log-rank §14.5	Cox proportional hazards §14.6; stratified Cox and time-varying covariates §14.6.3; competing risks §14.8
Agreement between raters	Cohen's kappa §16.3.3.1	CI of kappa §16.3.3.1	–
Agreement between two measurement methods	Mean difference, limits of agreement §16.3.5	Bland–Altman §16.3.5	–
Diagnostic accuracy	Sensitivity, specificity §5.5.1; ROC, AUC §15.5.1	CI for sensitivity and specificity §7.2.4; AUC CI §15.5.1	Prediction models §15.5.2

Table 8.4. The third column in every table: independent, paired or clustered.

Structure of the data	Consequence	Where
Independent individuals (simple random sample, individually randomised trial)	The ordinary tests and models above	Chapters 9–14

Structure of the data	Consequence	Where
Pairs (same person before and after, matched pairs)	Analyse within-pair differences or use matched methods	§9.2.3, §10.3, §10.8, §13.2.9
Clusters (households within villages, cluster-randomised trials, survey PSUs)	SE too small if ignored; inflate by the design effect or model the clusters	§6.2.6, §15.2.2, GEE §15.2.5, mixed models §15.2.4, survey analysis §15.2.6
Repeated measures over time on the same people	Correlated observations	§9.5.5, §15.2.4, §15.2.5

MUST-KNOW NOTES (MD-LEVEL)

- The design comes first. A case-control study gives the odds ratio of exposure, not a risk ratio (recall §1.5), whatever test is run on it.
- Paired and clustered data break the independence assumption of the ordinary tests. Ignoring pairing usually wastes power; ignoring clustering gives SEs and p-values that are too small.
- "Non-normal data, so use a non-parametric test" is too simple. With large samples the t-test is robust through the Central Limit Theorem (recall §6.6), and the choice depends also on whether the mean or the median answers the question. §10.1 (when to use non-parametric tests) sets out the choice.
- A correctly computed test on the wrong design is still wrong. Choosing the method to fit the design and outcome matters as much as the arithmetic.

CHECKPOINT 8.3

1. Name the five questions you ask to choose a statistical test, and say why the design question comes first.
2. Using Tables 8.1–8.4, choose the test and the adjusting model for: (a) haemoglobin before and after iron–folic acid in the same 12 women, with skewed differences; (b) a case-control study of neonatal sepsis with each case matched to a neighbourhood control, exposure "home delivery"; (c) full-immunisation coverage compared between urban and rural children in a 30-cluster survey.
3. In a cohort, 30 of 150 exposed and 15 of 150 unexposed people develop the outcome. Following the seven steps of §8.7, compute the risk difference, the pooled SE under H_0, z and the two-sided p, and the 95% CI of the difference (unpooled SE). State the conclusion.

(Answers in the Appendix at the end of the book.)

Comparing Groups – Parametric Tests

This chapter puts the testing framework of Chapter 8 to work on the commonest research question: do these groups differ? It starts with z-tests for a mean and for proportions (§9.1), moves to the t-test family for means (§9.2), checks the Normality assumption these tests lean on and what to do when it fails (§9.3–§9.4), and ends with ANOVA for three or more groups, including how to handle many comparisons at once (§9.5). Every test here is "parametric": it assumes the data, or the sampling distribution of the statistic, is approximately Normal. Chapter 10 gives the alternatives when that assumption is not defensible.

9.1 z-test

DEFINITION

The **z-test** compares a sample statistic with a hypothesised value by asking how many standard errors apart they are, and reads the p-value from the standard Normal distribution. For a mean with known population SD σ,

$z = (\bar{x} - \mu_0) \div (\sigma \div \sqrt{n})$.

SIMPLIFIED EXPLANATION

This is a z-score (recall §5.7.2.2.1) computed for a sample statistic instead of a single child's value. Under H_0 the sample mean is centred on μ_0 with SD $\sigma/\sqrt{n}$ (recall §6.7), and by the Central Limit Theorem (§6.6) its sampling distribution is close to Normal. So z behaves like a standard Normal variable when H_0 is true, and a large $|z|$ is evidence against H_0.

ILLUSTRATION / EXAMPLE

A district's sample of 100 newborns has mean birthweight 2.9 kg. Test it against a hypothesised reference value of 3.0 kg, assuming σ = 0.5 kg is known from a large earlier survey.

WORKED CALCULATION

1. SE = $0.5 \div \sqrt{100} = 0.05$ kg.
2. $z = (2.9 - 3.0) \div 0.05 = \mathbf{-2.00}$.
3. Two-sided $p = 2 \times P(Z < -2.00) = 2 \times 0.0228 = \mathbf{0.046}$.

4. At $\alpha = 0.05$, reject H_0: a mean 2 SEs below 3.0 kg would be unusual if the true mean were 3.0 kg (recall the steps of §8.7).

MUST-KNOW NOTES (MD-LEVEL)

- The general form is z = (statistic − hypothesised value) ÷ SE. It needs either a known σ (rare) or a sample large enough that the Normal approximation holds. "n ≥ 30" is a convention, not a law: strongly skewed data need more (recall §6.6).
- When σ is estimated by s, use the t-test (§9.2.1) at any sample size (recall §5.7.2.3).
- For proportions σ is not a separate unknown: under H_0 it is fixed by p_0. That is why the z-test remains a standard large-sample test for proportions (§9.1.1–§9.1.2).

9.1.1 One-Proportion z-test

DEFINITION

The **one-proportion z-test** compares an observed proportion $\hat{p} = x/n$ with a hypothesised proportion p_0:

$z = (\hat{p} - p_0) \div \sqrt{[p_0(1 - p_0) \div n]}$.

When n is small, the **exact binomial test** replaces it and computes the p-value directly from the Binomial(n, p_0) distribution.

SIMPLIFIED EXPLANATION

Each person contributes a 0/1 outcome, a Bernoulli variable with variance $p(1 - p)$ (recall §5.7.1.1). The proportion $\hat{p}$ is the mean of these 0s and 1s, so its SE is $\sqrt{[p(1 - p)/n]}$ (recall §6.7). The test assumes H_0 is true, so it puts p_0, not $\hat{p}$, into the SE. (The confidence interval of §7.2.2 uses $\hat{p}$, because it makes no assumption about the true value.) The Normal approximation is adequate when np_0 and $n(1 - p_0)$ are both at least 5, a common convention (recall §5.7.2.2.4). Below that, count the binomial probabilities exactly.

ILLUSTRATION / EXAMPLE

Recall the infant cohort of §8.7 (200 infants followed up at two PHCs; the same data return in §10.6.3, §10.6.4 and §11.7): 120 were exclusively breastfed (EBF). The district's planning target, used here as a hypothetical H_0, is 70%.

WORKED CALCULATION

1. $\hat{p} = 120/200 = 0.60$. Check: $np_0 = 140$ and $n(1 - p_0) = 60$, both ≥ 5.
2. SE under $H_0 = \sqrt{(0.70 \times 0.30 \div 200)} = \sqrt{0.00105} = 0.0324$.
3. $z = (0.60 - 0.70) \div 0.0324 =$ **−3.09**; two-sided p = **0.002**. EBF coverage is below the target.

Small sample: one anganwadi. In one anganwadi, 7 of 15 infants were exclusively breastfed. Test against $p_0 = 0.70$.

1. Check: $np_0 = 10.5$ but $n(1 - p_0) = 4.5 < 5$. The Normal approximation is doubtful.

2. The exact test uses Binomial(15, 0.70). P(X = 7) = 0.0348. The two-sided exact p-value adds the probabilities of every outcome no more likely than the one observed: P(X ≤ 7) = 0.0500, plus P(X = 14) = 0.0305 and P(X = 15) = 0.0047, which gives **p = 0.085**.
3. The z approximation gives z = (0.467 − 0.70) ÷ √(0.21/15) = −1.97 and p = 0.049. It would call the result "significant"; the exact test does not. With small n, trust the exact test.

IN R (OPTIONAL)

```
eb ← read.csv("data/ebf_diarrhoea_cohort.csv")
x ← sum(eb$not_ebf == 0); n ← nrow(eb); c(x = x, n = n)
prop.test(x, n, p = 0.70, correct = FALSE)
binom.test(7, 15, p = 0.70)
```

```
  x   n
120 200

    1-sample proportions test without continuity correction
X-squared = 9.5238, df = 1, p-value = 0.002028
...
    Exact binomial test
number of successes = 7, number of trials = 15, p-value = 0.08528
95 percent confidence interval:
 0.2126667 0.7341387
```

`prop.test` reports X-squared = 9.52, which is $z^2 = (-3.086)^2$; its p-value, 0.002, is the z-test's. `binom.test` gives the exact p = 0.085 for the anganwadi and an exact 95% CI (recall §7.2.4) of 21% to 73%, which includes the 70% target.

MUST-KNOW NOTES (MD-LEVEL)

- SE in the test uses p_0; SE in the CI uses $\hat{p}$. With $\hat{p}$ far from p_0 the two differ.
- Check np_0 and $n(1 - p_0)$ before trusting z. The exact binomial test is widely available in statistical software and is the usual choice for small n; the sign test (§10.1.1) is this same test applied to the signs of paired differences.
- A survey that sampled villages rather than individuals needs an SE that allows for clustering (recall DEFF, §6.2.6); the formula above assumes independent individuals.

9.1.2 Two-Proportion z-test (and $z^2 = \chi^2$)

DEFINITION

The **two-proportion z-test** compares proportions $\hat{p}_1 = x_1/n_1$ and $\hat{p}_2 = x_2/n_2$ from two independent groups:

$z = (\hat{p}_1 - \hat{p}_2) \div \sqrt{[\bar{p}(1 - \bar{p})(1/n_1 + 1/n_2)]}$, where $\bar{p} = (x_1 + x_2) \div (n_1 + n_2)$

is the **pooled proportion**. For a 2×2 table, z^2 equals the Pearson chi-square statistic, so the two-sided z-test and the chi-square test of independence (§10.6.1, which compares observed and expected counts in each cell) are the same test.

SIMPLIFIED EXPLANATION

H_0 says both groups share one true proportion. The best estimate of that common value pools both groups, $\bar{p}$, and the SE under H_0 uses it. The confidence interval for the difference makes no such assumption, so it uses the separate, unpooled SE of §7.2.3. The same pattern as §9.1.1 holds: test under H_0, estimate without it.

DERIVATION

SE. Variances of independent estimates add (recall §7.2.3): $\text{Var}(\hat{p}_1 - \hat{p}_2) = p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2$. Under H_0, $p_1 = p_2 = p$, so this becomes $p(1 - p)(1/n_1 + 1/n_2)$, estimated with $\bar{p}$.

$z^2 = \chi^2$. Write $N = n_1 + n_2$, and m_1, m_2 for the column totals (with and without the outcome). The expected count for a cell is row total × column total ÷ N. For the top-left cell, $O - E = x_1 - n_1m_1/N = n_1n_2(\hat{p}_1 - \hat{p}_2)/N$, and every cell has the same $|O - E|$. Then

$$\chi^2 = \Sigma(O - E)^2/E = (O - E)^2 \times N^3/(n_1n_2m_1m_2) = N\, n_1n_2(\hat{p}_1 - \hat{p}_2)^2 \div (m_1m_2),$$

since $\Sigma 1/E = N(1/n_1 + 1/n_2)(1/m_1 + 1/m_2)$. Put $\bar{p} = m_1/N$ and $1 - \bar{p} = m_2/N$ into z^2: its denominator $\bar{p}(1 - \bar{p})(1/n_1 + 1/n_2) = m_1m_2/(N\, n_1n_2)$, which gives the same expression. So $z^2 = \chi^2$ with 1 degree of freedom, as $z^2 = \chi^2_1$ requires (recall §5.7.2.5).

ILLUSTRATION / EXAMPLE

The same 200-infant cohort, by feeding and diarrhoea during follow-up:

	Diarrhoea	No diarrhoea	Total
Not EBF	33	47	80
EBF	14	106	120
Total	47	153	200

WORKED CALCULATION

1. $\hat{p}_1 = 33/80 = 0.4125$ (not EBF); $\hat{p}_2 = 14/120 = 0.1167$ (EBF). Difference = 0.296.
2. $\bar{p} = 47/200 = 0.235$. $SE_0 = \sqrt{[0.235 \times 0.765 \times (1/80 + 1/120)]} = \sqrt{(0.1798 \times 0.02083)} = 0.0612$.
3. $z = 0.296 \div 0.0612 =$ **4.83**; two-sided $p = 1.3 \times 10^{-6}$ (the same z as in the steps of §8.7).
4. 95% CI for the difference, unpooled SE (recall §7.2.3): $\sqrt{(0.4125 \times 0.5875/80 + 0.1167 \times 0.8833/120)} = 0.0624$; $0.296 \pm 1.96 \times 0.0624 =$ **17.4 to 41.8 percentage points**.
5. Chi-square on the same table. Expected counts: $80 \times 47/200 = 18.8$, 61.2, 28.2 and 91.8. Every $|O - E| = 14.2$. $\chi^2 = 14.2^2 \times (1/18.8 + 1/61.2 + 1/28.2 + 1/91.8) = 201.64 \times 0.1159 =$ **23.37** $= 4.834^2$.

IN R (OPTIONAL)

```
eb ← read.csv("data/ebf_diarrhoea_cohort.csv")
tab ← table(not_ebf = factor(eb$not_ebf, levels = c(1, 0)),
            diarrhoea = factor(eb$diarrhoea, levels = c(1, 0)))
tab
prop.test(tab, correct = FALSE)
```

```
       diarrhoea
not_ebf   1   0
      1  33  47
      0  14 106

    2-sample test for equality of proportions without continuity
    correction
X-squared = 23.367, df = 1, p-value = 1.338e-06
95 percent confidence interval:
 0.1736206 0.4180461
sample estimates:
   prop 1    prop 2
0.4125000 0.1166667
```

The table reproduces the counts. `X-squared = 23.367` is z^2 (step 5), and the p-value matches step 3. The CI (0.174 to 0.418) is the unpooled interval of step 4. R reports the test as a chi-square; it is the same test as the z-test.

MUST-KNOW NOTES (MD-LEVEL)

- Pooled SE for the test, unpooled SE for the CI. The z-test and the uncorrected 2×2 chi-square give identical two-sided p-values; only z can be one-sided.
- The difference, 29.6 percentage points, is the risk difference. The risk ratio and odds ratio for the same table come in §10.6.4.
- This is an association in an observational cohort, not an effect of breastfeeding: mothers who breastfeed exclusively may differ in other ways, including which PHC they attend. §11.7 (stratified analysis) shows that PHC explains part of the association.
- With small expected counts use Fisher's exact test (§10.7, which computes exact probabilities for the 2×2 table).

CHECKPOINT 9.1

1. In a block, 45 of 60 sampled children are fully immunised. Test against a target of 85%. Check the Normal-approximation condition first, then give z and the two-sided p-value.
2. ORS use during the last diarrhoea episode was 60/100 in Block A and 45/100 in Block B. Compute the two-proportion z statistic, then the Pearson χ^2 from the expected counts, and show that $\chi^2 = z^2$.
3. Why does the one-proportion test use p_0 in the SE while the confidence interval uses $\hat{p}$?

(Answers in the Appendix at the end of the book.)

9.2 t-test

The t-test family compares means when σ is unknown and estimated by s, so the test statistic follows the t-distribution (recall §5.7.2.3). There are three forms: one mean against a fixed value (§9.2.1), two independent groups (§9.2.2) and two paired measurements (§9.2.3). All share the logic "difference ÷ SE, compared with t on the right degrees of freedom", and one set of assumptions (§9.2.4).

9.2.1 One-Sample t-test

DEFINITION

The **one-sample t-test** compares a sample mean with a hypothesised value μ_0:

$t = (\bar{x} - \mu_0) \div (s \div \sqrt{n})$, with $n - 1$ degrees of freedom.

SIMPLIFIED EXPLANATION

This is the z-test of §9.1 with s in place of σ. Estimating σ adds uncertainty, and the heavier-tailed t-distribution pays for it (recall §5.7.2.3). The $n - 1$ degrees of freedom are those of s itself (recall Bessel's correction, §3.6.2–§3.6.3).

ILLUSTRATION / EXAMPLE

A second district samples 25 newborns: mean 2.9 kg, s = 0.4 kg. Test against the same hypothesised reference value, 3.0 kg.

WORKED CALCULATION

1. SE = 0.4 ÷ √25 = 0.08 kg.
2. $t = (2.9 - 3.0) \div 0.08 =$ **−1.25**, with 24 df.
3. The 97.5% point of t_{24} is 2.064; $|t| < 2.064$. Two-sided p = **0.22**.
4. 95% CI: 2.9 ± 2.064 × 0.08 = **2.73 to 3.07 kg**.

COMMON MISREADING

"p = 0.22, so this district's mean is 3.0 kg." Compare §9.1. Both districts show the same 0.1 kg shortfall. In §9.1, SE = 0.05 kg and p = 0.046; here SE = 0.08 kg and p = 0.22. The difference did not change; only its precision did, mainly because n fell from 100 to 25 (the SE shrinks with √n). Had this district measured 100 newborns with the same s = 0.4 kg, SE would be 0.04 kg, t = −2.5 and p = 0.014. The CI tells the full story: 2.73 to 3.07 kg is compatible with no shortfall and with a shortfall of 0.27 kg. A non-significant p-value means the study could not tell, not that there is no difference (recall §8.5).

MUST-KNOW NOTES (MD-LEVEL)

- $t = (\bar{x} - \mu_0) \div (s/\sqrt{n})$, $df = n - 1$.
- Assumes independent observations from a roughly Normal population; for larger n the Central Limit Theorem relaxes the Normality requirement (§9.2.4).

- The p-value and the CI come from the same t-distribution: $p < 0.05$ exactly when the 95% CI excludes μ_0.

9.2.2 Independent Samples t-test

DEFINITION

The **independent samples t-test** compares the means of two unrelated groups. The **pooled (Student's) version** assumes equal population variances:

$s_p^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2] \div (n_1 + n_2 - 2)$, $t = (\bar{x}_1 - \bar{x}_2) \div [s_p \sqrt{(1/n_1 + 1/n_2)}]$, $df = n_1 + n_2 - 2$.

Welch's version does not assume equal variances:

$t = (\bar{x}_1 - \bar{x}_2) \div \sqrt{(s_1^2/n_1 + s_2^2/n_2)}$, $df = (v_1 + v_2)^2 \div [v_1^2/(n_1 - 1) + v_2^2/(n_2 - 1)]$, where $v_1 = s_1^2/n_1$ and $v_2 = s_2^2/n_2$.

The Welch df is the **Welch–Satterthwaite approximation**; it is usually not a whole number.

SIMPLIFIED EXPLANATION

The numerator is the observed difference; the denominator is the SE of that difference, built by adding the two groups' variances (recall §7.2.3). If the two populations share one variance, the best estimate of it averages both groups' s^2, weighted by their degrees of freedom: that is s_p^2. If they do not, keep each group's own s^2/n, as Welch does. Welch then shrinks the df to reflect that the SE is built from two separately estimated variances.

DERIVATION

1. $\text{Var}(\bar{x}_1 - \bar{x}_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2$ for independent groups (recall §7.2.3).
2. If $\sigma_1 = \sigma_2 = \sigma$, this is $\sigma^2(1/n_1 + 1/n_2)$. Each group's sum of squares, $(n_i - 1)s_i^2$, estimates $(n_i - 1)\sigma^2$ (recall §3.6.2), so pooling them and dividing by the total df, $n_1 + n_2 - 2$, gives s_p^2, the estimate of σ^2.
3. Without equal variances, the estimated variance $v_1 + v_2$ is not a scaled chi-square variable, so t is not exactly t-distributed. Satterthwaite matched its mean and variance to a chi-square with ν df and solved for ν, which gives the Welch df above (proof beyond this book).
4. The Welch df always lies between the smaller of $n_1 - 1$ and $n_2 - 1$ and $n_1 + n_2 - 2$.

ILLUSTRATION / EXAMPLE

Recall the urban and rural children of §7.2.3: urban $n_1 = 40$, mean 12 kg, SD 2.0 kg; rural $n_2 = 50$, mean 11 kg, SD 1.8 kg. Test H_0: no difference in mean weight.

WORKED CALCULATION

Pooled.

1. $s_p^2 = (39 \times 4.00 + 49 \times 3.24) \div 88 = (156.0 + 158.8) \div 88 = 3.577$; $s_p = 1.891$ kg.
2. $SE = 1.891 \times \sqrt{(1/40 + 1/50)} = 1.891 \times 0.2121 = 0.401$ kg.

3. t = 1 ÷ 0.401 = **2.49**, df = 88; t_{88} (97.5%) = 1.987. p = **0.015**. 95% CI: 1 ± 1.987 × 0.401 = 0.20 to 1.80 kg.

Welch.

1. v_1 = 4.00/40 = 0.100; v_2 = 3.24/50 = 0.0648. SE = √0.1648 = 0.406 kg (the SE of §7.2.3).
2. t = 1 ÷ 0.406 = **2.46**.
3. df = 0.1648^2 ÷ (0.100^2/39 + 0.0648^2/49) = 0.02716 ÷ (0.000256 + 0.0000857) = **79.4**. p = **0.016**. 95% CI: 1 ± 1.990 × 0.406 = 0.19 to 1.81 kg.

The two versions agree closely here because the SDs are similar. They part company when the variances differ and the group sizes differ too.

IN R (OPTIONAL)

```
cl ← read.csv("data/clinic_children.csv")
aggregate(weight_kg ~ sex, data = cl,
          FUN = function(x) c(n = length(x), mean = mean(x),
                              sd = sd(x)))
t.test(weight_kg ~ sex, data = cl, var.equal = TRUE)
t.test(weight_kg ~ sex, data = cl)
```

```
  sex weight_kg.n weight_kg.mean weight_kg.sd
1   F   50.000000      12.150000     1.559598
2   M   50.000000      12.154000     1.339389

    Two Sample t-test
t = -0.013758, df = 98, p-value = 0.9891
95 percent confidence interval:
 -0.580952  0.572952

    Welch Two Sample t-test
t = -0.013758, df = 95.814, p-value = 0.9891
95 percent confidence interval:
 -0.5811165  0.5731165
```

Girls and boys in the under-five clinic weigh 12.150 and 12.154 kg on average. `var.equal = TRUE` gives the pooled test with df = 50 + 50 − 2 = 98. The default, `t.test()` without that argument, is Welch: the same t (with equal group sizes the two SEs are identical) but df = 95.8. The 95% CI, −0.58 to 0.57 kg, is the useful result: any sex difference in mean weight in this clinic is under about 0.6 kg either way.

MUST-KNOW NOTES (MD-LEVEL)

- Use Welch by default. Delacre, Lakens and Leys (2017) argue that "Welch's t-test should be used as a default strategy": it loses little when variances are equal and protects the Type I error rate when they are not, especially with unequal group sizes. R, for one, already uses Welch by default.
- With $n_1 = n_2$ the pooled and Welch t statistics are identical; only the df differ.
- The groups must be independent. For the same people measured twice, or for matched pairs, use §9.2.3.

9.2.3 Paired t-test

DEFINITION

The **paired t-test** compares two related measurements by computing the difference within each pair, d, and running a one-sample t-test of mean(d) against 0: $t = \bar{d} \div (s_d/\sqrt{n})$, with $n - 1$ df, where n is the number of pairs.

SIMPLIFIED EXPLANATION

Two measurements on the same person are correlated: a woman with low haemoglobin before treatment tends to have relatively low haemoglobin after it. The paired test uses this. Reducing each pair to its difference removes the stable between-person variation, so s_d is usually much smaller than the SD of either measurement, and the test is more powerful.

ILLUSTRATION / EXAMPLE

A synthetic example (the numbers are made up for teaching, not taken from a study): ten adolescent girls have haemoglobin measured before and after 12 weeks of iron–folic acid (IFA) supplementation. The data are in `data/paired_hb_ifa.csv`.

Girl	1	2	3	4	5	6	7	8	9	10
Before (g/dL)	10.2	11.4	9.8	12.1	10.9	11.0	9.5	11.8	10.6	11.3
After (g/dL)	11.1	11.9	10.9	12.4	11.5	12.0	10.1	12.3	11.7	11.8
d = after − before	0.9	0.5	1.1	0.3	0.6	1.0	0.6	0.5	1.1	0.5

WORKED CALCULATION

1. Differences: $\Sigma d = 7.1$, so $\bar{d} = 7.1 \div 10 =$ **0.71 g/dL**.
2. $\Sigma(d - \bar{d})^2 = 0.749$, so $s_d^2 = 0.749 \div 9 = 0.0832$ and $s_d =$ **0.288 g/dL**.
3. $SE = s_d \div \sqrt{n} = 0.288 \div \sqrt{10} =$ **0.091 g/dL**.
4. $t = 0.71 \div 0.091 =$ **7.78**, $df = 10 - 1 = 9$. t_9 (97.5%) = 2.262, so $p =$ **2.8×10^{-5}**.
5. 95% CI: $0.71 \pm 2.262 \times 0.091 = 0.71 \pm 0.21 =$ **0.50 to 0.92 g/dL**.
6. For contrast, wrongly treat the columns as two independent groups (§9.2.2): means 10.86 and 11.57 g/dL, SDs 0.844 and 0.701 g/dL, $s_p^2 = (0.7116 + 0.4912) \div 2 = 0.601$, $SE = \sqrt{(0.601 \times 2/10)} = 0.347$ g/dL, $t = 0.71 \div 0.347 =$ **2.05**, $df = 18$, $p =$ **0.056**, 95% CI −0.02 to 1.44 g/dL.

The mean change is the same 0.71 g/dL in both analyses, but pairing gains power because the girls' starting levels vary a lot (SD 0.84 g/dL) while their changes vary little (s_d = 0.29 g/dL), and only the independent test leaves that between-girl spread in the SE, which is almost four times larger (0.347 vs 0.091 g/dL).

IN R (OPTIONAL)

```
hb ← read.csv("data/paired_hb_ifa.csv")   # synthetic teaching data
t.test(hb$hb_after, hb$hb_before, paired = TRUE)
# the wrong, unpaired analysis:
t.test(hb$hb_after, hb$hb_before, var.equal = TRUE)
```

```
    Paired t-test
t = 7.7829, df = 9, p-value = 2.756e-05
95 percent confidence interval:
 0.503632 0.916368
sample estimates:
mean difference
           0.71

    Two Sample t-test
t = 2.0472, df = 18, p-value = 0.05552
95 percent confidence interval:
 -0.0186226  1.4386226
```

With `paired = TRUE`, R subtracts the second vector from the first girl by girl, so the "mean difference" is $\bar{d} = 0.71$ g/dL (step 1), t = 7.78 on 9 df (step 4) and the CI 0.50 to 0.92 g/dL (step 5). The unpaired call on the same numbers gives t = 2.05 on 18 df and a CI that crosses 0 (step 6). The order of the vectors sets the sign: `t.test(hb$hb_before, hb$hb_after, paired = TRUE)` gives −0.71.

MUST-KNOW NOTES (MD-LEVEL)

- The unit of analysis is the pair. Use it for the same person at two time points, the two eyes of one patient, or **matched pairs**: two different people deliberately paired because they share characteristics such as age, sex and village.
- Analysing paired data with an independent-samples test ignores the correlation, inflates the SE and wastes power.
- The paired test assumes the *differences* are roughly Normal; the two raw measurements need not be. For clearly non-Normal but roughly symmetric differences, see the Wilcoxon signed-rank test (§10.3, a rank-based test on the differences); for skewed differences or ordinal scores, the sign test (§10.1.1, which uses only the direction of each change).

9.2.4 Assumptions of the t-test

DEFINITION

The t-tests assume (1) independent observations (apart from the deliberate pairing of §9.2.3), (2) approximate Normality of the data, or of the differences for the paired test, and (3) for the pooled independent-samples test only, equal population variances. Equal variances can be checked with the **variance-ratio F-test**, $F = s_1^2/s_2^2$ with $(n_1 - 1, n_2 - 1)$ df, or with **Levene's test**.

SIMPLIFIED EXPLANATION

Independence is a design property: check how the data were collected. Normality matters most for small samples; as n grows the Central Limit Theorem makes the sampling distribution of the mean close to Normal anyway (recall §6.6). Equal variance matters for the pooled test only, and Welch's test (§9.2.2) removes the need for it.

DERIVATION

For two Normal samples, $(n_i - 1)s_i^2/\sigma_i^2$ is chi-square with $n_i - 1$ df (recall §5.7.2.3). Dividing each by its df and taking the ratio gives an F variable (recall §5.7.2.5): $F = (s_1^2/\sigma_1^2) \div (s_2^2/\sigma_2^2)$. Under H_0: $\sigma_1 = \sigma_2$ the σ's cancel and $F = s_1^2/s_2^2$ with $(n_1 - 1, n_2 - 1)$ df.

ILLUSTRATION / EXAMPLE

Are the urban and rural variances of §9.2.2 different?

WORKED CALCULATION

1. Put the larger variance on top: $F = 2.0^2 \div 1.8^2 = 4.00 \div 3.24 =$ **1.23**, with (39, 49) df.
2. For a two-sided 5% test, compare with the upper 2.5% point of F(39, 49), which is 1.81. $F = 1.23 < 1.81$.
3. Two-sided p = **0.48**. No evidence that the variances differ.

Levene's test measures each observation's absolute distance from its group centre and runs a one-way ANOVA (§9.5) on those distances: if one group is more spread out, its distances are larger on average. The median-centred version (often called Brown–Forsythe) is the recommended form: "the definition based on the median is recommended as the choice that provides good robustness against many types of non-normal data while retaining good power" (NIST/SEMATECH, n.d.).

IN R (OPTIONAL)

```
cl ← read.csv("data/clinic_children.csv")
var.test(weight_kg ~ sex, data = cl)
car::leveneTest(weight_kg ~ factor(sex), data = cl)
```

```
    F test to compare two variances
F = 1.3559, num df = 49, denom df = 49, p-value = 0.29
95 percent confidence interval:
 0.7694131 2.3892651

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.7666 0.3834
      98
```

`var.test` gives $F = 1.5596^2 \div 1.3394^2 = 1.36$ (girls' variance over boys'), with p = 0.29. `leveneTest` uses the median-centred version by default and agrees (p = 0.38). Neither suggests unequal variances, which is why the pooled and Welch results in §9.2.2 were nearly identical.

MUST-KNOW NOTES (MD-LEVEL)

- **Independence** is the assumption no test can rescue. Children from the same household or village, or repeated visits by one patient, are not independent (recall clustering, §6.2.6).
- **Normality.** Judge it with a Q-Q plot (§9.3) and, for small samples, a formal test. With heavily skewed data and small n, consider a log transformation (§9.4) or the Mann–Whitney U test (§10.2, a rank-based comparison of two groups).
- **Equal variances.** The variance-ratio F-test assumes Normal data and "is very sensitive to deviations from normality" (Illowsky & Dean, 2013), as is Bartlett's test for several variances (NIST/SEMATECH, n.d.). Levene's test "is less sensitive than the Bartlett test to departures from normality" (NIST/SEMATECH, n.d.). Testing variances first and then choosing between pooled and Welch is not recommended; using Welch from the start avoids the two-step procedure (Delacre et al., 2017).

CHECKPOINT 9.2

1. Explain why a paired t-test is generally more powerful than an independent samples t-test for detecting the same size of change.
2. Group 1: n = 12, mean SBP 128 mmHg, SD 10. Group 2: n = 15, mean 120 mmHg, SD 18. Compute the pooled t (with df) and the Welch t (with Welch–Satterthwaite df). Which SE is larger here, and why?
3. Sixteen children have mean haemoglobin 11.2 g/dL, s = 1.6 g/dL. Test against a hypothesised value of 12 g/dL (t_{15}, 97.5% point = 2.131). State t and your conclusion.
4. Nine pregnant women have haemoglobin measured before and after a month of supervised IFA. The mean change (after − before) is 0.45 g/dL with s_d = 0.60 g/dL. Compute the paired t, its df and the 95% CI (t_8, 97.5% point = 2.306), and state your conclusion.

(Answers in the Appendix at the end of the book.)

9.3 Tests of Normality

DEFINITION

Tests of normality assess whether data are consistent with a Normal distribution. The main visual tool is the **Normal quantile–quantile (Q-Q) plot**: the sorted data are plotted against the values a Normal distribution would give at the same cumulative proportions. Formal tests (§9.3.1–§9.3.2) test H_0: the data come from a Normal population.

SIMPLIFIED EXPLANATION

Skewness and kurtosis (recall §3.9–§3.10) summarise shape in one number each. A Q-Q plot shows *where* the shape departs from Normal. If the data are Normal, the points lie close to a straight line. A curve that bends away at one end shows skew; an S-shape shows tails heavier or lighter than Normal.

ILLUSTRATION / EXAMPLE

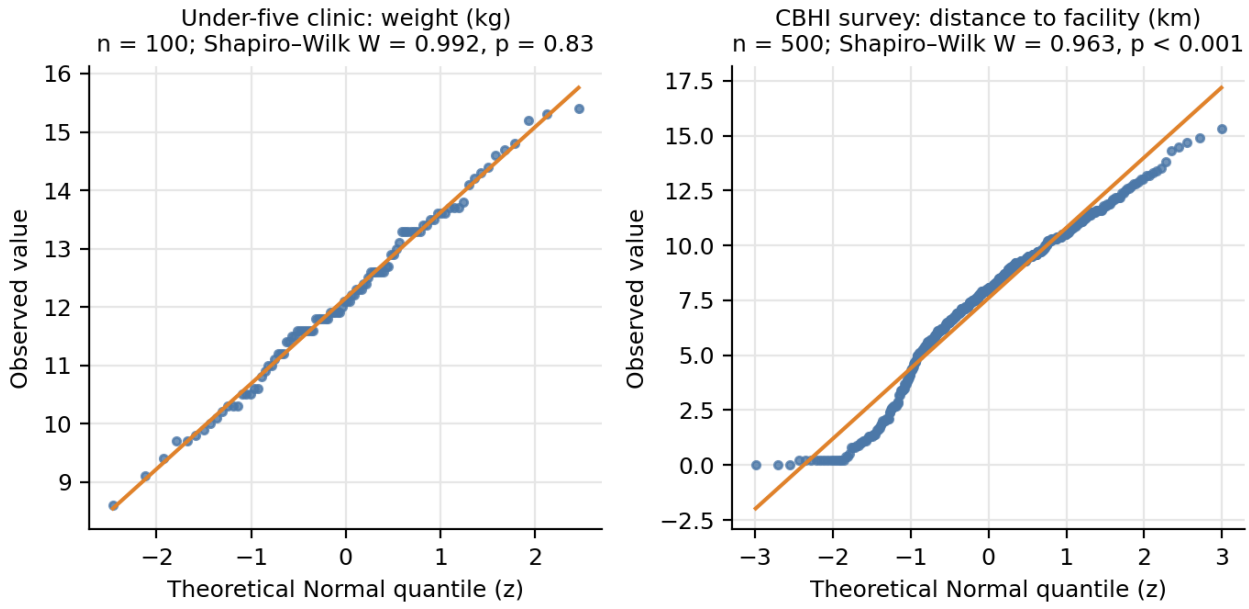


Figure 9.1. Normal Q-Q plots. Left: weights of the 100 children in the under-five clinic lie on the line. Right: distance to the nearest facility for the 500 CBHI households. The points fall below the line at the lower end: distances have a long lower tail that runs down to a floor at 0 km (24 households live within 1 km of a facility). The upper tail is shorter than Normal, so the distribution is skewed to the left.

MUST-KNOW NOTES (MD-LEVEL)

- The Q-Q plot is the primary check. It shows the kind and size of the departure, which a p-value cannot.
- A formal test's H_0 is Normality. A non-significant result does not prove Normality; it means the data could not rule it out (recall §8.5).
- With small samples the tests have low power and miss real departures; with large samples they flag departures too small to matter. The t-test itself becomes less dependent on Normality as n grows (recall §6.6). Use the plot and the size of the sample together, not the p-value alone.

9.3.1 Shapiro-Wilk Test

DEFINITION

The **Shapiro–Wilk test** measures how closely the sorted sample follows the pattern expected from a Normal sample. Its statistic W lies between 0 and 1; values near 1 are consistent with Normality.

SIMPLIFIED EXPLANATION

W is closely related to the squared correlation between the sorted data and the ideal Normal scores: a numerical version of how straight the Q-Q plot is. A small p-value says the plot bends more than chance would explain.

IN R (OPTIONAL)

```
cl ← read.csv("data/clinic_children.csv")
cb ← read.csv("data/cbhi_survey.csv")
shapiro.test(cl$weight_kg)
shapiro.test(cb$distance_km)
```

```
    Shapiro-Wilk normality test
data:  cl$weight_kg
W = 0.99212, p-value = 0.8299

    Shapiro-Wilk normality test
data:  cb$distance_km
W = 0.96269, p-value = 5.953e-10
```

Clinic weights: W = 0.992, p = 0.83, no evidence against Normality, matching the straight Q-Q plot. CBHI distances: W = 0.963 is still close to 1, but with n = 500 even this departure gives $p < 0.001$. The plot shows it comes from the long lower tail and the floor at 0 km. For a comparison of mean distance between two groups of this size, the Central Limit Theorem makes a t-test adequate despite the tiny p-value.

MUST-KNOW NOTES (MD-LEVEL)

- In a simulation comparison, "Shapiro-Wilk test is the most powerful normality test, followed by Anderson-Darling test, Lilliefors test and Kolmogorov-Smirnov test", but "the power of all four tests is still low for small sample size" (Razali & Wah, 2011).
- A small p-value with a large n often means a trivial departure (the CBHI example). Look at the plot.

9.3.2 Kolmogorov-Smirnov Test

DEFINITION

The **Kolmogorov–Smirnov (K-S) test** compares the sample's cumulative distribution with a fully specified theoretical cumulative distribution, using the largest vertical gap between the two curves (recall the ogive, §4.3.3).

SIMPLIFIED EXPLANATION

K-S is a general goodness-of-fit test for any continuous distribution; testing Normality is one use. Its catch: the comparison distribution must be fully specified in advance. The NIST/SEMATECH handbook calls this "perhaps the most serious limitation … if location, scale, and shape parameters are estimated from the data, the critical region of the K-S test is no longer valid", and adds that the test "tends to be more sensitive near the center of the distribution than at the tails".

MUST-KNOW NOTES (MD-LEVEL)

- When the mean and SD are estimated from the same sample, the usual situation, the **Lilliefors correction** gives valid p-values.
- For testing Normality, K-S is less powerful than Shapiro–Wilk (Razali & Wah, 2011). Prefer the Q-Q plot plus Shapiro–Wilk.

9.4 Data Transformation Techniques

DEFINITION

A **data transformation** re-expresses every value of a variable with the same function, such as the logarithm or square root, to reduce skewness, stabilise variance, or make a multiplicative relationship additive.

SIMPLIFIED EXPLANATION

When a variable fails the Normality checks, you can analyse it on a scale where it behaves better, or use a non-parametric test (Chapter 10). Transformation does not change the information in the data. It changes the ruler, and that changes what a "difference" means.

ILLUSTRATION / EXAMPLE

Hospital length of stay (LOS) is right-skewed (recall §3.9). On the log scale, a stay of 2 days becomes ln 2 = 0.69 and a stay of 32 days becomes ln 32 = 3.47. The long right tail is pulled in, and the log-LOS values are often close to Normal.

WORKED CALCULATION: BACK-TRANSFORMING A T-TEST ON LOGS

Mean ln(LOS) is 1.20 in ward A and 0.98 in ward B. A t-test on the logs gives a difference of 0.22 with 95% CI 0.05 to 0.39.

1. The exponential of the mean of logs is the geometric mean (recall §2.2.3): GM_A = e^1.20 and GM_B = e^0.98.
2. A difference of logs is the log of a ratio, so e^0.22 = GM_A ÷ GM_B = **1.25**. Ward A's geometric mean stay is 25% longer.
3. Back-transform the CI limits the same way: e^0.05 = 1.05 and e^0.39 = 1.48. The ratio of geometric means is 1.25 (95% CI 1.05 to 1.48).
4. The result is a ratio, not a difference in days: the back-transformed difference is not "0.22 days", and the arithmetic means cannot be recovered from it.

MUST-KNOW NOTES (MD-LEVEL)

- **Log:** for positive, right-skewed data (lab values, LOS, antibody titres, costs). Zeros need a small added constant, which changes results; state the constant used.

- **Square root:** a gentler compression, often used for counts, whose variance grows with the mean (recall the Poisson distribution, §5.7.1.3).
- **Box–Cox:** a family of power transformations that includes the log and square root; it estimates the power that makes the data closest to Normal.
- Report results back on the original scale. For logs this is a ratio of geometric means, as above. The same back-transformation applies to a regression on a log-transformed outcome: exp(slope) is then a ratio of geometric means.

CHECKPOINT 9.3

1. A dataset fails the Shapiro–Wilk test with n = 2,000. Should you conclude it is unfit for a parametric test? What should you look at instead?
2. A right-skewed variable, household income, is log-transformed before a t-test. The mean log income is 0.18 higher in village X than in village Y, with 95% CI 0.02 to 0.34. What does the difference represent on the original scale? Give the point estimate and CI.

(Answers in the Appendix at the end of the book.)

9.5 ANOVA

DEFINITION

ANOVA (analysis of variance) tests whether three or more group means are equal. It splits the total variability of the outcome into a between-group part and a within-group part and compares them with an F-ratio (recall §5.7.2.5).

SIMPLIFIED EXPLANATION

With four clinics there are six pairs. Testing each pair at $\alpha = 0.05$ inflates the chance of at least one false positive. If the six tests were independent and every H_0 true, that chance would be $1 - 0.95^6 =$ **0.265** (recall the complement rule, §5.2.3). ANOVA asks one question instead: is the spread of the group means larger than the spread within groups would produce by chance? One test, one α.

MUST-KNOW NOTES (MD-LEVEL)

- F = (between-group mean square) ÷ (within-group mean square). Under H_0 both estimate the same variance, so F is near 1.
- H_0: all population means are equal. A significant F says at least one differs; it does not say which. §9.5.3 covers comparing specific groups.
- Assumptions are those of the t-test (§9.2.4): independence, approximate Normality within groups, equal variances. Welch's ANOVA relaxes the equal-variance assumption, as Welch's t-test does.
- Forms covered: one-way (§9.5.2), two-way (§9.5.4), repeated measures (§9.5.5).

9.5.1 Logic of ANOVA – Partitioning Variance

DEFINITION

For k groups with n_i observations each (N in total), group means $\bar{y}_i$ and grand mean $\bar{y}$:

- $SST = \Sigma(y - \bar{y})^2$ (total), df $N - 1$
- $SSB = \Sigma\, n_i(\bar{y}_i - \bar{y})^2$ (between groups), df $k - 1$
- $SSW = \Sigma(y - \bar{y}_i)^2$ (within groups), df $N - k$
- $SST = SSB + SSW$; $MSB = SSB/(k - 1)$; $MSW = SSW/(N - k)$; $F = MSB/MSW$ with $(k - 1, N - k)$ df.

SIMPLIFIED EXPLANATION

Each observation's distance from the grand mean has two parts: its group mean's distance from the grand mean (signal) plus its own distance from its group mean (noise). Squaring and summing keeps the two parts separate. Dividing each sum by its df turns it into a variance estimate, a "mean square". MSW estimates the common within-group variance whatever the truth. MSB estimates the same variance only if H_0 is true, and is inflated if the means differ.

DERIVATION

1. $y - \bar{y} = (\bar{y}_i - \bar{y}) + (y - \bar{y}_i)$.
2. Square and sum over everyone: $\Sigma(y - \bar{y})^2 = \Sigma n_i(\bar{y}_i - \bar{y})^2 + \Sigma(y - \bar{y}_i)^2 + 2\Sigma(\bar{y}_i - \bar{y})(y - \bar{y}_i)$.
3. Within group i, $(\bar{y}_i - \bar{y})$ is constant and $\Sigma(y - \bar{y}_i) = 0$ (deviations from a mean sum to zero, recall §2.2.1), so the cross term is 0. Hence SST = SSB + SSW exactly.
4. The df add the same way: $(N - 1) = (k - 1) + (N - k)$.

ILLUSTRATION / EXAMPLE

Birthweights (kg) of five newborns at each of three PHCs:

PHC A	PHC B	PHC C
2.6, 2.8, 2.9, 3.0, 2.7	3.0, 3.1, 2.9, 3.3, 3.2	2.9, 3.0, 2.8, 3.1, 2.7
mean 2.80	mean 3.10	mean 2.90

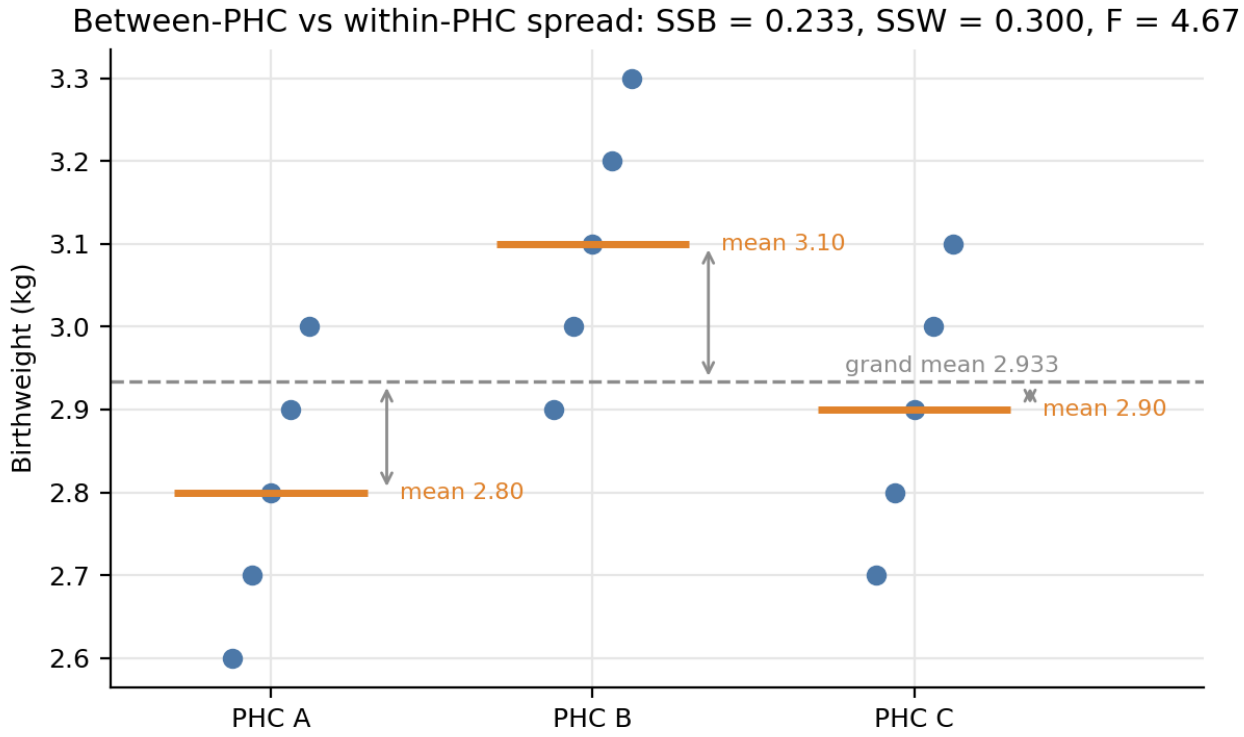


Figure 9.2. Each dot is one newborn. Orange bars are the PHC means; the dashed line is the grand mean, 2.933 kg. Arrows show the between-group deviations that make up SSB; the scatter of dots around each orange bar makes up SSW.

WORKED CALCULATION

1. Grand mean $\bar{y} = 44.0 \div 15 = 2.933$ kg.
2. SSB $= 5 \times [(2.80 - 2.933)^2 + (3.10 - 2.933)^2 + (2.90 - 2.933)^2] = 5 \times (0.01778 + 0.02778 + 0.00111) =$ **0.2333**.
3. SSW: within PHC A, deviations −0.2, 0, 0.1, 0.2, −0.1 give 0.10; PHC B and PHC C also give 0.10 each. SSW = **0.300**.
4. SST = 0.2333 + 0.300 = 0.5333 (check: $\Sigma(y - 2.933)^2 = 0.5333$).
5. ANOVA table:

Source	SS	df	MS	F
Between PHCs	0.2333	2	0.1167	4.67
Within PHCs	0.3000	12	0.0250	
Total	0.5333	14		

1. The 95% point of F(2, 12) is **3.89**. F = 4.67 > 3.89, so reject H_0 at the 5% level; p = 0.032.

MUST-KNOW NOTES (MD-LEVEL)

- SST = SSB + SSW is an identity, not an approximation; this split is what "analysis of variance" means.
- MSB uses k − 1 df, MSW uses N − k. Always report both df with F: here F(2, 12) = 4.67.

- The same between-versus-within comparison measures clustering in the intraclass correlation coefficient (§15.2.1: the share of total variance lying between clusters).

9.5.2 One-Way ANOVA

DEFINITION

One-way ANOVA applies §9.5.1 when the groups are defined by one categorical factor.

SIMPLIFIED EXPLANATION

"One-way" means one factor, such as PHC or wealth quintile. With exactly two groups, one-way ANOVA is the pooled t-test: $F = t^2$ (recall $t^2_\nu = F(1, \nu)$, §5.7.2.5), and the p-values are identical.

IN R (OPTIONAL)

```
cl ← read.csv("data/clinic_children.csv")
cl$quintile ← factor(cl$wealth_quintile)
tapply(cl$weight_kg, cl$quintile, mean)
fit ← aov(weight_kg ~ quintile, data = cl)
summary(fit)
```

```
       1        2        3        4        5
11.34615 11.99615 11.99048 12.39000 12.81000
             Df Sum Sq Mean Sq F value Pr(>F)
quintile      4  19.41   4.853   2.457 0.0508 .
Residuals    95 187.68   1.976
```

Mean weight rises from 11.35 kg in the poorest quintile to 12.81 kg in the richest. The table has the layout of §9.5.1: SSB = 19.41 on 5 − 1 = 4 df, SSW = 187.68 on 100 − 5 = 95 df, F = 4.853 ÷ 1.976 = 2.46, p = 0.051. `factor()` matters: without it, R would treat quintile as a number and fit a straight-line trend instead of five separate means.

MUST-KNOW NOTES (MD-LEVEL)

- F(k − 1, N − k); report the group means and SDs, not only F and p.
- A significant F says the means are not all equal. A non-significant F does not show they are equal (recall §8.5). Here p = 0.051 is weak evidence, not evidence of no difference.
- Wealth and weight here are observational: the gradient is an association, not an effect of wealth.

9.5.3 Post-Hoc Tests and Multiple Comparisons

DEFINITION

Multiple comparison procedures test several hypotheses together while controlling an error rate for the whole set. The **family-wise error rate (FWER)** is the probability of at least one false positive among the k tests. The **false discovery rate (FDR)** is the expected proportion of false positives

among the hypotheses rejected. **Post-hoc tests** are multiple comparison procedures for pairs of group means after ANOVA.

SIMPLIFIED EXPLANATION

Every test at $\alpha = 0.05$ carries a 5% false-positive risk when its H_0 is true, and the risks accumulate. Procedures fall into two families. Bonferroni-type methods adjust p-values or α and work for any set of tests. Methods built for group means, such as Tukey's and Dunnett's, use the joint distribution of the comparisons and are more powerful for that job.

DERIVATION

1. *Independent tests, all H_0 true* (recall the multiplicity note in §8.5): P(no false positive) = $(1 - \alpha)^k$ by the multiplication rule (recall §5.2.2), so FWER = $1 - (1 - \alpha)^k$.
2. *Any tests:* P(at least one of k events) ≤ the sum of their probabilities (the addition rule without the overlap term, recall §5.2.1). Testing each at α/k gives FWER ≤ $k \times \alpha/k = \alpha$. That is **Bonferroni**, and it needs no independence.
3. *Holm:* sort the p-values $p_{(1)} \le \ldots \le p_{(k)}$. Compare $p_{(1)}$ with α/k, $p_{(2)}$ with $\alpha/(k-1)$, and so on; stop at the first non-rejection. It controls the FWER and rejects everything Bonferroni rejects, and sometimes more (Holm, 1979).

WORKED CALCULATION 1: HOW FAST THE FWER GROWS (A = 0.05, INDEPENDENT TESTS)

Tests, k	FWER = $1 - 0.95^k$	Bonferroni α/k
3	0.143	0.0167
5	0.226	0.010
6	0.265	0.0083
10	0.401	0.005

WORKED CALCULATION 2: BONFERRONI, HOLM AND BH ON FOUR P-VALUES

A trial reports four secondary outcomes with p = 0.004, 0.015, 0.020 and 0.20.

1. *Bonferroni:* multiply by 4: 0.016, 0.060, 0.080, 0.80. One outcome significant.
2. *Holm:* compare 0.004 with 0.05/4 = 0.0125 (reject); 0.015 with 0.05/3 = 0.0167 (reject); 0.020 with 0.05/2 = 0.025 (reject); 0.20 with 0.05 (stop). Three significant. Adjusted p-values: 0.016, 0.045, 0.045, 0.20.
3. *Benjamini–Hochberg (BH):* compare $p_{(i)}$ with $(i/k) \times 0.05$ = 0.0125, 0.025, 0.0375, 0.05. The largest i satisfying it is i = 3, so reject the first three. Adjusted p-values: 0.016, 0.027, 0.027, 0.20.

WORKED CALCULATION 3: TUKEY'S HSD ON THE THREE PHCS (§9.5.1)

Tukey's **honestly significant difference** uses the **studentized range distribution**: the distribution of (largest − smallest of k sample means) ÷ (their estimated SE) when all true means are equal.

1. From that distribution, q(0.95; k = 3, df = 12) = 3.773.
2. HSD = q × √(MSW/n) = 3.773 × √(0.025/5) = 3.773 × 0.0707 = **0.267 kg**.
3. Pairwise differences: B − A = 0.30 (> 0.267, significant); B − C = 0.20; C − A = 0.10 (both < 0.267). Only PHC B differs from PHC A.
4. For contrast, Fisher's **least significant difference** (LSD) is an ordinary pooled t-test for each pair: $t_{12} \times \sqrt{(2 \times MSW/n)} = 2.179 \times 0.1 = 0.218$ kg. It makes no correction for the three comparisons.

IN R (OPTIONAL)

```
cl ← read.csv("data/clinic_children.csv")
cl$quintile ← factor(cl$wealth_quintile)
fit ← aov(weight_kg ~ quintile, data = cl)
TukeyHSD(fit)
p.adjust(c(0.004, 0.015, 0.020, 0.20), method = "holm")
p.adjust(c(0.004, 0.015, 0.020, 0.20), method = "BH")
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
$quintile
            diff        lwr      upr     p adj
2-1  0.650000000 -0.6776891 1.977689 0.6536177
...
4-1  1.043846154 -0.3486459 2.436338 0.2352979
5-1  1.463846154  0.0713541 2.856338 0.0343073
...
5-4  0.420000000 -0.8160113 1.656011 0.8785081

[1] 0.016 0.045 0.045 0.200
[1] 0.01600000 0.02666667 0.02666667 0.20000000
```

All ten pairwise differences of §9.5.2 get intervals with a joint 95% coverage. Only the richest vs poorest quintile (5 − 1) excludes 0: 1.46 kg (0.07 to 2.86), adjusted p = 0.034. This happened although the overall F in §9.5.2 had p = 0.051. There is no contradiction: F spreads its power over every way the five means could differ, while Tukey concentrates on pairs. The `p.adjust` lines reproduce the Holm and BH columns of Worked Calculation 2.

MUST-KNOW NOTES (MD-LEVEL)

- **A significant F is not a precondition for Tukey, Dunnett, Bonferroni or Holm.** These control the FWER on their own. GraphPad's statistics guide states that "all the multiple comparisons tests offered by Prism are valid even if the overall ANOVA did not find a significant difference among means", and quotes Maxwell and Delaney: requiring a significant omnibus test first "only serves to lower alpha ... below the desired level ... and hence inappropriately decreases power". The F-test gate belongs to **Fisher's LSD**, which makes no correction of its own: "protected" LSD

is run only after a significant F. The same guide calls protection "no longer recommended" because it controls the FWER poorly.

- **Tukey vs Dunnett.** Tukey compares every pair. Dunnett compares each treatment with one control only, "comparing several treatments simultaneously with a control or standard treatment" while keeping a joint significance level (Dunnett, 1964). It makes fewer comparisons, so it is more powerful for that question: four treatment arms against placebo give 4 comparisons instead of Tukey's 10.
- **Bonferroni vs Holm.** Holm is uniformly at least as powerful and equally valid, so there is little reason to report plain Bonferroni except for its simplicity.
- **FDR.** In screening many hypotheses (for example, dozens of biomarkers), controlling the FWER is very strict. Benjamini and Hochberg (1995) proposed "controlling the expected proportion of falsely rejected hypotheses — the false discovery rate". It gives more power when many hypotheses are false, and it suits exploratory screening rather than a trial's confirmatory outcome.
- **Planned vs unplanned.** A few comparisons specified in the protocol before seeing the data can be tested directly, with a correction for that small number only. Comparisons chosen after looking at the data need a full correction.
- **Subgroup analyses** are multiple comparisons too. Testing an effect in 10 subgroups gives a FWER of 0.40 when there is no effect in any (Worked Calculation 1). An effect "significant in women but not in men" is not evidence that the effect differs by sex; that needs a test of interaction (§9.5.4).

CHECKPOINT 9.4

1. A one-way ANOVA comparing four clinics is significant. A colleague concludes "all four clinics differ from each other". What is wrong, and what should be done next?
2. A study tests 8 independent hypotheses at $\alpha = 0.05$, all nulls true. What is the FWER? What is the Bonferroni per-test α?
3. A one-way ANOVA has 4 groups and $N = 24$, with $SSB = 6.0$ and $SSW = 20.0$. Complete the ANOVA table and decide at $\alpha = 0.05$ (the 95% point of $F(3, 20)$ is 3.10).

(Answers in the Appendix at the end of the book.)

9.5.4 Two-Way ANOVA

DEFINITION

Two-way ANOVA tests the effects of two categorical factors on a quantitative outcome at the same time: the main effect of each factor and their **interaction**, meaning that the effect of one factor depends on the level of the other.

SIMPLIFIED EXPLANATION

The one-way split of §9.5.1 becomes SST = SS(factor 1) + SS(factor 2) + SS(interaction) + SS(within), each with its own F. The interaction asks something neither one-way analysis can: is the gap between supplemented and unsupplemented mothers the same at every clinic?

ILLUSTRATION / EXAMPLE

Mean birthweight (kg) by clinic and maternal supplementation (hypothetical):

Clinic	No supplementation	Supplementation	Gain
A	2.80	2.85	0.05
B	2.85	3.30	0.45
C	2.90	3.00	0.10
D	2.75	2.95	0.20

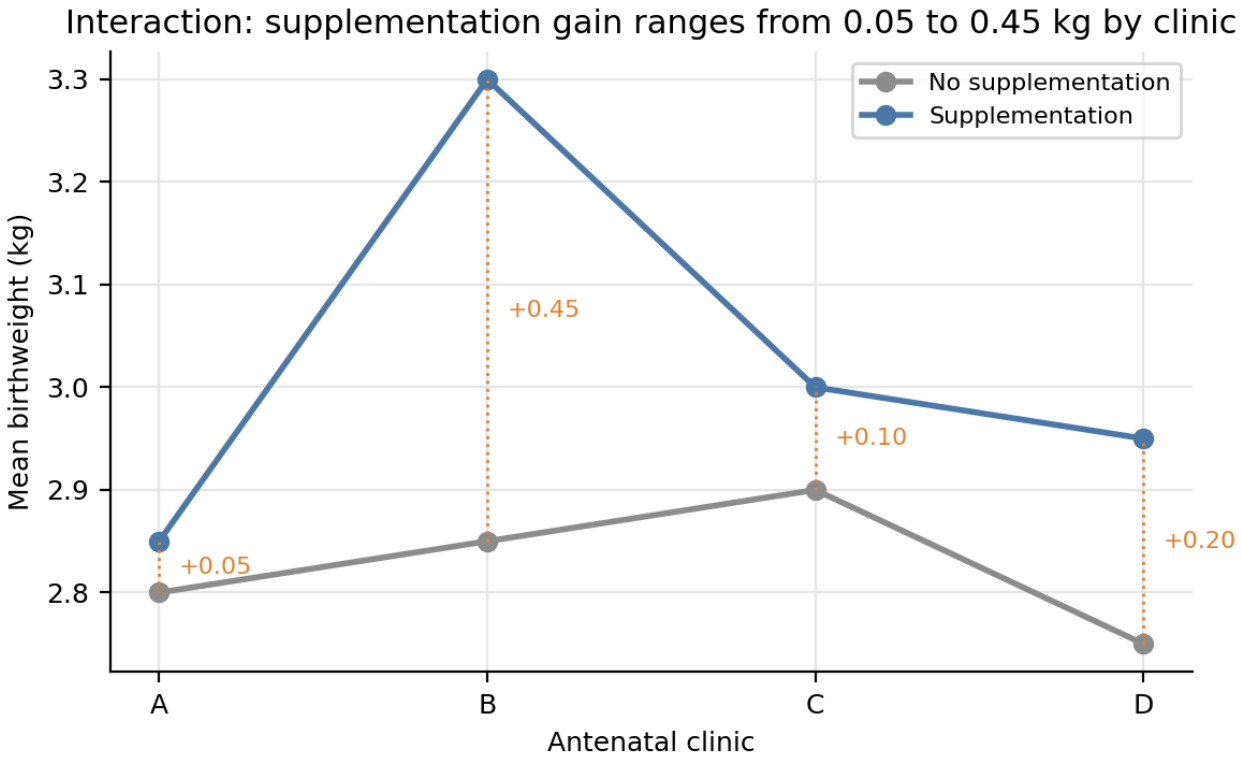


Figure 9.3. Mean birthweight at four clinics with and without supplementation. Parallel lines would mean no interaction; here the gain ranges from 0.05 kg (clinic A) to 0.45 kg (clinic B).

MUST-KNOW NOTES (MD-LEVEL)

- Three F-tests: factor 1, factor 2, interaction (df = (a − 1)(b − 1) for factors with a and b levels).
- If the interaction is real, a single "effect of supplementation" (here an average gain of 0.20 kg) misdescribes every clinic; report the effect by clinic.
- With unequal cell sizes the sums of squares depend on the order of entry; software offers alternatives. Two-way ANOVA is a special case of multiple regression with categorical predictors (Chapter 12), which handles unbalanced designs more transparently.

9.5.5 Repeated Measures ANOVA

DEFINITION

Repeated measures ANOVA (RM-ANOVA) compares the means of the same subjects measured at three or more time points or conditions. It removes between-subject variation from the error term, as the paired t-test does for two measurements (§9.2.3).

SIMPLIFIED EXPLANATION

Measurements on one person are correlated, so treating time points as independent groups would be wrong. RM-ANOVA treats each subject as a **block**: a set of related measurements whose shared level is removed before the time points are compared. It adds an assumption, **sphericity**: the variances of the differences between every pair of time points are equal. When sphericity fails, the Greenhouse–Geisser correction reduces the df to keep the Type I error near α.

ILLUSTRATION / EXAMPLE

Systolic BP in 6 patients at baseline, 1 month and 3 months after starting treatment:

Patient	0 m	1 m	3 m	Change per month (3 m − 0 m) ÷ 3
1	152	146	140	−4.0
2	160	158	151	−3.0
3	148	147	145	−1.0
4	156	150	147	−3.0
5	164	160	152	−4.0
6	150	149	147	−1.0

WORKED CALCULATION: THE SUMMARY-MEASURE APPROACH

RM-ANOVA asks "do the three time-point means differ?". The clinical question is usually "how fast does BP fall?". Matthews, Altman, Campbell and Royston (1990) recommend a two-stage method: reduce each subject's series to one clinically meaningful number, "such as a rate of change or an area under a curve", then analyse those numbers "by simple statistical techniques as though they were raw data".

1. Stage 1: each patient's change per month (last column).
2. Stage 2: mean = −2.67 mmHg/month, SD = 1.37, SE = 1.37/√6 = 0.558.
3. One-sample t-test against 0 (§9.2.1): t = −2.67 ÷ 0.558 = **−4.78**, 5 df, p = 0.005. 95% CI: −2.67 ± 2.571 × 0.558 = **−4.10 to −1.23 mmHg/month**.

Using all three visits to fit a straight line for each patient gives a better slope; that uses regression (Chapter 12).

MUST-KNOW NOTES (MD-LEVEL)

- RM-ANOVA generalises the paired t-test, as one-way ANOVA generalises the independent t-test.

- **Limits.** RM-ANOVA needs every subject measured at the same fixed times. A subject with one missed visit is dropped entirely, which wastes data and can bias the result if drop-outs differ from completers. It cannot use unequally spaced or subject-specific visit times, and its F-test answers a vague question ("do the means differ anywhere?").
- **Alternatives.** The summary-measure approach above is simple and valid (Matthews et al., 1990). Linear mixed models (§15.2.4: regression models with a random intercept or slope for each person) use every available measurement and allow unequal timing, so they suit longitudinal data with missing visits.

CHECKPOINT 9.5

1. Using the table in §9.5.4, compute the supplementation gain at clinic A and at clinic B, and the difference between the two gains. What does a non-zero difference of this kind represent?
2. Haemoglobin is measured in 40 pregnant women at 0, 2 and 4 months; 9 miss the 4-month visit. What does RM-ANOVA do with those 9 women, and name two better approaches.
3. Four children's weights gain 0.20, 0.30, 0.10 and 0.20 kg per month. Using the summary-measure approach, compute the mean gain and its SE.

(Answers in the Appendix at the end of the book.)

Comparing Groups – Non-Parametric Tests

Chapter 9's tests assume a quantitative outcome with a roughly Normal distribution. This chapter covers what to use when that assumption fails or the outcome is not quantitative at all. §10.1–10.5 give the rank- and sign-based tests, one for each design in Chapter 9. §10.6–10.8 turn to counts in tables: the chi-square tests, the measures of risk and odds read from a 2 × 2 table, the exact test for small tables and the tests for paired binary data.

10.1 When to Use Non-Parametric Tests

DEFINITION

A **non-parametric** (distribution-free) test does not assume that the data come from a particular distribution such as the Normal. Most such tests replace each value by its rank or its sign. Use one when the outcome is ordinal, when a small sample is clearly non-Normal, or when a few extreme outliers would dominate a mean-based test.

SIMPLIFIED EXPLANATION

The t-test and ANOVA rely on roughly Normal data (recall §9.2.4). Non-parametric tests drop that requirement by working with the order of the values rather than their size. The price is some loss of power (recall §8.4) when the data really are Normal.

ILLUSTRATION / EXAMPLE

Hospital length of stay is right-skewed (recall §3.9). In a small study comparing two wards, a few patients who stay for weeks would pull each ward's mean far to the right. The Mann-Whitney U test (§10.2) sees only who ranks above whom, so a 40-day stay counts as "the longest", no more.

MUST-KNOW NOTES (MD-LEVEL)

- Use them for ordinal outcomes (means are not meaningful, recall §2.5 and §1.1.2.2), small samples with clear non-Normality, or outliers you cannot justify removing.
- Ranking makes a test resistant to outliers, because an extreme value can only take the top (or bottom) rank. It does not make the test blind to shape: for example, the Mann-Whitney test reads as a comparison of medians only when the two groups' distributions have a similar shape (§10.2).

- When the parametric assumptions hold, the non-parametric test is somewhat less powerful. When they fail badly, it can be the more powerful of the two.
- The rank tests mirror Chapter 9: sign test and Wilcoxon signed-rank ↔ paired t-test; Mann-Whitney ↔ independent-samples t-test; Kruskal-Wallis ↔ one-way ANOVA; Friedman ↔ repeated-measures ANOVA. For categorical outcomes the chi-square family (§10.6), Fisher's exact test (§10.7) and McNemar's test (§10.8) take over.

10.1.1 The Sign Test

DEFINITION

The **sign test** is a test for paired data. It records only whether each pair's difference is positive or negative and asks whether the two signs are equally likely. Under the null hypothesis (H_0, recall §8.2) that the median difference is zero, the number of positive differences among n non-zero differences follows a Binomial(n, 0.5) distribution (recall §5.7.1.2). A test of a proportion against a fixed value using this exact distribution is the **exact binomial test** (recall §9.1.1); the sign test applies it to the signs of paired differences.

SIMPLIFIED EXPLANATION

If an intervention does nothing, each person is as likely to go up as to go down, like a coin toss. Count the rises. If there are far more (or far fewer) than half, the coin looks biased. Pairs with a difference of exactly zero carry no sign and are dropped, so n is the number of non-zero differences. The test uses only the direction of each change, so it is valid for purely ordinal scores, where "improved by two points" and "improved by one point" cannot be compared in size.

DERIVATION

1. Under H_0, each non-zero difference is positive with probability 0.5, independently of the others.
2. So X = number of positives ~ Binomial(n, 0.5), and $P(X = k) = C(n, k) \times 0.5^n$ (recall §5.7.1.2).
3. The p-value is the probability, under H_0, of a split at least as lopsided as the one observed. Binomial(n, 0.5) is symmetric, so the two-sided p-value is twice the smaller tail (recall §8.6).

ILLUSTRATION / EXAMPLE

Twelve anaemic women at a PHC had their haemoglobin (Hb) measured before and after a course of iron–folic acid tablets. The changes (after − before, g/dL) were: +0.8, +0.4, +1.1, −0.3, +0.7, +0.9, +0.6, −0.2, +1.0, +1.2, +0.5, +1.3. Ten rose and two fell. The mean change is 0.67 g/dL and the median change 0.75 g/dL.

WORKED CALCULATION

1. n = 12 non-zero differences; X = 10 positives, 2 negatives.
2. The smaller count is 2. $P(X \le 2) = [C(12, 0) + C(12, 1) + C(12, 2)] \div 2^{12} = (1 + 12 + 66) \div 4{,}096 = 79 \div 4{,}096 =$ **0.0193**.
3. Two-sided $p = 2 \times 0.0193 =$ **0.039**.

4. At the 5% level you reject H_0: the rises outnumber the falls by more than chance alone would usually give.

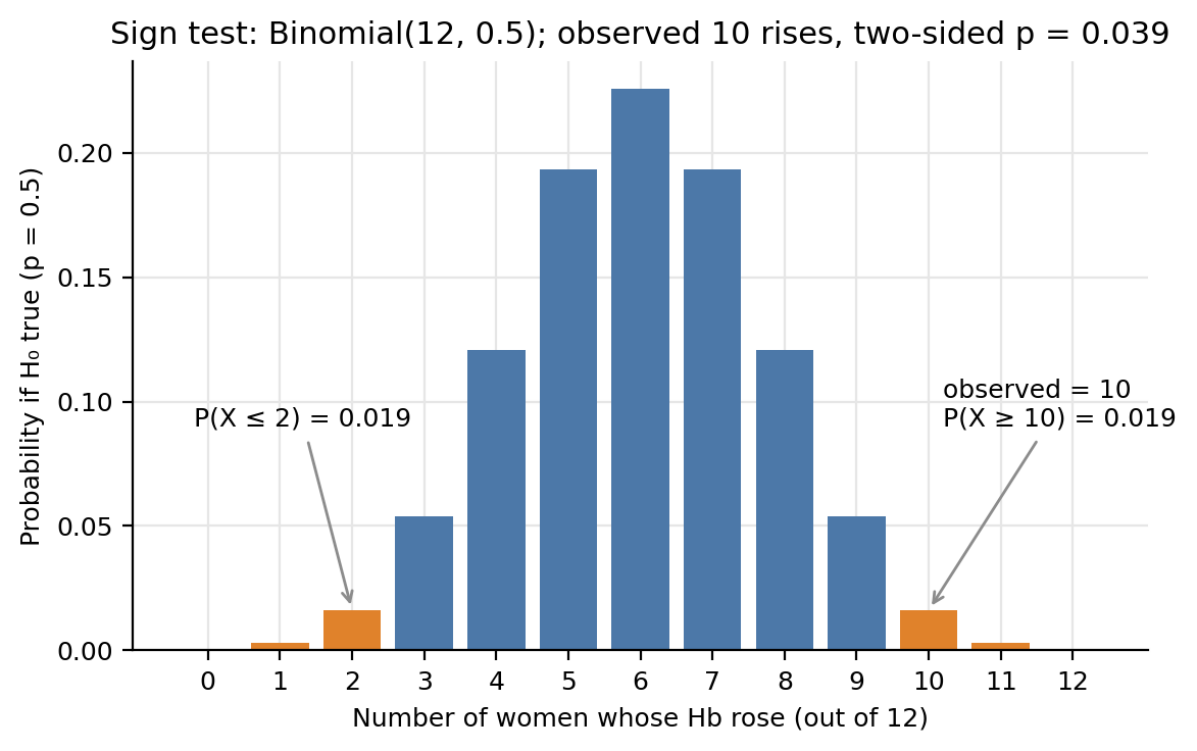


Figure 10.1. The sign test for the 12 women. If iron had no effect, the number of rises would follow Binomial(12, 0.5). The shaded tails (≤ 2 and ≥ 10) each hold probability 0.019, giving a two-sided p of 0.039.

IN R (OPTIONAL)

```
hb ← read.csv("data/sign_test_hb.csv")
d  ← hb$hb_after - hb$hb_before       # change in Hb, g/dL
table(sign(d))                        # -1 = fell, 1 = rose
binom.test(sum(d > 0), sum(d ≠ 0))    # exact sign test
```

```
-1  1
 2 10

    Exact binomial test

data:  sum(d > 0) and sum(d ≠ 0)
number of successes = 10, number of trials = 12, p-value = 0.03857
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5158623 0.9791375
sample estimates:
probability of success
             0.8333333
```

`table` gives the 2 falls and 10 rises. `binom.test` gives the two-sided p of step 3 (0.0386). It also estimates the proportion of women whose Hb rises, 10/12 = 0.83, with an exact 95% CI of 0.52 to 0.98 (recall §7.2.4).

MUST-KNOW NOTES (MD-LEVEL)

- The sign test assumes only that the pairs are independent of each other. It makes no assumption about the shape of the differences and needs only their direction.
- It throws away the size of each change, so it has low power. On these same 12 women the Wilcoxon signed-rank test, which also uses the sizes, gives p = 0.002 (§10.3). Use the sign test when sizes are not meaningful (ordinal scores) or cannot be trusted.
- The same exact binomial calculation tests any single proportion against a fixed value when n is small (recall §5.7.2.2.4 for when the Normal approximation is adequate). It is also the exact form of McNemar's test (§10.8, a test for paired yes/no outcomes).

CHECKPOINT 10.1

1. In a sign test on 10 pairs, 8 differences are positive, 1 is negative and 1 is zero. What are n and X, and what is the two-sided p-value?
2. Why is the sign test valid for a 5-point pain score when a test using the sizes of the differences may not be?

(Answers in the Appendix at the end of the book.)

10.2 Mann-Whitney U Test

DEFINITION

The **Mann-Whitney U test** (also called the Wilcoxon rank-sum test) compares two independent groups. It ranks all observations together and tests whether values in one group tend to be larger than values in the other. Its null hypothesis is that a randomly chosen value from group 1 is equally likely to be larger or smaller than a randomly chosen value from group 2.

SIMPLIFIED EXPLANATION

Pool both groups and rank every value from smallest (rank 1) to largest, giving tied values the average of their ranks. If the groups do not differ, each group's ranks should be spread evenly through the pool. If one group's values tend to be larger, its ranks will pile up at the top. U counts, over all possible pairs of one person from each group, how many times the group-1 person has the larger value.

DERIVATION

1. Suppose every member of group 1 (size n_1) has a smaller value than everyone in group 2. Their ranks are 1, 2, … , n_1, with sum $n_1(n_1 + 1)/2$, and they exceed nobody: $U_1 = 0$.
2. Each time a group-1 value moves above one group-2 value, that person's rank rises by 1 and U_1 rises by 1.
3. So $U_1 = R_1 - n_1(n_1 + 1)/2$, where R_1 is group 1's rank sum. With n_1n_2 pairs in all, $U_1 + U_2 = n_1n_2$.

4. Under H_0, U_1 has mean $n_1n_2/2$ and SD $\sqrt{[n_1n_2(N+1)/12]}$, where $N = n_1 + n_2$; for larger groups, $z = (U_1 - n_1n_2/2) \div \text{SD}$ is referred to the Normal curve.

ILLUSTRATION / EXAMPLE

In the CBHI household survey (500 households, 150 members), is the distance from the health facility different for member and non-member households? Median distance is 6.9 km for members and 8.5 km for non-members.

WORKED CALCULATION

1. Rank all 500 distances. The members' rank sum is $R_1 = 29{,}758.5$ (mean rank 198.4; non-members' mean rank 272.8).
2. $U_1 = 29{,}758.5 - 150 \times 151/2 = 29{,}758.5 - 11{,}325 = 18{,}433.5$. $U_0 = 150 \times 350 - 18{,}433.5 = 34{,}066.5$.
3. $U_1 \div (n_1n_0) = 18{,}433.5 \div 52{,}500 = 0.35$. Pick one member and one non-member at random: the member lives farther away only about 35% of the time (ties counted as half). Under H_0 this would be 50%.
4. Under H_0, mean $U = 26{,}250$ and SD $= \sqrt{(52{,}500 \times 501/12)} = 1{,}480.5$. $z = (18{,}433.5 - 26{,}250) \div 1{,}480.5 = -5.28$, $p \approx 1 \times 10^{-7}$.

IN R (OPTIONAL)

```
cb ← read.csv("data/cbhi_survey.csv")
tapply(cb$distance_km, cb$cbhi_member, median)  # 0 = no, 1 = member
wilcox.test(distance_km ~ cbhi_member, data = cb)
```

```
  0   1
8.5 6.9

    Wilcoxon rank sum test with continuity correction

data:  distance_km by cbhi_member
W = 34066, p-value = 1.293e-07
alternative hypothesis: true location shift is not equal to 0
```

R's W is U for the first group listed (non-members), 34,066.5 from step 2, printed rounded. The p-value (1.3×10^{-7}) matches step 4 after R's small corrections for ties and continuity.

MUST-KNOW NOTES (MD-LEVEL)

- Assumptions: observations independent within and between groups; an outcome that can at least be ordered.
- It is a test of medians only when the two distributions have the same shape and differ just by a shift. When shapes or spreads differ, it can be significant even with similar medians (Hart, 2001). Here the member distances are more spread out (IQR 3.5–9.2 km) than the non-member distances (6.6–10.2 km), so report the result as "members tend to live closer", with the medians as a description, not as "the medians differ".

10.3 Wilcoxon Signed-Rank Test

DEFINITION

The **Wilcoxon signed-rank test** is a paired-data test. It ranks the absolute sizes of the paired differences, attaches each difference's sign to its rank, and tests whether the positive and negative ranks balance. Its H_0 is that the differences are symmetric about zero.

SIMPLIFIED EXPLANATION

The sign test (§10.1.1) counts rises and falls. The signed-rank test also asks how big they were: a few small falls and many large rises is stronger evidence than the same count of rises and falls of similar size. That extra information buys power, but it has a cost. Ranking differences by size assumes that a difference of 1.2 is really bigger than a difference of 0.8, which is true for Hb in g/dL (an interval scale, recall §1.1.2.3) but not for a pain score, where a change from 2 to 4 need not equal a change from 6 to 8.

ILLUSTRATION / EXAMPLE

For the 12 women's Hb changes (§10.1.1), rank the absolute changes: 0.2 gets rank 1, 0.3 rank 2, 0.4 rank 3, and so on up to 1.3 with rank 12. The two falls (−0.2, −0.3) hold the two smallest ranks.

For a 0–10 pain score measured before and after physiotherapy, the scores are ordinal. The sign test (§10.1.1) is the safe choice: it asks only whether more patients improved than worsened.

WORKED CALCULATION

1. Ranks of the negative differences: 1 and 2, so $W^- = 3$. The ranks 1 to 12 sum to $12 \times 13/2 = 78$, so $W^+ = 75$.
2. Under H_0 each rank is equally likely to carry a + or a − sign: $2^{12} = 4{,}096$ equally likely patterns.
3. Patterns with $W^- \le 3$: no negative rank; {1}; {2}; {3}; {1, 2}. That is 5 patterns, so $P(W^- \le 3) = 5 \div 4{,}096 = 0.0012$.
4. Two-sided $p = 2 \times 0.0012 =$ **0.002**, much smaller than the sign test's 0.039 on the same data, because the two falls were also the two smallest changes.

IN R (OPTIONAL)

```
hb ← read.csv("data/sign_test_hb.csv")
wilcox.test(hb$hb_after, hb$hb_before, paired = TRUE)
```

```
    Wilcoxon signed rank exact test

data:  hb$hb_after and hb$hb_before
V = 75, p-value = 0.002441
alternative hypothesis: true location shift is not equal to 0
```

V is $W^+ = 75$ from step 1, and the exact p-value 0.0024 is $10 \div 4{,}096$ from steps 3 and 4.

MUST-KNOW NOTES (MD-LEVEL)

- Two requirements: the differences must be meaningful in size (interval data), and their distribution should be roughly symmetric. Under symmetry the test is a test of the median difference. For purely ordinal scores, use the sign test.
- It is the usual non-parametric alternative to the paired t-test (recall §9.2.3) when the differences are clearly non-Normal in a small sample.
- Zero differences are dropped, as in the sign test; many ties or zeros make software switch from exact to approximate p-values.

10.4 Kruskal-Wallis Test

DEFINITION

The **Kruskal-Wallis test** compares three or more independent groups. It ranks all observations together and tests whether the groups' mean ranks differ more than chance would allow. It is the rank-based counterpart of one-way ANOVA (recall §9.5.2).

SIMPLIFIED EXPLANATION

It is the Mann-Whitney idea applied to k groups. Under H_0 every group's mean rank should be close to the overall mean rank, $(N + 1)/2$. The statistic H adds up the squared distances of the group mean ranks from that centre, weighted by group size, and scales them so that H is approximately χ^2 with $k - 1$ degrees of freedom (recall §5.7.2.4).

$$H = [12 \div (N(N + 1))] \times \Sigma\, n_i \bar{R}^2 - 3(N + 1)$$

where n_i and $\bar{R}$ are the size and mean rank of group i.

ILLUSTRATION / EXAMPLE

In the CBHI survey, does distance to the facility differ by maternal education (none, primary, secondary or higher)? Median distances are 9.0, 8.5 and 7.1 km.

WORKED CALCULATION

1. Rank all 500 distances. Mean ranks: none 289.5 (n = 168), primary 263.3 (n = 177), secondary+ 193.6 (n = 155). The overall mean rank is 501/2 = 250.5.
2. $\Sigma\, n_i \bar{R}^2 = 168 \times 289.52^2 + 177 \times 263.32^2 + 155 \times 193.56^2 = 32{,}162{,}647$.
3. $H = 12 \div (500 \times 501) \times 32{,}162{,}647 - 3 \times 501 = 1{,}540.73 - 1{,}503 =$ **37.73**, on $k - 1 = 2$ df.
4. The 5% critical value of χ^2 with 2 df is 5.99. H far exceeds it ($p < 0.001$): distance differs across education groups.

Dunn's post-hoc test. Like ANOVA, a significant H says only that at least one group differs. To find which, **Dunn's test** (Dunn, 1964) compares each pair's mean ranks using the pooled ranking from the Kruskal-Wallis test:

$$z = (\bar{R}_i - \bar{R}_j) \div \sqrt{[S^2 \times (1/n_i + 1/n_j)]}, \text{ with } S^2 = N(N + 1)/12 \text{ (less a small correction for ties)}$$

and then adjusts the p-values for the number of comparisons, for example by Bonferroni (recall §9.5.3). For none vs secondary+: $S^2 = 500 \times 501/12 = 20{,}875$; $z = (289.5 - 193.6) \div \sqrt{[20{,}875 \times (1/168 + 1/155)]} = 95.9 \div 16.1 = 5.96$.

IN R (OPTIONAL)

```
cb ← read.csv("data/cbhi_survey.csv")
kruskal.test(distance_km ~ maternal_education, data = cb)
# Dunn's post-hoc test (base R): compare mean ranks pair by pair
r  ← rank(cb$distance_km); g ← factor(cb$maternal_education)
N  ← length(r); t ← table(cb$distance_km)
# rank variance, tie-corrected
S2 ← N * (N + 1) / 12 - sum(t^3 - t) / (12 * (N - 1))
Rm ← tapply(r, g, mean); n ← table(g)
pr ← combn(levels(g), 2)
z  ← apply(pr, 2, function(k)
  (Rm[k[1]] - Rm[k[2]]) / sqrt(S2 * (1/n[k[1]] + 1/n[k[2]])))
p  ← 2 * pnorm(-abs(z))
data.frame(pair = paste(pr[1, ], "vs", pr[2, ]), z = round(z, 2),
           p = signif(p, 3),
           p_bonferroni = signif(p.adjust(p, "bonferroni"), 3))
```

```
    Kruskal-Wallis rank sum test

data:  distance_km by maternal_education
Kruskal-Wallis chi-squared = 37.733, df = 2, p-value = 6.404e-09

                  pair    z        p p_bonferroni
1       none vs primary 1.68 9.23e-02     2.77e-01
2    none vs secondary+ 5.96 2.46e-09     7.37e-09
3 primary vs secondary+ 4.39 1.13e-05     3.40e-05
```

The first block is H = 37.73 on 2 df from step 3. Dunn's table shows where the difference lies: secondary+ households live closer than both other groups (Bonferroni $p < 0.001$), while none and primary do not differ clearly (adjusted $p = 0.28$). The packages `dunn.test` and `FSA` do the same in one line.

MUST-KNOW NOTES (MD-LEVEL)

- Same assumptions as Mann-Whitney, for k groups; the same shape caveat applies to reading it as a comparison of medians.
- With two groups, Kruskal-Wallis gives the same p-value as the Mann-Whitney test (without continuity correction), just as one-way ANOVA with two groups reproduces the t-test.
- Use Dunn's test (with a multiplicity adjustment) for pairwise follow-up, not repeated Mann-Whitney tests with separate rankings.

10.5 Friedman Test

DEFINITION

The **Friedman test** compares three or more related measurements, such as the same patients at several time points. It ranks each subject's own measurements separately and tests whether some condition tends to get the higher ranks. It is the rank-based counterpart of repeated-measures ANOVA (recall §9.5.5).

SIMPLIFIED EXPLANATION

Kruskal-Wallis ranks everybody together. Friedman ranks within each person, so differences between people (one patient always reports more pain than another) drop out, exactly as pairing does in the paired t-test. Only the order of each person's own scores is used, so an ordinal outcome is fine. With n subjects, k conditions and rank sums R_1 … R_k:

$Fr = [12 \div (nk(k+1))] \times \Sigma R_j^2 - 3n(k+1)$, approximately χ^2 with $k - 1$ df.

ILLUSTRATION / EXAMPLE

Six patients score pain (0–10) before physiotherapy, immediately after, and one month later. Patient 1 scores 7, 4, 3, so the ranks are 3, 2, 1. Across the six patients, "before" gets rank 3 every time.

WORKED CALCULATION

1. Rank sums (ties within a patient share ranks): before 18, after 10.5, one month 7.5.
2. $Fr = 12 \div (6 \times 3 \times 4) \times (18^2 + 10.5^2 + 7.5^2) - 3 \times 6 \times 4 = (1/6) \times 490.5 - 72 =$ **9.75**, on 2 df.
3. The critical value is 5.99, so $p < 0.05$. A correction for the one tied pair, which software applies automatically, raises Fr to 10.17 ($p = 0.006$). With only six patients the χ^2 approximation is rough, and this small example is for arithmetic only.

IN R (OPTIONAL)

```
pain ← matrix(c(7,4,3,  6,5,4,  8,5,5,  5,3,4,  7,6,3,  6,4,2),
              ncol = 3, byrow = TRUE,
              dimnames = list(paste0("P", 1:6),
                              c("before", "after", "month1")))
t(apply(pain, 1, rank))        # ranks within each patient
friedman.test(pain)
```

```
   before after month1
P1      3   2.0    1.0
P2      3   2.0    1.0
P3      3   1.5    1.5
P4      3   1.0    2.0
P5      3   2.0    1.0
P6      3   2.0    1.0

    Friedman rank sum test

data:  pain
Friedman chi-squared = 10.174, df = 2, p-value = 0.006177
```

The rank table is step 1 (column sums 18, 10.5, 7.5). The statistic is the tie-corrected version of step 2.

MUST-KNOW NOTES (MD-LEVEL)

- Ranking happens within each subject, never across subjects. That is the structural difference from Kruskal-Wallis.
- A significant result says only that the conditions are not all alike. Pairwise follow-up uses Wilcoxon or sign tests on the pairs, with a multiplicity adjustment (recall §9.5.3).

CHECKPOINT 10.2

1. Match each test to its parametric counterpart: Mann-Whitney U, Wilcoxon signed-rank, Kruskal-Wallis, Friedman.
2. Two groups of 5 are ranked together. Group 1 holds ranks 1, 2, 4, 5 and 7. Compute R_1, U_1 and U_2.
3. A Mann-Whitney test is significant, but the two groups have almost identical medians. How can that happen, and how should you report it?

(Answers in the Appendix at the end of the book.)

10.6 Chi-Square Test

DEFINITION

The **chi-square (χ^2) test** compares the observed counts (O) in the categories of a table with the counts expected (E) if a null hypothesis were true. Pearson's statistic is

$X^2 = \Sigma\,(O - E)^2 \div E$, summed over every cell.

Under H_0, and when expected counts are not too small (§10.6.1), X^2 approximately follows a χ^2 distribution (recall §5.7.2.4), whose degrees of freedom count the cells that are free to vary.

SIMPLIFIED EXPLANATION

Earlier tests in this chapter handled quantitative or ordinal outcomes. Here the data are counts of people in categories. For each cell, O − E is how far reality departs from the null hypothesis. A count behaves

roughly like a Poisson variable, whose variance equals its mean (recall §5.7.1.3), so $\sqrt{E}$ is roughly its SD and $(O - E) \div \sqrt{E}$ is roughly a standardised deviation, like a z-score. Squaring and adding these gives a sum of approximately squared standard Normals: exactly the recipe for a χ^2 variable. That is the whole logic. It is not a partition of variance, which is ANOVA's logic (recall §9.5.1).

ILLUSTRATION / EXAMPLE

In the CBHI survey, is membership (member or not) associated with maternal education? §10.6.1 works out the expected counts and X^2 for this 3×2 table.

MUST-KNOW NOTES (MD-LEVEL)

- Two uses: the **test of independence** between two categorical variables (§10.6.1) and the **goodness-of-fit** test of one variable against hypothesised proportions (§10.6.2).
- The χ^2 distribution is only an approximation for counts. Its validity depends on the expected counts, not the sample size as such (§10.6.1). When the expected counts are too small, use an exact test (§10.7).
- X^2 tests whether there is an association. It does not measure how strong it is; for a 2×2 table, §10.6.4 gives the measures.

10.6.1 Chi-Square Test of Independence

DEFINITION

The **chi-square test of independence** tests H_0: two categorical variables are independent (recall §5.4), against H_1: they are associated. In a table with r rows and c columns, the expected count in row i, column j is

E_{ij} = (row i total × column j total) ÷ N,

and X^2 is referred to the χ^2 distribution with $(r - 1)(c - 1)$ degrees of freedom.

SIMPLIFIED EXPLANATION

If membership had nothing to do with education, the 30% overall membership rate would hold in every education group. So the expected number of members among the 168 households with no maternal education is 30% of 168 = 50.4. That is the formula: (168 × 150) ÷ 500.

DERIVATION

Expected counts. Under independence, P(row i and column j) = P(row i) × P(column j) (recall §5.4). Estimate each probability by its margin: (row total ÷ N) × (column total ÷ N). Multiply by N to turn it into a count: E_{ij} = row total × column total ÷ N.

Degrees of freedom. The expected counts are built from the margins, so X^2 asks how the counts are arranged *given* the row and column totals. Count the cells you can fill freely once the totals are fixed:

1. In a 2×2 table with fixed totals, choose the top-left count a. Then b = row 1 total − a, c = column 1 total − a, and d = row 2 total − c. One free cell: df = 1.

2. In an r × c table, fill the first r − 1 rows and c − 1 columns freely. The last column in each row is set by the row totals, and the last row by the column totals. Free cells: (r − 1)(c − 1).

ILLUSTRATION / EXAMPLE

CBHI membership by maternal education (500 households):

Maternal education	Member	Non-member	Total
None	31 (E 50.4)	137 (E 117.6)	168
Primary	50 (E 53.1)	127 (E 123.9)	177
Secondary+	69 (E 46.5)	86 (E 108.5)	155
Total	150	350	500

WORKED CALCULATION

1. Expected counts, e.g. none/member: 168 × 150 ÷ 500 = 50.4; secondary+/member: 155 × 150 ÷ 500 = 46.5. The others are in the table.
2. Cell contributions $(O - E)^2/E$. Members: $(31 - 50.4)^2/50.4 = 7.467$; $(50 - 53.1)^2/53.1 = 0.181$; $(69 - 46.5)^2/46.5 = 10.887$. Non-members: 3.200; 0.078; 4.666.
3. $X^2 = 7.467 + 0.181 + 10.887 + 3.200 + 0.078 + 4.666 =$ **26.48**.
4. df = (3 − 1)(2 − 1) = 2. The 5% critical value is 5.99; $p = 1.8 \times 10^{-6}$.
5. Reading the contributions: most of X^2 comes from the extremes of education, with fewer members than expected among mothers with no schooling and more among secondary+.

IN R (OPTIONAL)

```
cb  ← read.csv("data/cbhi_survey.csv")
edu ← factor(cb$maternal_education,
             levels = c("none", "primary", "secondary+"))
tab ← table(edu, member = cb$cbhi_member)
res ← chisq.test(tab)
round(res$expected, 1)
res
```

```
            member
edu              0    1
  none       117.6 50.4
  primary    123.9 53.1
  secondary+ 108.5 46.5

    Pearson's Chi-squared test

data:  tab
X-squared = 26.479, df = 2, p-value = 1.779e-06
```

`res$expected` reproduces step 1; X-squared, df and p are steps 3 and 4. (R applies Yates' correction, §10.6.3, only to 2 × 2 tables.)

MUST-KNOW NOTES (MD-LEVEL)

- **When is the approximation adequate?** A widely used convention, traced to Cochran (1954) (Rayson et al., 2004), is that no more than 20% of cells should have an expected count below 5 and no cell an expected count below 1 (Kim, 2017). For a 2 × 2 table this is usually read as all four expected counts ≥ 5. It is a rule of thumb, not a theorem, and it concerns *expected* counts: a small observed count, even zero, is not by itself a problem. Here the smallest expected count is 46.5, so the rule is met easily.
- When the rule fails: use an exact test (Fisher's, §10.7), or merge adjacent categories if the merged categories still make sense. Yates' correction (§10.6.3) is not a remedy.
- A significant X^2 shows association, not causation or direction. Education is itself linked to wealth and distance, so this association may partly reflect those other factors (confounding, the subject of §11.6).

10.6.2 Chi-Square Goodness of Fit

DEFINITION

The **chi-square goodness-of-fit test** asks whether the counts of a single categorical variable match a hypothesised set of proportions. The expected count in category i is $E_i = N \times p_i$, where p_i is the hypothesised proportion, and $X^2 = \Sigma(O - E)^2/E$ has k − 1 degrees of freedom for k categories.

SIMPLIFIED EXPLANATION

There is only one variable, so the expected counts come from the hypothesis, not from row and column totals. Once N is fixed, the first k − 1 counts determine the last, so df = k − 1. If you also estimate parameters from the data to set the p_i, you lose one further df for each one.

ILLUSTRATION / EXAMPLE

A PHC recorded 500 deliveries in a year: 140, 110, 115 and 135 in the four quarters. Is there a seasonal pattern, or are deliveries spread evenly? Treating the quarters as equal in length, H_0 gives $p_i = 0.25$ in each.

WORKED CALCULATION

1. E = 500 × 0.25 = 125 in each quarter.
2. $(O - E)^2/E$: $(15)^2/125 = 1.8$; $(-15)^2/125 = 1.8$; $(-10)^2/125 = 0.8$; $(10)^2/125 = 0.8$.
3. $X^2 = 1.8 + 1.8 + 0.8 + 0.8 =$ **5.2**, df = 4 − 1 = 3.
4. The 5% critical value with 3 df is 7.81; p = 0.16. These data do not show clear seasonality.

MUST-KNOW NOTES (MD-LEVEL)

- Expected counts come from the hypothesis times N; df = k − 1 (minus one more for each parameter estimated from the data).

- A common use is to check a sample's make-up against known population proportions (by age group, sex or residence from the Census, recall §1.4.4). A clear misfit warns that the sample is unrepresentative, which biases descriptive estimates (recall §6.4) and may limit generalisability (§17.1.2: external validity).
- A non-significant result does not prove the hypothesised distribution is right; it shows only that these data cannot rule it out (recall §8.5).

10.6.3 Yates' Correction

DEFINITION

Yates' continuity correction modifies Pearson's X^2 for a 2 × 2 table by reducing each absolute difference $|O - E|$ by 0.5 before squaring:

$X^2_Yates = \Sigma\ (|O - E| - 0.5)^2 \div E.$

It is the 2 × 2 version of the continuity correction for approximating a discrete count by a continuous curve (recall §5.7.2.2.4).

SIMPLIFIED EXPLANATION

Counts move in whole steps, but the χ^2 curve is smooth. Pearson's X^2 tends to give p-values a little smaller than the exact ones in small 2 × 2 tables. Shrinking each gap by half a count makes X^2 smaller and p larger. The correction often overshoots, making the test conservative (p too large); the small pilot table in §10.7 (Fisher's exact test, an exact calculation for small tables) shows both errors side by side. It does not rescue a table whose expected counts are too small: that problem is solved only by an exact test (§10.7).

ILLUSTRATION / EXAMPLE

Recall the infant cohort of §8.7: 200 infants were followed from birth; 80 were not exclusively breastfed (not EBF) and 120 were. Diarrhoea during follow-up:

	Diarrhoea	No diarrhoea	Total
Not EBF	a = 33	b = 47	80
EBF	c = 14	d = 106	120
Total	47	153	200

WORKED CALCULATION

The uncorrected test on this table was worked cell by cell in §9.1.2 (step 5): expected counts 18.8, 61.2, 28.2 and 91.8, all ≥ 5; $|O - E| = 14.2$ in every cell (in a 2 × 2 table this is always so: with one free cell, every cell is off by the same amount); Pearson $X^2 =$ **23.37** $= 4.83^2$, df = 1, $p = 1.3 \times 10^{-6}$. Now add Yates' correction.

1. Shortcut formula for Pearson's X^2, same answer: $X^2 = N(ad - bc)^2 \div [(a + b)(c + d)(a + c)(b + d)] = 200 \times (3{,}498 - 658)^2 \div (80 \times 120 \times 47 \times 153) = 200 \times 2{,}840^2 \div 69{,}033{,}600 = 23.37$.

2. Yates: $(14.2 - 0.5)^2 = 13.7^2 = 187.69$, so $X^2_Yates = 187.69 \times (1/18.8 + 1/61.2 + 1/28.2 + 1/91.8) =$ **21.75**, $p = 3.1 \times 10^{-6}$. The shortcut form is $N(|ad - bc| - N/2)^2 \div [(a + b)(c + d)(a + c)(b + d)] = 200 \times 2{,}740^2 \div 69{,}033{,}600 = 21.75$.
3. Both far exceed 3.84, the 5% point of χ^2 with 1 df. With expected counts this large, the correction barely matters.

IN R (OPTIONAL)

```
e  ← read.csv("data/ebf_diarrhoea_cohort.csv")
tb ← table(not_ebf = factor(e$not_ebf, levels = 1:0),
           diarrhoea = factor(e$diarrhoea, levels = 1:0))
tb
chisq.test(tb, correct = FALSE)   # Pearson
chisq.test(tb)                    # Yates-corrected (R's 2x2 default)
```

```
       diarrhoea
not_ebf   1   0
      1  33  47
      0  14 106

    Pearson's Chi-squared test

data:  tb
X-squared = 23.367, df = 1, p-value = 1.338e-06


    Pearson's Chi-squared test with Yates' continuity correction

data:  tb
X-squared = 21.751, df = 1, p-value = 3.105e-06
```

The table is the one above; the two statistics are Pearson's 23.37 (§9.1.2) and the Yates-corrected 21.75 of step 2. Note that R applies Yates by default to every 2 × 2 table, so state which version you report.

MUST-KNOW NOTES (MD-LEVEL)

- Yates' correction applies only to 2 × 2 tables and makes X^2 smaller (when $|O - E|$ is below 0.5, only $|O - E|$ is subtracted, as R does, so the correction never increases X^2).
- It is a continuity correction, not a small-sample fix. When expected counts are below the Cochran threshold, use Fisher's exact test (§10.7), not a corrected X^2.
- Report which version you used; the two can straddle 0.05 when the result is borderline.

CHECKPOINT 10.3

1. A 2 × 3 table has 500 observations. How many degrees of freedom does the test of independence have, and why, in terms of free cells?
2. In a 2 × 2 table, a = 20, b = 30, c = 10, d = 40. Compute the expected count for cell a and the Pearson X^2. Is it significant at 5%?

3. "My 2 × 2 table has an expected count of 2.1, so I applied Yates' correction." What is wrong with this?

(Answers in the Appendix at the end of the book.)

10.6.4 Risk, Odds and Measures of Association and Impact in the 2 × 2 Table

DEFINITION

Lay out a 2 × 2 table with exposure in rows and outcome in columns: a = exposed with outcome, b = exposed without, c = unexposed with outcome, d = unexposed without.

- **Risk** in each group (cohort data only): $R_1 = a/(a + b)$ in the exposed, $R_0 = c/(c + d)$ in the unexposed (recall risk, §1.5).
- **Risk ratio (RR)** = R_1/R_0. **Risk difference (RD)** = $R_1 - R_0$, also called the **attributable risk (AR)**.
- **Odds** of the outcome (recall §5.5.3): a/b in the exposed, c/d in the unexposed. **Odds ratio (OR)** = (a/b) ÷ (c/d) = ad/bc.
- **Attributable fraction in the exposed (AR%)** = $(R_1 - R_0)/R_1 = (RR - 1)/RR$.
- **Population attributable risk (PAR)** = $R_t - R_0$, where R_t is the risk in the whole population. **Population attributable fraction (PAF)** = PAR/R_t. With p_e the proportion of the population exposed, **Levin's formula** is $PAF = p_e(RR - 1) \div [1 + p_e(RR - 1)]$ (Levin, 1953).
- **Number needed to treat (NNT)** = 1/|RD| for a beneficial exposure; for a harmful one it is the **number needed to harm (NNH)** (Laupacis et al., 1988).

SIMPLIFIED EXPLANATION

The χ^2 test (§10.6.1) asks only whether there is an association. These measures say how strong it is and how much disease it accounts for. Ratios (RR, OR) describe strength: how many times more likely. Differences (RD, NNT) describe absolute impact: how many extra cases per 100 exposed. The PAF asks what share of all cases in the population would not occur if the exposure were removed, which depends both on the RR and on how common the exposure is.

Which measures a table can give depends on the design (recall §1.5). A **cohort** samples on exposure and follows people, so the row totals are real groups at risk and R_1, R_0, RR and RD are all valid. A **case-control** study samples on outcome: the investigator fixes the column totals (how many cases, how many controls). The rows no longer contain a population at risk, so risks, RR and RD cannot be computed. Only the OR survives, and it survives exactly (derivation below). A cross-sectional survey gives prevalences and their ratio (§13.7 estimates prevalence ratios directly).

DERIVATION

The OR survives case-control sampling. Suppose a case-control study takes all cases but only a fraction f of the non-cases. The table becomes a, fb, c, fd. The OR is a(fd) ÷ (fb × c) = ad/bc, unchanged. The "risk" a/(a + fb), by contrast, depends on f, which the investigator chose.

The exposure OR equals the disease OR. A case-control study compares the odds of exposure in cases (a/c) with that in controls (b/d): (a/c) ÷ (b/d) = ad/bc. This is the same number as the odds ratio of disease, so a case-control study estimates the cohort's OR.

When OR ≈ RR. OR = RR × $(1 - R_0)/(1 - R_1)$. When the outcome is rare in both groups, both brackets are near 1 and OR ≈ RR. When it is common, the OR lies further from 1 than the RR.

Levin's formula. The population risk is a weighted average: $R_t = p_e R_1 + (1 - p_e)R_0 = R_0[1 + p_e(RR - 1)]$. So $PAF = (R_t - R_0)/R_t = p_e(RR - 1) \div [1 + p_e(RR - 1)]$.

Standard errors on the log scale. Ratios are skewed (they cannot go below 0 but can go far above 1), so their CIs are built for ln RR and ln OR and then exponentiated back.

1. A small change in p changes ln p by about (change in p) ÷ p. So $\text{Var}(\ln \hat{p}) \approx \text{Var}(\hat{p})/p^2 = [p(1 - p)/n]/p^2 = 1/(np) - 1/n$, estimated by $1/a - 1/(a + b)$ for R_1.
2. The two groups are independent, so variances add: $\text{SE}(\ln RR) = \sqrt{[1/a - 1/(a + b) + 1/c - 1/(c + d)]}$.
3. Similarly, a small change in p changes $\ln[p/(1 - p)]$ by about (change in p) ÷ $[p(1 - p)]$, so $\text{Var}(\ln \text{odds}) \approx 1/(np) + 1/(n(1 - p))$, estimated by $1/a + 1/b$. Adding both groups gives **Woolf's** formula $\text{SE}(\ln OR) = \sqrt{(1/a + 1/b + 1/c + 1/d)}$ (Woolf, 1955).
4. 95% CI: exp(ln estimate ± 1.96 × SE) (recall §7.2).

ILLUSTRATION / EXAMPLE

The EBF cohort from §10.6.3: 80 infants not exclusively breastfed, 33 with diarrhoea; 120 exclusively breastfed, 14 with diarrhoea. Here "exposure" is *not* being exclusively breastfed.

WORKED CALCULATION

1. Risks: $R_1 = 33/80 =$ **0.412**; $R_0 = 14/120 =$ **0.117**.
2. RR = 0.4125/0.1167 = **3.54**. Not-EBF infants had 3.5 times the risk of diarrhoea during follow-up.
3. $\text{SE}(\ln RR) = \sqrt{(1/33 - 1/80 + 1/14 - 1/120)} = \sqrt{0.0809} = 0.284$. ln RR = 1.263. CI: exp(1.263 ± 1.96 × 0.284) = exp(0.705) to exp(1.820) = **2.02 to 6.17**.
4. Odds: 33/47 = 0.702 and 14/106 = 0.132. OR = (33 × 106)/(47 × 14) = 3,498/658 = **5.32**.
5. $\text{SE}(\ln OR) = \sqrt{(1/33 + 1/47 + 1/14 + 1/106)} = \sqrt{0.1324} = 0.364$. ln OR = 1.671. CI: exp(1.671 ± 0.713) = **2.61 to 10.85**.
6. OR (5.32) is well above RR (3.54) because diarrhoea is common here (23.5% overall): 3.54 × (1 − 0.117)/(1 − 0.412) = 3.54 × 1.503 = 5.32.
7. RD (attributable risk) = 0.4125 − 0.1167 = **0.296**: 29.6 extra infants with diarrhoea per 100 not exclusively breastfed. SE(RD) = √(0.4125 × 0.5875/80 + 0.1167 × 0.8833/120) = 0.062, so 95% CI 0.174 to 0.418 (recall §7.2.3).
8. NNH = 1/0.296 = **3.4**: for about every 3 to 4 infants not exclusively breastfed, one extra has diarrhoea (95% CI 2.4 to 5.8, the reciprocals of the RD limits). Read the other way, if the association were causal, about 3 to 4 infants would need to be exclusively breastfed to prevent one case.
9. AR% = (3.54 − 1)/3.54 = **71.7%**: of the diarrhoea cases among not-EBF infants, about 72% would be attributable to not being exclusively breastfed, if the association is causal.
10. $p_e = 80/200 = 0.40$; $R_t = 47/200 = 0.235$. PAR = 0.235 − 0.117 = **0.118**. PAF = 0.118/0.235 = 50.4%, and Levin's formula gives the same: 0.40 × 2.54 ÷ (1 + 0.40 × 2.54) = 1.014 ÷ 2.014 = **50.4%**.
11. Case-control check: keep all 47 cases but only half the 153 controls (f = 0.5, so 23.5 and 53 on average). The OR is 33 × 53 ÷ (23.5 × 14) = 5.32, unchanged, but the "risk" in the exposed becomes 33/56.5 = 0.58 instead of 0.41, which is meaningless.

IN R (OPTIONAL)

```
library(epiR)
e  ← read.csv("data/ebf_diarrhoea_cohort.csv")
tb ← table(not_ebf = factor(e$not_ebf, levels = 1:0),
           diarrhoea = factor(e$diarrhoea, levels = 1:0))
epi.2by2(tb, method = "cohort.count")
```

```
             Outcome +    Outcome -      Total                 Inc risk *
Exposed +           33           47         80     41.25 (30.35 to 52.82)
Exposed -           14          106        120      11.67 (6.53 to 18.80)
Total               47          153        200     23.50 (17.81 to 30.00)

Point estimates and 95% CIs:
-------------------------------------------------------------------
Inc risk ratio                                 3.54 (2.02, 6.17)
Inc odds ratio                                 5.32 (2.61, 10.85)
Attrib risk in the exposed *                   29.58 (17.36, 41.80)
Attrib fraction in the exposed (%)            71.72 (50.61, 83.80)
Attrib risk in the population *                11.83 (3.62, 20.05)
Attrib fraction in the population (%)         50.35 (25.89, 66.74)
-------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 23.367 Pr>chi2 = <0.001
...
 * Outcomes per 100 population units
```

`epi.2by2` expects the exposed row first and the outcome-present column first, which is why the factor levels are set to `1:0`. Each line matches a step: risks per 100 (step 1), RR and OR with the log-scale CIs (steps 3 and 5), attributable risk 29.58 per 100 (step 7), AR% (step 9), PAR 11.83 per 100 and PAF 50.35% (step 10), and Pearson's X^2 (§10.6.3). `epitools::riskratio()` and `epitools::oddsratio()` give the same RR and OR one at a time.

MUST-KNOW NOTES (MD-LEVEL)

- Cohort → risk, RR, RD, NNT, AR%, PAF. Case-control → OR only (it estimates the cohort OR). Cross-sectional → prevalence, prevalence ratio or prevalence OR.
- The OR is not a risk ratio. It approximates the RR only when the outcome is rare; with a common outcome, calling an OR of 5.32 "five times the risk" exaggerates a real RR of 3.54.
- The null value of a ratio is 1 and of a difference is 0. A CI for RR or OR is asymmetric on the ordinary scale because it is symmetric on the log scale.
- AR%, PAF and NNT assume the association is causal and unconfounded. They are statements about what removing the exposure *would* do. The EBF result has not yet been checked for confounding by PHC (§11.7 does this).
- PAF needs p_e from the population. Using the cohort's 40% assumes the cohort is representative.
- A quick link to the χ^2 test: a **test-based** SE is ln OR ÷ $\sqrt{X^2}$ = 1.671 ÷ 4.83 = 0.346, which gives a CI of 2.70 to 10.47, close to Woolf's. Both tests address the same H_0 (OR = 1).

10.6.5 Chi-Square Test for Trend

DEFINITION

When the rows of a 2-column table are ordered categories (education level, wealth quintile, age group), the **chi-square test for trend** (the Cochran–Armitage test) tests whether the proportion with the outcome rises or falls steadily across the ordered categories. Give each category a score s_i (usually 0, 1, 2, ...). With x_i outcomes out of n_i in category i, $\bar{p} = \Sigma x_i/N$ and $\bar{s} = \Sigma n_i s_i/N$:

$T = \Sigma x_i s_i - \bar{p}\Sigma n_i s_i$, $V = \bar{p}(1 - \bar{p})[\Sigma n_i s_i^2 - (\Sigma n_i s_i)^2/N]$, $X^2_trend = T^2/V$, on 1 df.

SIMPLIFIED EXPLANATION

The ordinary r × 2 test (§10.6.1) asks whether the proportions differ in any way, and spends r − 1 df to do it. If you expect a steady gradient, put all the evidence into one question, "is there a slope?", which costs only 1 df. The test fits a straight line through the proportions (weighted by group size) and asks whether its slope differs from zero (Armitage, 1955). What is left, X^2(total) − X^2_trend on r − 2 df, tests **departure from linear trend**.

DERIVATION

1. Because $\bar{p}\Sigma n_i s_i = \Sigma x_i \times \bar{s}$, $T = \Sigma x_i(s_i - \bar{s})$: it compares the scores of the people with the outcome with the average score.
2. Treat each person's outcome as 0 or 1 and fit a straight line of outcome on score. Its slope is b = T ÷ S_ss, with $S_ss = \Sigma n_i(s_i - \bar{s})^2 = \Sigma n_i s_i^2 - (\Sigma n_i s_i)^2/N$ (the least-squares slope, a formula derived in Chapter 12).
3. Under H_0 (no trend) each outcome has variance $\bar{p}(1 - \bar{p})$ (recall §5.7.1.1), so $Var(T) = \bar{p}(1 - \bar{p})S_ss = V$.
4. $T \div \sqrt{V}$ is approximately standard Normal, so $X^2_trend = T^2/V$ is approximately χ^2 with 1 df (recall §5.7.2.4).

ILLUSTRATION / EXAMPLE

The CBHI membership table from §10.6.1: members were 31/168 (18.5%) with no maternal education, 50/177 (28.2%) with primary and 69/155 (44.5%) with secondary or higher. Scores 0, 1, 2.

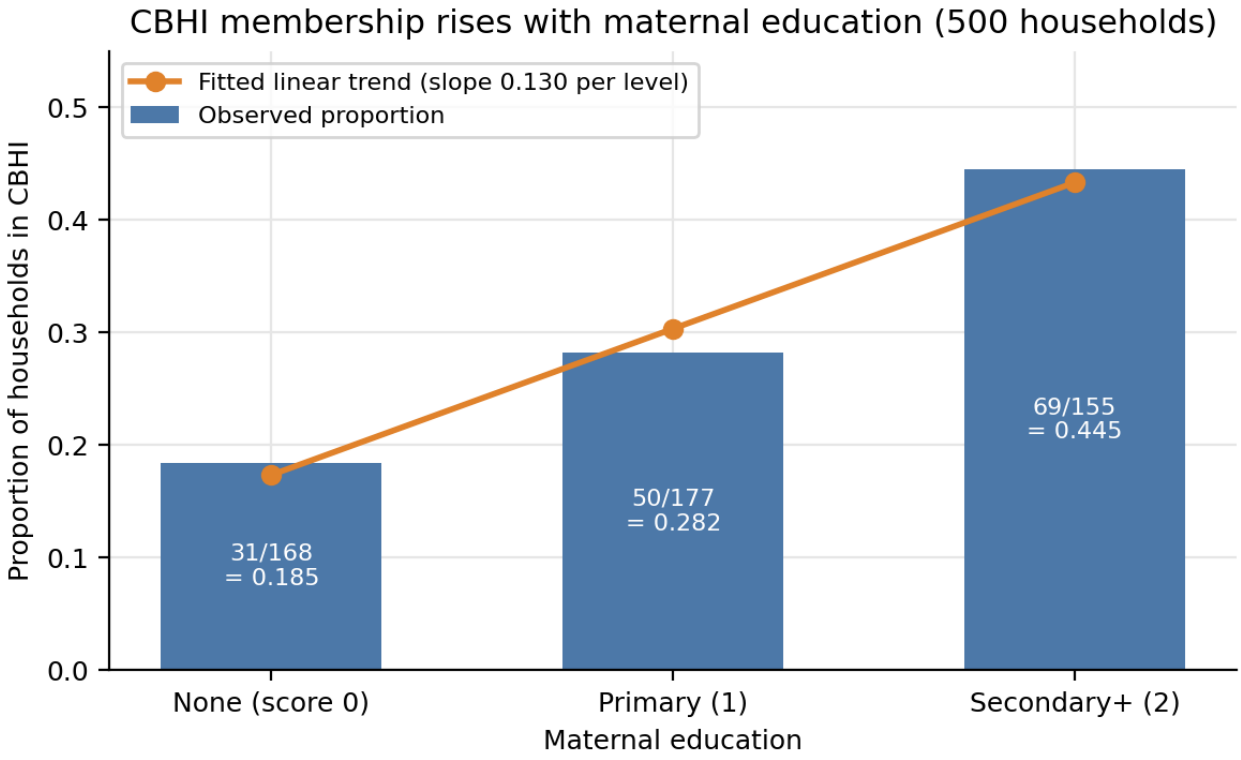


Figure 10.2. CBHI membership by maternal education (500 households). The line is the fitted linear trend, rising 0.130 per education level; the bars sit close to it, so there is little departure from linearity.

WORKED CALCULATION

1. $\bar{p} = 150/500 = 0.30$. $\Sigma x_i s_i = 0 \times 31 + 1 \times 50 + 2 \times 69 = 188$. $\Sigma n_i s_i = 0 \times 168 + 177 + 2 \times 155 = 487$. $\Sigma n_i s_i^2 = 177 + 4 \times 155 = 797$.
2. $T = 188 - 0.30 \times 487 = 188 - 146.1 = 41.9$.
3. $S_{ss} = 797 - 487^2/500 = 797 - 474.34 = 322.66$. $V = 0.30 \times 0.70 \times 322.66 = 67.76$.
4. $X^2_trend = 41.9^2 \div 67.76 = 1{,}755.6 \div 67.76 =$ **25.91**, 1 df, $p = 3.6 \times 10^{-7}$.
5. Slope $b = 41.9 \div 322.66 = 0.130$: membership rises by about 13 percentage points per education level.
6. Departure from trend: $26.48 - 25.91 =$ **0.57** on $3 - 2 = 1$ df, $p = 0.45$. The rise is well described by a straight line.

IN R (OPTIONAL)

```
cb  ← read.csv("data/cbhi_survey.csv")
edu ← factor(cb$maternal_education,
             levels = c("none", "primary", "secondary+"))
x ← tapply(cb$cbhi_member, edu, sum); n ← table(edu)
prop.trend.test(x, n, score = 0:2)     # 1 df: linear trend
chisq.test(cbind(x, n - x))$statistic  # 2 df: any difference
```

```
    Chi-squared Test for Trend in Proportions

data:  x out of n ,
 using scores: 0 1 2
X-squared = 25.91, df = 1, p-value = 3.578e-07

X-squared
 26.47934
```

`prop.trend.test` gives step 4 (25.91 on 1 df). The second line is the overall X^2 of §10.6.1; their difference is the departure-from-trend χ^2 of step 6.

MUST-KNOW NOTES (MD-LEVEL)

- Use it when categories are ordered and a gradient is the question (a dose–response pattern). It is more powerful than the r × c test against a steady trend, and blind to a U-shape.
- The result depends on the scores chosen. Equally spaced scores assume equal steps between levels; state your scores.
- A significant trend is an association. Here education travels with wealth and distance, so the gradient is not yet an effect of education (§11.6). §12.4.5 extends trend tests to regression models.

CHECKPOINT 10.4

1. In a cohort, 30 of 100 smokers and 10 of 100 non-smokers develop a cough. Compute RR, OR, RD and NNH. Why do RR and OR differ?
2. Why can a case-control study not estimate the risk ratio directly, although it can estimate the odds ratio?
3. Using Levin's formula, find the PAF when 20% of the population is exposed and RR = 3.

(Answers in the Appendix at the end of the book.)

10.7 Fisher's Exact Test

DEFINITION

Fisher's exact test computes the exact p-value for association in a 2 × 2 table, treating the row and column totals as fixed. With those totals fixed, the count a follows a **hypergeometric distribution**: the distribution of the number of "successes" when you draw a fixed number of items *without replacement* from a group holding a known number of successes and failures. With row 1 total r_1, column 1 total c_1 and grand total N,

$P(a) = C(c_1, a) \times C(N - c_1, r_1 - a) \div C(N, r_1)$.

SIMPLIFIED EXPLANATION

Picture N cards, c_1 of them marked "outcome". Under H_0 (no association) the exposure labels are dealt at random: r_1 cards go to the exposed row. How many marked cards land in that row? Choose which a of the c_1 marked cards ($C(c_1, a)$ ways) and which $r_1 - a$ of the unmarked ones ($C(N - c_1, r_1 - a)$ ways), out of $C(N, r_1)$ equally likely deals (recall combinations, §5.7.1.2). This is the binomial idea without replacement: each card dealt changes what is left. The p-value adds up the probabilities of all tables at least as unlikely as the one observed. No large-sample approximation is used, so it is valid however small the counts.

ILLUSTRATION / EXAMPLE

A pilot study at an anganwadi follows 14 infants: 6 not exclusively breastfed, of whom 4 get diarrhoea; 8 exclusively breastfed, of whom 1 does.

	Diarrhoea	None	Total
Not EBF	4	2	6
EBF	1	7	8
Total	5	9	14

The expected counts are 2.14, 3.86, 2.86 and 5.14. Three of the four (75%) are below 5, far outside Cochran's convention (§10.6.1), so the χ^2 approximation cannot be trusted.

WORKED CALCULATION

1. Margins fixed: $r_1 = 6$, $c_1 = 5$, $N = 14$. $C(14, 6) = 3{,}003$ equally likely deals.
2. $P(a = 4) = C(5, 4) \times C(9, 2) \div 3{,}003 = 5 \times 36 \div 3{,}003 = \mathbf{0.0599}$.
3. All possible tables: $P(a = 0, 1, 2, 3, 4, 5) = 0.0280, 0.2098, 0.4196, 0.2797, 0.0599, 0.0030$.
4. Two-sided p = sum of P for tables no more probable than the observed one: a = 0, 4, 5: 0.0280 + 0.0599 + 0.0030 = **0.091**.
5. Compare: Pearson X^2 gives p = 0.036 (a false "significant" result from a poor approximation), and Yates-corrected X^2 gives p = 0.126 (overshooting the other way). Only the exact test is right here.

IN R (OPTIONAL)

```
pilot ← matrix(c(4, 2,
                 1, 7), nrow = 2, byrow = TRUE,
               dimnames = list(c("not EBF", "EBF"),
                               c("diarrhoea", "none")))
fisher.test(pilot)
dhyper(0:5, m = 5, n = 9, k = 6)   # P(a = 0..5) with margins fixed
```

```
    Fisher's Exact Test for Count Data

data:  pilot
p-value = 0.09091
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.6418261 779.1595463
sample estimates:
odds ratio
  10.98111

[1] 0.027972028 0.209790210 0.419580420 0.279720280 0.059940060 0.002997003
```

The p-value 0.091 is step 4, and `dhyper` lists the six probabilities of step 3. R's OR (10.98) is a conditional estimate, slightly different from ad/bc = 14. Its CI of 0.64 to 779 shows how little 14 infants can tell you.

MUST-KNOW NOTES (MD-LEVEL)

- Use it when the expected counts fail Cochran's convention (§10.6.1). It is valid for any sample size; software computes it for large tables too.
- Because a can take only a few values, the achievable p-values jump in steps, and the test tends to be conservative (its true Type I error rate is at or below α). The **mid-p** version counts only half the probability of the observed table and is less conservative (here 0.066, computed as twice the one-sided mid-p).
- It conditions on both margins. In most studies only one margin (a cohort's exposure groups, or a case-control study's cases and controls) is fixed by design; the test is still widely used and valid, just conservative.
- Extensions to larger r × c tables exist, and software computes them.

10.8 McNemar's Test

DEFINITION

McNemar's test (McNemar, 1947) tests paired binary data: the same subjects measured twice, two raters classifying the same subjects, or matched case-control pairs. Arrange the pairs in a 2 × 2 table of pairs. Let b and c be the two kinds of **discordant pairs** (the members of the pair differ). Then

X^2_McNemar = $(b - c)^2 \div (b + c)$, on 1 df.

For a matched case-control study the **matched-pair odds ratio** is OR = b/c, with SE(ln OR) = $\sqrt{(1/b + 1/c)}$.

SIMPLIFIED EXPLANATION

The chi-square test of independence assumes independent observations. Paired data are not independent: a household measured before and after an education drive is the same household. Pairs in which

both members agree (**concordant pairs**) say nothing about which way things shifted. Only the discordant pairs carry information, and under H_0 each of them is equally likely to go either way.

DERIVATION

1. Under H_0, each of the b + c discordant pairs is of type "b" with probability 0.5. So b ~ Binomial(b + c, 0.5): this is the sign test (§10.1.1) applied to the discordant pairs.
2. For large b + c, b is approximately Normal with mean (b + c)/2 and SD $\sqrt{[(b + c)/4]}$ (recall §5.7.2.2.4).
3. $z = [b - (b + c)/2] \div \sqrt{[(b + c)/4]} = (b - c) \div \sqrt{(b + c)}$, so $z^2 = (b - c)^2 \div (b + c)$.
4. In a matched case-control study, the concordant pairs match cases and controls on exposure and so cannot contribute to a contrast within the pair. Among discordant pairs, the odds that the case (rather than the control) is the exposed one is b/c, which is the matched OR. The same estimate comes out of conditional logistic regression (§13.2.9, a regression model that compares each case only with its own matched controls).

ILLUSTRATION / EXAMPLE

Before and after. To test whether CBHI membership changes after a community education drive in the same 100 households, a χ^2 test of independence would wrongly treat "before" and "after" as two independent samples. McNemar's test compares the households that joined with those that dropped out.

Matched case-control. Eighty children with diarrhoea (cases) were each matched to a control child of the same age from the same village. Exposure: not exclusively breastfed in infancy.

	Control exposed	Control not exposed
Case exposed	16 (concordant)	b = 24
Case not exposed	c = 10	30 (concordant)

WORKED CALCULATION

1. Discordant pairs: b = 24, c = 10; b + c = 34. The 46 concordant pairs drop out.
2. $X^2 = (24 - 10)^2 \div 34 = 196 \div 34 =$ **5.76**, 1 df, p = 0.016.
3. With a continuity correction: $(|24 - 10| - 1)^2 \div 34 = 169 \div 34 = 4.97$, p = 0.026. The exact binomial version (the sign test of §10.1.1 on the 34 discordant pairs) gives p = 0.024.
4. Matched OR = 24/10 = **2.4**. SE(ln OR) = $\sqrt{(1/24 + 1/10)} = 0.376$. CI: $\exp(0.875 \pm 1.96 \times 0.376) =$ **1.15 to 5.02**.
5. Ignoring the matching: pooled into an ordinary 2 × 2 table, 40 of 80 cases and 26 of 80 controls were exposed, and ad/bc = (40 × 54)/(40 × 26) = 2.08. In this dataset the unmatched analysis gives a different, smaller OR. Matched data must be analysed as matched.

IN R (OPTIONAL)

```
m  ← read.csv("data/matched_casecontrol.csv")
cs ← m[m$case == 1, ]; ct ← m[m$case == 0, ]
pp ← merge(cs, ct, by = "pair_id", suffixes = c(".case", ".ctrl"))
pt ← table(case = factor(pp$not_ebf.case, levels = 1:0),
           control = factor(pp$not_ebf.ctrl, levels = 1:0))
pt
mcnemar.test(pt, correct = FALSE)
b ← pt[1, 2]; c ← pt[2, 1]
# matched OR and its 95% CI
exp(log(b / c) + c(0, -1.96, 1.96) * sqrt(1/b + 1/c))
```

```
    control
case  1  0
   1 16 24
   0 10 30

    McNemar's Chi-squared test

data:  pt
McNemar's chi-squared = 5.7647, df = 1, p-value = 0.01635

[1] 2.400000 1.147690 5.018775
```

`merge` puts each case beside its own control, and the table of pairs is the one above. McNemar's X^2 is step 2, and the last line is the matched OR with its CI from step 4. (`mcnemar.test(pt)` without `correct = FALSE` gives the corrected 4.97.)

MUST-KNOW NOTES (MD-LEVEL)

- Paired binary data → McNemar; independent groups → χ^2 or Fisher. It is the categorical counterpart of the paired t-test.
- Only discordant pairs enter the test and the matched OR. When b + c is small, use the exact binomial version (step 3), for the same reason as in §10.1.1.
- For paired data with more than two categories, Bowker's test of symmetry extends the same discordant-pairs logic.
- Chance-corrected agreement between two raters on the same subjects is a different question, answered by kappa (§16.3.3.1); McNemar tests only whether one rater says "yes" more often than the other.

CHECKPOINT 10.5

1. A 2 × 2 table has three expected counts below 5. Which test should you use, and why is Yates' correction not the answer?
2. In a before–after survey of 150 households, 20 joined CBHI and 8 left; the rest did not change. Compute McNemar's X^2 and state your conclusion at the 5% level.
3. In a matched case-control study, b = 30 and c = 12. Compute the matched OR and its 95% CI.

(Answers in the Appendix at the end of the book.)

Chapter map: which test for which comparison

Outcome	Two independent groups	Paired / matched	Three or more independent groups	Three or more related
Ordinal, or quantitative non-Normal	Mann-Whitney (§10.2)	Sign test (ordinal, §10.1.1); Wilcoxon signed-rank (interval, §10.3)	Kruskal-Wallis + Dunn (§10.4)	Friedman (§10.5)
Binary or categorical	χ^2 (§10.6.1); Fisher if expected counts are small (§10.7); RR, OR, RD (§10.6.4)	McNemar; matched OR = b/c (§10.8)	r × c χ^2 (§10.6.1); trend test if ordered (§10.6.5)	Cochran's Q (not covered)

Correlation

The scatter diagram (recall §4.4) showed that older children in the under-five clinic tend to weigh more. This chapter turns that picture into numbers. It defines covariance and Pearson's r with their test and confidence interval (§11.2), the rank-based coefficients of Spearman and Kendall (§11.3) and the coefficients for binary variables (§11.4). It then asks what an association can and cannot tell you about cause (§11.5), builds the idea of confounding and draws it with causal diagrams (§11.6), and closes with the first method that removes a confounder's distortion: stratified analysis with the Mantel–Haenszel estimators (§11.7).

11.1 The Idea of Association

DEFINITION

Two variables are **associated** when the values of one tend to change in a systematic way as the values of the other change. The association is **positive** if they tend to rise together, **negative** if one tends to fall as the other rises, and absent if there is no such tendency.

SIMPLIFIED EXPLANATION

Association is a description before it is a number. Ask: as one variable goes up, does the other tend to move, and in which direction? "Tend to" matters. Taller people are heavier on average, but some short people outweigh some tall ones.

ILLUSTRATION / EXAMPLE

Height and weight are positively associated. Hours of sunshine and a household's heating costs are negatively associated. Shoe size and IQ in adults show no meaningful association.

MUST-KNOW NOTES (MD-LEVEL)

- An association can be **linear** (points scatter about a straight line; the target of Pearson's r, §11.2), **monotonic** (always rising or always falling, but not necessarily straight; the target of Spearman's and Kendall's coefficients, §11.3) or neither. A U-shaped relationship (falling, then rising) can give a coefficient near 0 although the two variables are strongly related. Draw the scatter diagram first.
- Association is not effect. Whether one variable causes the other is a separate question, taken up in §11.5–11.7.

11.2 Pearson Correlation Coefficient

DEFINITION

The **covariance** of two quantitative variables X and Y, measured on the same n individuals, is

$$\text{cov}(X, Y) = \Sigma(x - \bar{x})(y - \bar{y}) \div (n - 1)$$

The **Pearson correlation coefficient** r is the covariance divided by the product of the two standard deviations:

$$r = \text{cov}(X, Y) \div (s_x \times s_y) = \Sigma(x - \bar{x})(y - \bar{y}) \div \sqrt{[\Sigma(x - \bar{x})^2 \times \Sigma(y - \bar{y})^2]}$$

It measures the strength and direction of the *linear* relationship and always lies between −1 and +1.

SIMPLIFIED EXPLANATION

Covariance is the two-variable cousin of the variance (recall §3.6). The variance averages $(x - \bar{x}) \times (x - \bar{x})$, a deviation times itself. Covariance averages $(x - \bar{x}) \times (y - \bar{y})$, one variable's deviation times the other's.

Look at the sign of each product. Draw a vertical line at $\bar{x}$ and a horizontal line at $\bar{y}$ (Figure 11.1). A child who is older *and* heavier than average sits in the upper right: plus × plus = plus. A child younger *and* lighter sits in the lower left: minus × minus = plus. Children in the other two corners give negative products. When most points fall in the two "plus" corners, the sum of products is large and positive. When the points spread evenly over all four corners, the products cancel and the covariance is near 0.

Covariance has awkward units (months × kg) and changes if you weigh in pounds. Dividing by $s_x \times s_y$ removes the units. That is r.

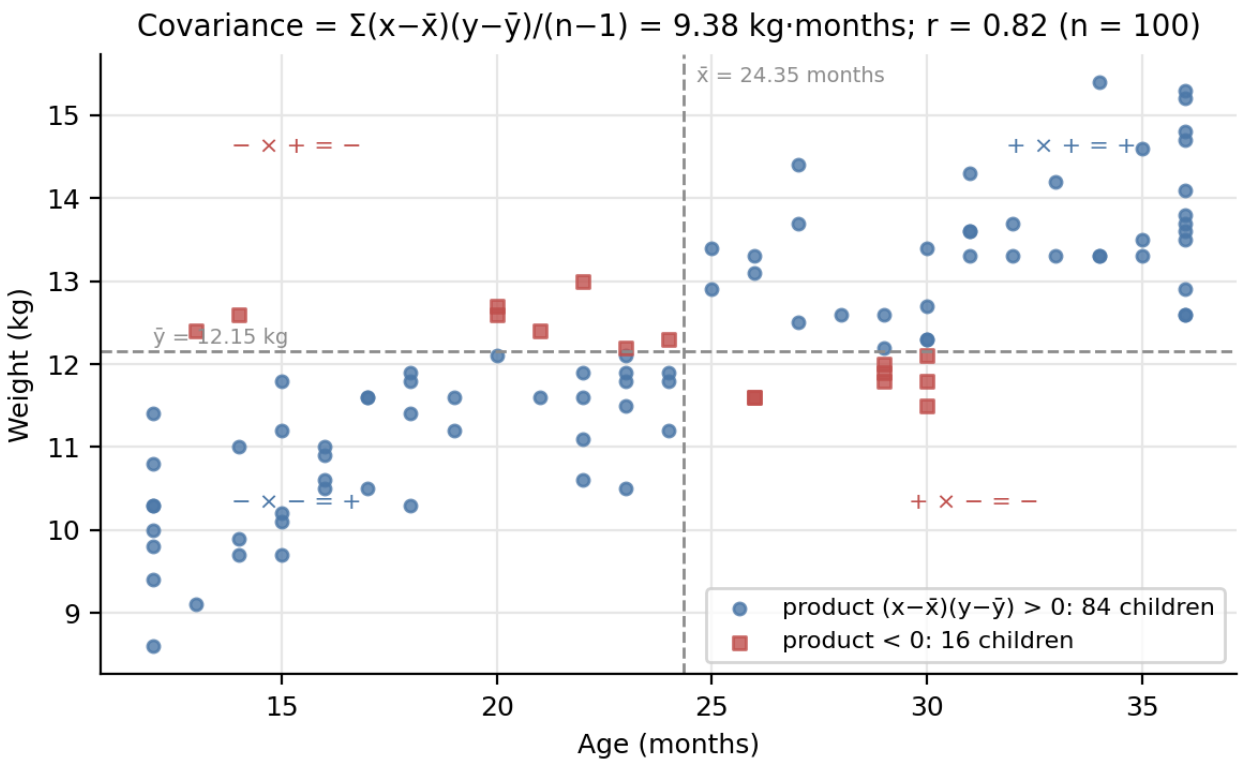


Figure 11.1. Age against weight for the 100 children in the clinic register, with dashed lines at the two means. The 84 blue points lie in the two quadrants where $(x - \bar{x})(y - \bar{y})$ is positive; the 16 red points give negative products. The positive products dominate, so the covariance (9.38 kg × months) and r (0.82) are positive.

DERIVATION

Why r cannot exceed 1 in size. Write each value as a z-score (recall §5.7.2.2.1): $z_x = (x - \bar{x})/s_x$ and $z_y = (y - \bar{y})/s_y$. Then:

1. $r = \Sigma z_x z_y \div (n - 1)$: r is the average product of the two z-scores.
2. Each set of z-scores has $\Sigma z^2 = n - 1$, because $\Sigma(x - \bar{x})^2 = (n - 1)s_x^2$.
3. A sum of squares cannot be negative: $\Sigma(z_x - z_y)^2 \geq 0$.
4. Expand: $\Sigma z_x^2 + \Sigma z_y^2 - 2\Sigma z_x z_y = (n - 1) + (n - 1) - 2(n - 1)r = 2(n - 1)(1 - r) \geq 0$, so $r \leq 1$.
5. The same argument with $\Sigma(z_x + z_y)^2 \geq 0$ gives $r \geq -1$.
6. $r = +1$ only when every $z_x = z_y$, that is, when all points lie exactly on a rising straight line.

ILLUSTRATION / EXAMPLE

The first eight children to arrive at the clinic today had these heights (x, cm) and weights (y, kg):

Child	1	2	3	4	5	6	7	8
Height (cm)	80.5	78.5	96.2	74.4	89.9	96.4	85.5	86.4
Weight (kg)	11.2	9.8	14.1	10.0	12.3	13.6	11.1	11.9

WORKED CALCULATION

Use the computing forms $S_{xy} = \Sigma xy - (\Sigma x)(\Sigma y)/n$, $S_{xx} = \Sigma x^2 - (\Sigma x)^2/n$ and $S_{yy} = \Sigma y^2 - (\Sigma y)^2/n$, which give the same sums of products and squares without subtracting a mean from each value.

1. Totals: $\Sigma x = 687.8$, $\Sigma y = 94.0$, $\Sigma x^2 = 59{,}582.48$, $\Sigma y^2 = 1{,}121.36$, $\Sigma xy = 8{,}165.34$; $n = 8$.
2. $S_{xy} = 8{,}165.34 - (687.8 \times 94.0)/8 = 8{,}165.34 - 8{,}081.65 = 83.69$.
3. $S_{xx} = 59{,}582.48 - 687.8^2/8 = 59{,}582.48 - 59{,}133.605 = 448.875$. $S_{yy} = 1{,}121.36 - 94.0^2/8 = 1{,}121.36 - 1{,}104.50 = 16.86$.
4. Covariance $= 83.69 \div 7 = 11.96$ cm·kg. Every one of the eight products $(x - \bar{x})(y - \bar{y})$ is positive: each child is either taller and heavier, or shorter and lighter, than the average of the eight.
5. $r = 83.69 \div \sqrt{(448.875 \times 16.86)} = 83.69 \div \sqrt{7{,}568.0} = 83.69 \div 86.99 =$ **0.96**.

Height and weight in these eight children lie close to a rising straight line. For all 100 children in the register, the covariance of age and weight is 9.38 kg·months, $s_x = 7.91$ months and $s_y = 1.45$ kg, so $r = 9.38 \div (7.91 \times 1.45) = 9.38 \div 11.47 =$ **0.82**.

Testing r and putting a confidence interval on it

The r of a sample estimates the population correlation ρ (rho). To test H_0: $\rho = 0$ (recall the steps of §8.7), use

$t = r\sqrt{(n - 2)} \div \sqrt{(1 - r^2)}$, with $n - 2$ degrees of freedom.

If both variables are Normally distributed, t follows the t-distribution on $n - 2$ degrees of freedom when $\rho = 0$ (R Core Team, documentation of R's correlation test; derivation beyond this book, see Bland, 2015). Two degrees of freedom are lost because the same t tests whether the slope of the best-fitting straight line is zero (§12.2.4, the regression slope test), and fixing a straight line uses up two estimated quantities: its level (intercept) and its slope.

A confidence interval cannot be built as r ± 1.96 × SE, because r is trapped between −1 and +1 and its sampling distribution is skewed when ρ is far from 0. Fisher's **z-transformation** fixes this (Fisher, 1915, derived the distribution of r):

$z = \frac{1}{2} \ln[(1 + r) \div (1 - r)]$, with $SE(z) = 1 \div \sqrt{(n - 3)}$

z is approximately Normal. Build the interval on the z scale and transform its ends back with $r = (e^{2z} - 1) \div (e^{2z} + 1)$.

WORKED CALCULATION

1. Eight children, r = 0.962: $t = 0.962 \times \sqrt{6} \div \sqrt{(1 - 0.9254)} = 2.356 \div 0.273 =$ **8.63** on 6 df, p = 0.0001.
2. Fisher $z = \frac{1}{2} \ln(1.962 \div 0.038) = 1.972$. $SE = 1 \div \sqrt{5} = 0.447$. 95% limits on the z scale: $1.972 \pm 1.96 \times 0.447 = 1.096$ to 2.849. Back-transformed: **0.80 to 0.99**.
3. All 100 children, age and weight, r = 0.820: $t = 0.820 \times \sqrt{98} \div \sqrt{(1 - 0.673)} = 8.119 \div 0.572 =$ **14.19** on 98 df, p < 0.001.
4. $z = \frac{1}{2} \ln(1.820 \div 0.180) = 1.157$. $SE = 1 \div \sqrt{97} = 0.102$. Limits $1.157 \pm 0.199 = 0.958$ to 1.356. Back-transformed: **0.74 to 0.88**.

The small sample gives a wide interval (0.80 to 0.99) even though its r is higher. The interval is not symmetric about r: the upper limit cannot pass 1.

r and the regression slope

The slope of the regression line of weight on age (the kilograms gained per month of age, built in §12.2) is $b = r \times s_y \div s_x = 0.82 \times 1.45 \div 7.91 =$ **0.150 kg/month**. The two answer different questions. r is unit-free and symmetric: it measures how tightly the points hug a line. b has units and direction: it says how much y changes per unit of x. A shallow line with little scatter gives r near 1 and a small b. The t of 14.19 above is exactly the t that §12.2.4 obtains for b, because testing ρ = 0 and testing that the slope is zero ask the same question.

IN R (OPTIONAL)

```
cl ← read.csv("data/clinic_children.csv")
cov(cl$age_months, cl$weight_kg)
cor.test(cl$age_months, cl$weight_kg)
```

```
[1] 9.379596

    Pearson's product-moment correlation

data:  cl$age_months and cl$weight_kg
t = 14.188, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7434410 0.8754941
sample estimates:
      cor
0.8201033
```

`cov` returns the covariance (9.38 kg·months). `cor.test` prints r (0.82), the t statistic of step 3 with its 98 df, and the Fisher-z interval of step 4 (0.74 to 0.88).

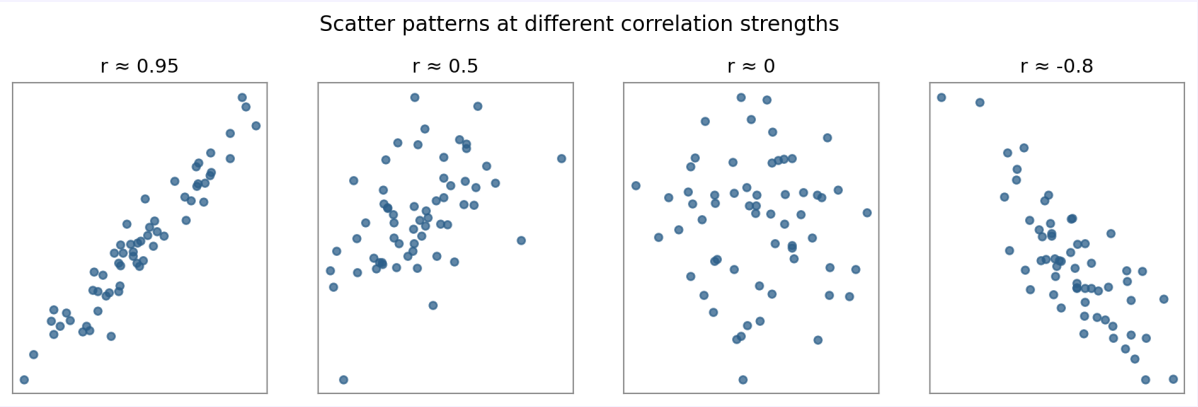


Figure 11.2. Scatter patterns at different correlation strengths.

MUST-KNOW NOTES (MD-LEVEL)

- r^2 is the proportion of the variance in one variable that is accounted for by its linear relationship with the other. For age and weight, $r^2 = 0.82^2 = 0.67$: 67% of the variation in weight goes with age. It reappears as R^2 in §12.2.5.
- Verbal labels for the size of r ("weak", "moderate", "strong") are conventions and differ between authors. Mukaka (2012), for example, calls 0.2 "negligible" and 0.3 "low positive". Report r with its confidence interval rather than a label alone.
- The t-test and Fisher interval assume the pairs are independent and the two variables are jointly Normal (Schober et al., 2018). For skewed data, ordinal data or outliers, use a rank coefficient (§11.3).
- r is sensitive to outliers because it is built from means and squared deviations. One extreme point can create or destroy a correlation. Look at the scatter diagram.
- r depends on the range sampled. Restrict the clinic register to the 26 children aged 24–30 months and r falls from 0.82 to −0.02: over six months the age effect is small beside child-to-child scatter, although children still gain weight with age.
- A significant r is not a large r. With n = 1,000, r = 0.07 gives p ≈ 0.03 but explains 0.5% of the variance.
- r measures association, not agreement. Two methods of measuring blood pressure can correlate at 0.95 while one reads 10 mmHg higher than the other. Bland and Altman (1986) showed why correlation is misleading for method comparison; the right method (limits of agreement, built on the differences between the two readings) is taught in §16.3.5.
- Correlations computed across groups (districts, states) need not hold for individuals: the ecological fallacy (recall §1.5).

11.3 Spearman Rank Correlation

DEFINITION

Spearman's rank correlation coefficient (ρ_s, or r_s) is Pearson's r computed on the *ranks* of the two variables instead of their raw values. With no tied ranks it reduces to

$$\rho_s = 1 - 6\Sigma d^2 \div [n(n^2 - 1)]$$

where d is the difference between a person's two ranks. **Kendall's tau** (τ) counts pairs of people instead:

$$\tau = (C - D) \div [n(n - 1)/2]$$

where C is the number of **concordant** pairs (the person higher on X is also higher on Y) and D the number of **discordant** pairs (Kendall, 1938). Both measure a *monotonic* relationship and lie between −1 and +1.

SIMPLIFIED EXPLANATION

Recall that the Mann–Whitney test (§10.2) replaces values by ranks. Spearman's coefficient does the same for correlation. Ranks ignore how far apart the values are, so an outlier can do no more than take the top rank, and a curved but steadily rising relationship ranks perfectly.

Kendall's τ asks a plainer question. Take every possible pair of people. Does the taller of the two also weigh more? τ and ρ_s are on different scales, so a τ from one study cannot be set beside a ρ_s from another. τ has a direct reading: the proportion of agreeing pairs minus the proportion of disagreeing pairs.

ILLUSTRATION / EXAMPLE

The Kuppuswamy socioeconomic scale and a health literacy score are both ordinal (recall §1.1.2). Their intervals are not equal, so Pearson's r is not appropriate. Rank both and use ρ_s or τ.

WORKED CALCULATION

The same eight children as in §11.2:

Child	1	2	3	4	5	6	7	8
Height rank	3	2	7	1	6	8	4	5
Weight rank	4	1	8	2	6	7	3	5
d	−1	1	−1	−1	0	1	1	0
d^2	1	1	1	1	0	1	1	0

1. $\Sigma d^2 = 6$; $n = 8$, so $n(n^2 - 1) = 8 \times 63 = 504$.
2. $\rho_s = 1 - (6 \times 6) \div 504 = 1 - 0.071 =$ **0.93**.
3. Kendall: there are $8 \times 7 \div 2 = 28$ pairs of children. Checking each pair, 25 are concordant and 3 discordant (children 1 and 7, 2 and 4, 3 and 6: in each pair the taller child is the lighter one).
4. $\tau = (25 - 3) \div 28 =$ **0.79**.

For all 100 children, age against weight gives ρ_s = **0.83** and τ = **0.66**, close to Pearson's 0.82 because the relationship is close to linear and has no extreme outliers.

IN R (OPTIONAL)

```
cl ← read.csv("data/clinic_children.csv")
cor.test(cl$age_months, cl$weight_kg, method = "spearman",
         exact = FALSE)
cor.test(cl$age_months, cl$weight_kg, method = "kendall",
         exact = FALSE)
```

```
    Spearman's rank correlation rho
...
S = 28518, p-value < 2.2e-16
sample estimates:
     rho
0.828872

    Kendall's rank correlation tau
...
z = 9.3613, p-value < 2.2e-16
sample estimates:
      tau
0.6550698
```

ρ_s = 0.83 and τ = 0.66. Ages are recorded in whole months, so many children share a rank; `exact = FALSE` asks for the large-sample p-values, which handle ties. R's τ with ties is the tie-corrected version (tau-b).

MUST-KNOW NOTES (MD-LEVEL)

- Use a rank coefficient when at least one variable is ordinal, when the relationship is monotonic but curved, or when outliers make r unreliable (Schober et al., 2018).
- The d^2 formula is exact only without ties. With ties, compute Pearson's r on the ranks (software such as R does this).
- When the data are Normal and linear, rank coefficients have somewhat less power than r, the same trade-off as between parametric and non-parametric tests (recall §10.1).

11.4 Point-Biserial and Phi Coefficient

DEFINITION

The **point-biserial coefficient** r_pb is Pearson's r between a continuous variable and a binary variable coded 0/1. The **phi coefficient** ϕ is Pearson's r between two binary variables coded 0/1; from a 2 × 2 table, $\phi = \sqrt{(\chi^2 \div N)}$ in size.

SIMPLIFIED EXPLANATION

Neither needs a new formula: code the binary variable 0/1 and compute r. The names survive from the days of hand calculation.

ILLUSTRATION / EXAMPLE

In the CBHI survey (500 households, 150 members), members live on average 6.41 km from the facility and non-members 8.12 km; the SD of distance is 3.25 km. Coding membership 1/0:

$r_{pb} = [(\bar{y}_1 - \bar{y}_0) \div s_y] \times \sqrt{[p(1-p)]} \times \sqrt{[n/(n-1)]} = (-1.71 \div 3.25) \times \sqrt{(0.30 \times 0.70)} \times 1.001 = \mathbf{-0.24}$

Its $t = r\sqrt{(n-2)}/\sqrt{(1-r^2)} = 5.54$ on 498 df is identical in size to the pooled two-sample t-test comparing mean distance between members and non-members (recall §9.2.2): the two analyses are one.

For phi, take the breastfeeding cohort of §10.6.4 (200 infants; $\chi^2 = 23.37$): $\phi = \sqrt{(23.37 \div 200)} = \mathbf{0.34}$.

MUST-KNOW NOTES (MD-LEVEL)

- The point-biserial t-test and the two-sample t-test give the same p-value; report whichever the reader will understand (usually the difference in means).
- ϕ can reach ±1 only when the two variables have the same marginal proportions, so it is a poor guide to strength. For a 2 × 2 table the risk ratio or odds ratio (recall §10.6.4) is usually the more useful summary.

CHECKPOINT 11.1

1. For 10 women, $S_{xx} = 250$, $S_{yy} = 40$ and $S_{xy} = 80$ (haemoglobin in g/dL against daily iron tablets taken over a month). Compute the covariance, r and the t statistic for H_0: $\rho = 0$. Is r significant at the 5% level?
2. A study of 28 anganwadi workers reports r = 0.50 between years of service and a knowledge score. Compute the 95% confidence interval for ρ by Fisher's z.
3. A researcher correlates ranked examination scores with hours of study. Which coefficient(s) should she use, and why?

(Answers in the Appendix at the end of the book.)

11.5 Correlation vs Causation

DEFINITION

An association between two variables, however strong or statistically significant, does not by itself establish that one causes the other. Four alternative explanations must be weighed first: **chance**, **bias**, **confounding** and **reverse causation**.

SIMPLIFIED EXPLANATION

A strong r, or a significant chi-square test of independence (recall §10.6.1), says that two variables move together. It says nothing about *why*.

1. **Chance.** Test enough pairs of variables and some will correlate by accident (recall §8.3 and the multiplicity point in §8.5). The p-value and confidence interval address this explanation, and only this one.
2. **Bias.** A systematic error in how people were selected (selection bias) or how the variables were measured (information bias) can create or hide an association. Bias is taken up in Chapter 17 (§17.2).
3. **Confounding.** A third variable causes both (§11.6).
4. **Reverse causation.** The outcome causes the "exposure". A cross-sectional survey that finds lower physical activity among people with arthritis cannot say which came first.

ILLUSTRATION / EXAMPLE

Ice cream sales and drowning deaths rise and fall together across the months of the year. Ice cream does not cause drowning: hot weather drives both. In a district, the number of health workers correlates positively with the recorded disease burden. More health workers do not cause disease. Districts with more disease may be sent more staff (reverse causation), bigger districts have more of both (confounding by population size), and more staff find and record more cases (information bias).

MUST-KNOW NOTES (MD-LEVEL)

- Bradford Hill (1965) listed nine "viewpoints" from which to study an association "before we cry causation": strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment and analogy. He wrote that none of them "can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a *sine qua non*". Temporality is the one that follows from logic: a cause must come before its effect. §17.2.3 treats these viewpoints in full.
- Hill also noted that significance tests "remind us of the effects that the play of chance can create" but "contribute nothing to the 'proof' of our hypothesis" (Hill, 1965).
- A randomised trial answers the confounding question by design: randomisation makes the groups alike, on average, in measured and unmeasured characteristics (§15.9, randomised trials, develops this).
- The same caution applies to every measure of association in this book: chi-square tests, risk and odds ratios, and the regression coefficients of Chapters 12–14.

11.6 Confounding – A Working Introduction

DEFINITION

A **confounder** is a variable that is associated with the exposure, is a cause of the outcome (or a proxy for one) through a route that does not pass through the exposure, and is not itself a consequence of the exposure. Its presence distorts the observed exposure–outcome association.

SIMPLIFIED EXPLANATION

Return to ice cream and drowning. The arithmetic is right; the comparison is wrong. Comparing high-ice-cream days with low-ice-cream days is secretly comparing hot days with cold days. A comparison is clean only if the groups are alike in every other respect that affects the outcome. When they are not, the difference you measure mixes the effect you want with the effects of everything else that differed. That mixing is confounding.

Three conditions make a variable a confounder:

1. It is associated with the exposure.
2. It affects the outcome other than through the exposure.
3. It is not on the causal path from the exposure to the outcome.

The third condition separates a confounder from two other kinds of third variable.

A **mediator** lies on the causal path: it is part of *how* the exposure acts. If maternal education improves child nutrition partly through better hygiene, hygiene is a mediator. Adjusting for it removes part of the effect you want to measure. This error is **overadjustment**.

A **collider** is *caused by* two variables on a path, for example by both the exposure and the outcome. Adjusting for a collider, or selecting study participants on it, creates an association that was not there. Variables measured after the exposure and outcome, or plausibly caused by them, are not automatically safe to adjust for.

Which variables belong in an analysis is therefore a question about how the variables are related in the world. The data cannot tell a cause from an effect. That judgement comes from subject knowledge and must be made before the analysis, not discovered during it.

ILLUSTRATION / EXAMPLE

A survey finds that households owning a radio have better-nourished children. Three candidate third variables:

- **Household wealth.** Wealthier households are more likely to own a radio (condition 1), feed their children better for reasons that have nothing to do with the radio (condition 2), and do not become wealthy because they own a radio (condition 3). Wealth is a **confounder**. Ignoring it overstates the benefit of the radio.
- **Exposure to nutrition messages.** If the radio delivers feeding advice that improves nutrition, the messages lie on the path from radio to nutrition. They are a **mediator**. Adjusting for them would hide a real effect.
- **Enrolment in a nutrition programme.** If radio-owning households hear about the programme and better-nourished children's families are more likely to stay enrolled, enrolment is caused by both. It is a **collider**. Studying only enrolled households would distort the radio–nutrition association.

Three variables, each correlated with both radio ownership and nutrition, need three different decisions.

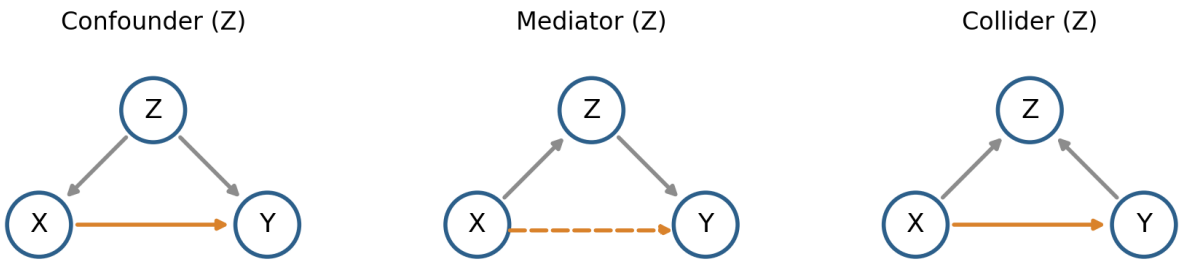


Figure 11.3. Confounder, mediator and collider as causal diagrams. X is the exposure, Y the outcome and Z the third variable.

MUST-KNOW NOTES (MD-LEVEL)

- **Confounder: adjust for it. Mediator: do not, if you want the total effect of the exposure. Collider: do not.** Correlations alone cannot tell you which is which.
- Confounding reflects how variables are related in the population, not an error in how the study was run. That is why a *measured* confounder can be dealt with at the analysis stage: by stratification (§11.7), by standardisation (recall §7.4.2) or by regression (§12.4.1, multiple regression).
- **Unmeasured confounding** comes from a confounder that was never recorded. No analysis of the data at hand can remove it.
- **Residual confounding** is what remains *after* adjusting for a measured confounder, because it was measured with error or grouped too coarsely (for example, wealth recorded as just "poor/not poor") (Fewell et al., 2007; Tripepi et al., 2010). The two are different problems: residual confounding can shrink with better measurement or finer strata; unmeasured confounding cannot shrink without new data.
- Chapter 17 (§17.3) returns to confounding as a threat to validity and to what study design can do about it.

11.6.1 Causal Diagrams (DAGs)

DEFINITION

A **causal diagram**, or **directed acyclic graph (DAG)**, draws each variable as a node and each assumed direct cause as an arrow from cause to effect. "Directed" means every arrow has one direction; "acyclic" means you can never follow arrows back to where you started.

- A **path** is any route between two nodes along arrows, whatever their direction, that passes through no node more than once.
- A **causal path** from exposure E to outcome W follows arrows forwards all the way (E → ... → W).
- A **backdoor path** is a path from E to W that starts with an arrow *into* E (E ← ...). An open backdoor path carries confounding.
- On a path, a node where two arrowheads meet (→ C ←) is a **collider on that path**.
- A **descendant** of E is any node you can reach from E by following arrows forwards (in Figure 11.4 below: A, P, H and W).

- A path is **open** (transmits association) unless it is **blocked**. It is blocked if it contains a non-collider that you adjust for, or a collider that you do *not* adjust for (and none of whose descendants you adjust for).

The **backdoor criterion**: a set of variables Z is sufficient to remove confounding of the effect of E on W if (1) no variable in Z is a descendant of E, and (2) Z blocks every backdoor path from E to W (Greenland, Pearl & Robins, 1999). A **minimal adjustment set** is a sufficient set from which no variable can be removed.

SIMPLIFIED EXPLANATION

Association flows along open paths like water through pipes. The causal paths carry the effect you want. The backdoor paths carry confounding. Adjusting for a variable in the middle of a pipe closes it. A collider is a pipe that is closed until you adjust for it; adjusting opens it. So the job is: keep the causal paths untouched, close every backdoor path, and open no collider. The diagram makes the job mechanical once you have drawn it; the hard part is drawing it honestly from subject knowledge.

ILLUSTRATION / EXAMPLE

Does maternal education (E) raise birthweight (W)? Figure 11.4 records the assumptions of a hypothetical study team.

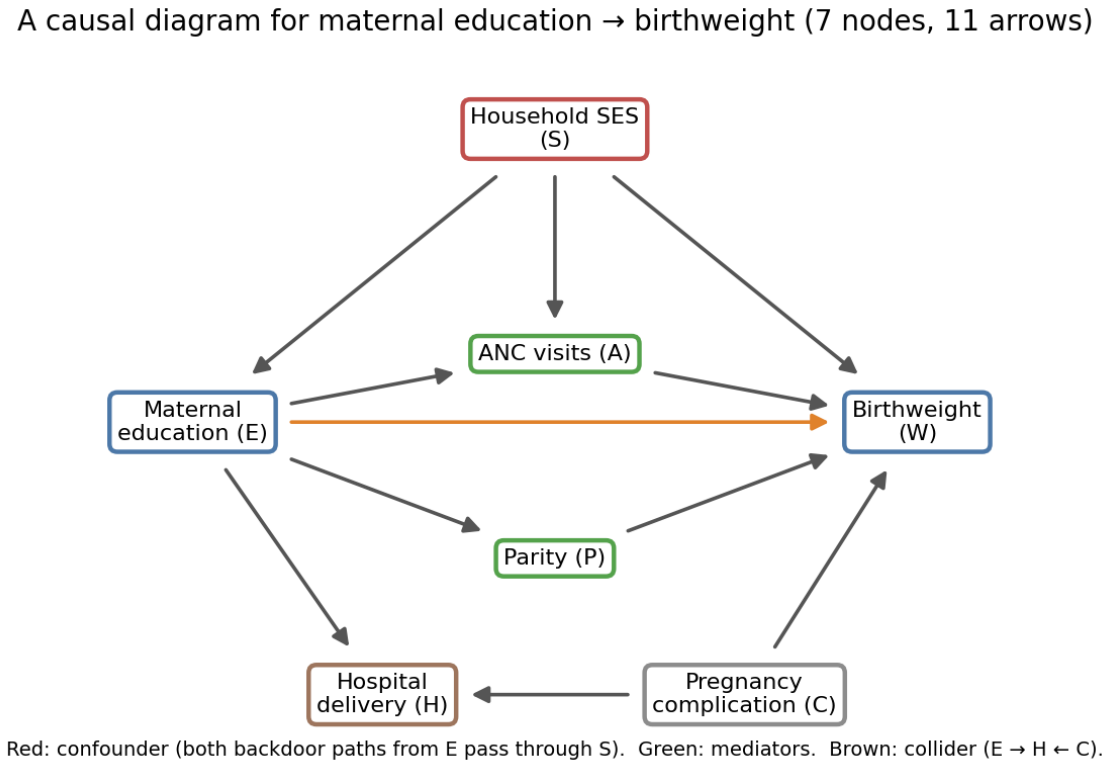


Figure 11.4. Assumed causal structure. Household socioeconomic status (S) affects education, antenatal care (A) and birthweight. Education acts on birthweight directly and through ANC visits and parity (P). Hospital delivery (H) is affected by education and by pregnancy complications (C), which also lower birthweight.

WORKED CALCULATION

1. List every path from E to W (seven in all):
   - E → W; E → A → W; E → P → W (causal paths)
   - E ← S → W; E ← S → A → W (backdoor paths)
   - E → A ← S → W; E → H ← C → W (non-causal paths that start forwards; each contains a collider, A or H)
2. With no adjustment, the three causal paths and the two backdoor paths are open. The two paths through a collider are blocked.
3. Both backdoor paths pass through S. Adjusting for S blocks both. S is not a descendant of E. So {S} satisfies the backdoor criterion, and nothing can be removed from it: {S} is the **minimal adjustment set** for the total effect of education.
4. Adjusting for A (ANC) instead is wrong twice. A is a mediator, so the path E → A → W is cut and part of the effect is lost. A is also a collider on E → A ← S → W, so adjusting for it opens that path; with S left unadjusted, a non-causal association flows through it.
5. Restricting the study to hospital deliveries (conditioning on H) opens E → H ← C → W. Among hospital births, less-educated mothers are more likely to be there *because* of a complication, and complications lower birthweight. A spurious association appears even after adjusting for S. This is selection bias with a collider structure (Hernán et al., 2004).
6. Parity (P) is a mediator here: adjust for it only if you want the effect of education *other than* through family size.

The free browser tool DAGitty (dagitty.net), "a browser-based environment for creating, editing, and analyzing causal diagrams", lists the paths and minimal adjustment sets for larger diagrams.

MUST-KNOW NOTES (MD-LEVEL)

- A DAG is a statement of assumptions, not a finding. Two teams can draw different diagrams; the diagram makes the disagreement visible and checkable.
- The traditional three-condition test of §11.6 can mislead when there are several possible confounders; the backdoor criterion corrects it (Greenland, Pearl & Robins, 1999).
- Do not adjust for a descendant of the exposure when estimating its total effect (condition 1 of the backdoor criterion).
- An arrow you leave out is a strong claim: it says there is *no* direct effect.
- Hernán et al. (2004) show that many selection biases have the same structure: conditioning on a common effect (a collider). Chapter 17 (§17.2) uses this.

CHECKPOINT 11.2

1. In the district example of §11.5 (more health workers, more recorded disease), name one explanation of each of the four kinds (chance, bias, confounding, reverse causation).
2. In Figure 11.4, how many paths join E and W, and how many of them are backdoor paths? A colleague adjusts for ANC visits and analyses only hospital deliveries. Which non-causal paths does she open? What should she adjust for instead?

3. Wealth is recorded only as "above or below the poverty line". After adjusting for it, the radio–nutrition association shrinks but does not vanish. Is what remains unmeasured or residual confounding? What would reduce it?

(Answers in the Appendix at the end of the book.)

11.7 Stratified Analysis – Mantel–Haenszel, Effect Modification and Simpson's Paradox

DEFINITION

Stratified analysis splits the data into strata (layers) defined by a third variable, computes the exposure–outcome measure within each stratum, and, if the stratum measures agree, combines them into one **adjusted** estimate. The **Mantel–Haenszel (MH)** estimators (Mantel & Haenszel, 1959) are weighted averages of the stratum-specific measures. For stratum i with the 2 × 2 layout of §10.6.4 (a_i = exposed with outcome, b_i = exposed without, c_i = unexposed with, d_i = unexposed without, n_i = stratum total):

$$OR_{MH} = \Sigma(a_i d_i / n_i) \div \Sigma(b_i c_i / n_i)$$

$$RR_{MH} = \Sigma[a_i(c_i + d_i)/n_i] \div \Sigma[c_i(a_i + b_i)/n_i]$$

Effect modification is present when the true exposure effect differs between strata. **Simpson's paradox** is the extreme case of confounding in which the crude association points the opposite way to the association in every stratum.

SIMPLIFIED EXPLANATION

Stratification is the simplest way to compare like with like. If the third variable is a confounder, then within each of its strata it cannot confound, because it does not vary there. You compute the risk ratio separately in each stratum, check that the stratum answers agree, and pool them.

The stratified analysis answers two separate questions:

1. **Do the stratum-specific measures differ from each other?** If yes, beyond what chance would produce, the third variable is an **effect modifier**. Report the stratum-specific estimates; a single pooled number would hide the finding.
2. **If they agree, does the pooled (MH) estimate differ from the crude one?** If yes, the third variable was **confounding** the crude estimate. Report the MH estimate.

The two questions are distinct, not opposite. A variable can be a confounder, an effect modifier, both, or neither. Confounding is a distortion of the crude estimate, and you remove it. Effect modification is a real feature of the effect, and you report it.

DERIVATION

Why the MH OR is an average of the stratum ORs. Write $OR_i = a_i d_i/(b_i c_i)$ and choose weights $w_i = b_i c_i / n_i$. Then:

1. $w_i \times OR_i = (b_i c_i / n_i) \times a_i d_i/(b_i c_i) = a_i d_i / n_i$.

2. Weighted average = $\Sigma w_i\, OR_i \div \Sigma w_i = \Sigma(a_i d_i/n_i) \div \Sigma(b_i c_i/n_i) = OR_MH$.
3. The weights are larger for bigger, more balanced strata, and a stratum with a zero cell still contributes, which a simple average of the ORs cannot do.

The same algebra with $w_i = c_i(a_i + b_i)/n_i$ gives RR_MH as a weighted average of the stratum RRs.

The MH chi-square test. Under H_0 (no association in any stratum), and with the margins of each table fixed, a_i has expected value $E(a_i) = (a_i + b_i)(a_i + c_i)/n_i$ and variance $V(a_i) = (a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i) \div [n_i^2(n_i - 1)]$. Summing over strata:

$\chi^2_MH = [\Sigma a_i - \Sigma E(a_i)]^2 \div \Sigma V(a_i)$, on 1 degree of freedom.

It asks whether the exposed have more outcomes than expected, pooling the evidence from all strata.

A test of homogeneity. Woolf's method (Woolf, 1955) weights each stratum's log RR (or log OR) by the inverse of its variance, $w_i = 1/SE^2(\ln RR_i)$, with $SE^2(\ln RR) = 1/a - 1/(a + b) + 1/c - 1/(c + d)$ (recall §10.6.4). Then

$Q = \Sigma w_i(\ln RR_i - \ln RR_w)^2$, where $\ln RR_w = \Sigma w_i \ln RR_i \div \Sigma w_i$,

is compared with χ^2 on (number of strata − 1) df. The Breslow–Day test (Breslow & Day, 1980) is an alternative for odds ratios. Both can have low power when strata are small, so a non-significant result is weak evidence that the effects are the same.

ILLUSTRATION / EXAMPLE

Recall the cohort of 200 infants in §8.7 and §10.6.4: not exclusively breastfed (not EBF) versus EBF, followed for diarrhoea. Crude RR = 3.54 (95% CI 2.02 to 6.17); crude OR = 5.32. The infants came from two PHCs:

Stratum	Exposure	Diarrhoea	No diarrhoea	Total	Risk
PHC Urban	Not EBF	9	21	30	0.300
	EBF	9	81	90	0.100
PHC Rural	Not EBF	24	26	50	0.480
	EBF	5	25	30	0.167
Both (crude)	Not EBF	33	47	80	0.413
	EBF	14	106	120	0.117

WORKED CALCULATION

1. Stratum RRs: urban 0.300 ÷ 0.100 = **3.00**; rural 0.480 ÷ 0.167 = **2.88**. Stratum ORs: urban (9 × 81)/(21 × 9) = 3.86; rural (24 × 25)/(26 × 5) = 4.62.
2. Homogeneity (Woolf, RR). Urban: $SE^2 = 1/9 - 1/30 + 1/9 - 1/90 = 0.178$, w = 5.63, ln RR = 1.099. Rural: $SE^2 = 1/24 - 1/50 + 1/5 - 1/30 = 0.188$, w = 5.31, ln RR = 1.058. Pooled ln RR_w = 1.079. $Q = 5.63(0.020)^2 + 5.31(0.021)^2 =$ **0.005** on 1 df, p = **0.95**. No evidence of effect modification on the RR scale. (For the ORs, the Breslow–Day test gives $\chi^2 = 0.054$, p = 0.82.)
3. MH RR: numerator 9 × 90/120 + 24 × 30/80 = 6.75 + 9.00 = 15.75; denominator 9 × 30/120 + 5 × 50/80 = 2.25 + 3.125 = 5.375. RR_MH = 15.75 ÷ 5.375 = **2.93**. Its 95% CI, 1.60 to 5.35, uses the

Mantel–Haenszel variance formula for the RR, which is long by hand and usually left to software.

4. MH OR: numerator 9 × 81/120 + 24 × 25/80 = 6.075 + 7.500 = 13.575; denominator 21 × 9/120 + 26 × 5/80 = 1.575 + 1.625 = 3.200. OR_MH = 13.575 ÷ 3.200 = **4.24**.
5. MH χ^2: E(a) = 30 × 18/120 = 4.50 (urban) and 50 × 29/80 = 18.125 (rural), total 22.625, against 33 observed. V = 2.893 + 4.388 = 7.281. χ^2_MH = $(33 - 22.625)^2 \div 7.281 = 107.64 \div 7.281 =$ **14.78**, p = 0.0001.
6. Crude versus adjusted: 3.54 against 2.93, a ratio of 1.21. PHC was confounding: not-EBF infants were concentrated in the rural PHC (50 of 80 rural infants, 62.5%, against 30 of 120 urban, 25%), and diarrhoea was commoner there even among EBF infants (16.7% against 10.0%). Part of the crude excess was the rural PHC, not feeding.

Conclusion: after allowing for PHC, not being exclusively breastfed is associated with about three times the risk of diarrhoea (RR_MH 2.93, 95% CI 1.60 to 5.35), and the association is similar in both PHCs.

Whether a crude-versus-adjusted gap of this size "matters" is a judgement. Comparing the change in estimate against a pre-stated size is one way to decide which confounders to keep; in simulations it tended to do better than significance testing when the right decision was not obvious (Mickey & Greenland, 1989). The decision to treat PHC as a potential confounder, however, comes from the causal reasoning of §11.6.1, not from the size of the change.

Simpson's paradox with numbers. An illustrative example (hypothetical counts). Two PHCs treat children admitted with diarrhoea; the outcome is treatment failure (referral onwards). PHC A has an inpatient ward and receives most of the severely dehydrated children.

Severity	PHC	Failures	Children	Risk
Severe	A	12	100	0.120
	B	3	20	0.150
Not severe	A	1	50	0.020
	B	6	200	0.030
Both (crude)	A	13	150	0.087
	B	9	220	0.041

1. Within each severity stratum PHC A does *better*: RR (A vs B) = 0.12 ÷ 0.15 = 0.80 for severe, 0.02 ÷ 0.03 = 0.67 for non-severe.
2. Crude, PHC A looks *worse*: RR = 0.087 ÷ 0.041 = 2.12.
3. The reason: two-thirds of A's children (100 of 150) were severe, against one-eleventh of B's (20 of 220). Severity is a cause of failure and is associated with the PHC. It confounds the crude comparison strongly enough to reverse it.
4. RR_MH = (12 × 20/120 + 1 × 200/250) ÷ (3 × 100/120 + 6 × 50/250) = (2.0 + 0.8) ÷ (2.5 + 1.2) = 2.8 ÷ 3.7 = **0.76**. Adjusted for severity, A's failure risk is about three-quarters of B's.

The same reversal appeared with district death rates in §7.4.2, where age played the part of severity.

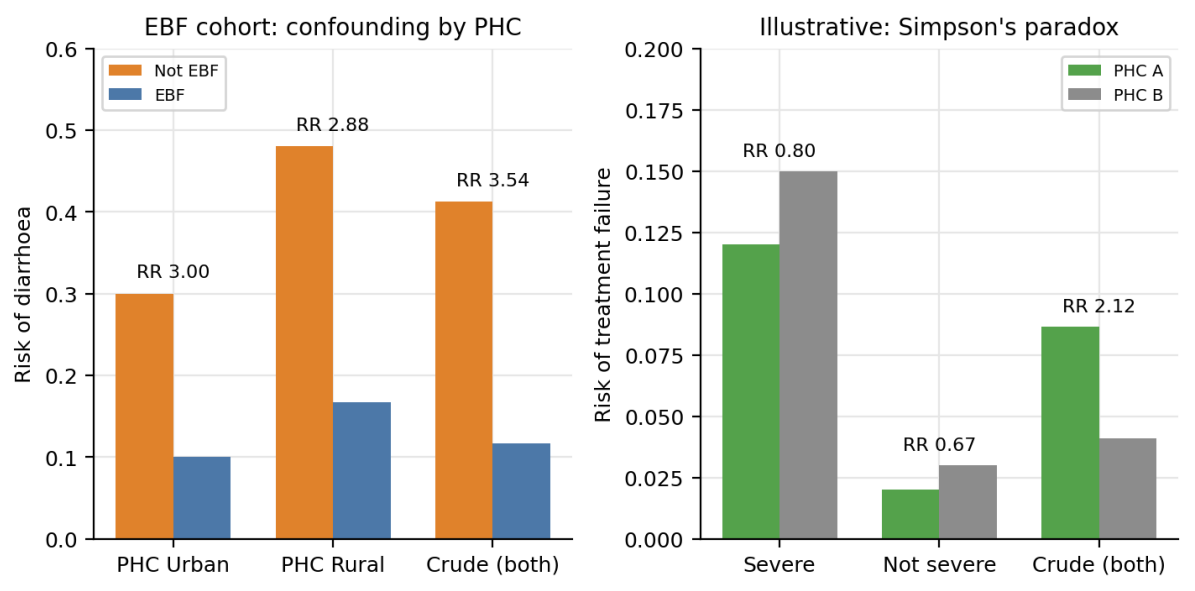


Figure 11.5. Left: diarrhoea risk by feeding within each PHC and crude (canonical cohort); the stratum RRs (3.00, 2.88) agree with each other and are lower than the crude 3.54: confounding. Right: the illustrative Simpson's paradox table; PHC A has the lower risk in each severity stratum but the higher crude risk.

IN R (OPTIONAL)

```
eb ← read.csv("data/ebf_diarrhoea_cohort.csv")
tab ← table(exposure  = factor(eb$not_ebf, levels = c(1, 0),
                               labels = c("Not EBF", "EBF")),
            diarrhoea = factor(eb$diarrhoea, levels = c(1, 0),
                               labels = c("Yes", "No")),
            phc       = eb$phc)
mantelhaen.test(tab, correct = FALSE)
library(epiR)
epi.2by2(tab, method = "cohort.count")
```

```
    Mantel-Haenszel chi-squared test without continuity correction

data:  tab
Mantel-Haenszel X-squared = 14.784, df = 1, p-value = 0.0001205
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.981805 9.080689
sample estimates:
common odds ratio
         4.242187
...
Point estimates and 95% CIs:
-------------------------------------------------------------------
Inc risk ratio (crude)                          3.54 (2.02, 6.17)
Inc risk ratio (M-H)                            2.93 (1.60, 5.35)
Inc risk ratio (crude:M-H)                      1.21
Inc odds ratio (crude)                          5.32 (2.61, 10.85)
Inc odds ratio (M-H)                            4.24 (1.98, 9.08)
Inc odds ratio (crude:M-H)                      1.25
...
-------------------------------------------------------------------
 M-H test of homogeneity of IRRs: chi2(1) = 0.005 Pr>chi2 = 0.946
 M-H test of homogeneity of ORs: chi2(1) = 0.027 Pr>chi2 = 0.871
 Test that M-H adjusted OR = 1:  chi2(1) = 14.784 Pr>chi2 = <0.001
```

The table is built with "Not EBF" and "Yes" first so that cell a is exposed-with-outcome, as in §10.6.4. `mantelhaen.test` gives χ^2_MH = 14.78 (step 5) and OR_MH = 4.24 (step 4) with its CI. `epi.2by2` adds the crude and MH risk ratios (3.54 and 2.93, step 3), their ratio 1.21 (step 6), and the homogeneity test of the RRs (χ^2 = 0.005, p = 0.95, step 2). By default `mantelhaen.test` applies a continuity correction; `correct = FALSE` matches the hand calculation. The OR homogeneity line (χ^2 = 0.027) is not the Breslow–Day test of step 2 (0.054): `epi.2by2` uses Woolf's method on the log ORs after adding 0.5 to every cell. Both find no evidence that the ORs differ.

MUST-KNOW NOTES (MD-LEVEL)

- **Decision rule.** First test homogeneity. If the stratum estimates clearly differ, report them separately (effect modification). If they agree, compare crude and MH: a material difference means confounding, so report the MH estimate; little difference means the variable did not confound, and the crude estimate may be reported.
- Interaction and confounding are distinct, not opposite. PHC in the cohort confounded (3.54 vs 2.93) without modifying (3.00 vs 2.88).
- **Effect modification depends on the scale.** The same cohort shows almost equal RRs (3.00, 2.88) but different risk differences (0.20 urban, 0.31 rural). Say which measure you mean, and give the stratum-specific RRs, ORs or RDs so readers can judge (Knol & VanderWeele, 2012).
- Stratification works for one or two categorical confounders. With several, strata multiply and empty out; a continuous confounder must be cut into categories, which leaves residual confounding within the categories (Tripepi et al., 2010). Regression (Chapters 12–13) adjusts for many variables at once.
- With matched pairs as strata, the MH OR equals the matched-pair OR b/c (recall §10.8): for the 80 pairs in `matched_casecontrol.csv` both give 2.4.

- The MH OR is valid for case-control data (it uses only odds); the MH RR needs cohort or cross-sectional data, where risks can be estimated (recall §10.6.4).

CHECKPOINT 11.3

1. Two strata of a cohort give (a, b, c, d) = (10, 20, 5, 40) and (15, 10, 20, 30), with 75 people in each. Compute the stratum ORs, the crude OR and OR_MH. Did the stratifying variable confound the crude OR, and in which direction?
2. In a study of tobacco chewing and oral leukoplakia, the RR is 1.2 among non-drinkers and 4.5 among drinkers; the homogeneity test gives p = 0.002. What should the authors report, and why is an MH RR inappropriate?
3. In the illustrative Simpson's paradox table of §11.7, apply PHC A's severity-specific failure risks to PHC B's mix of children (20 severe, 200 not severe). What overall failure risk would A have had? Compare it with B's crude risk and explain the paradox in one sentence.

(Answers in the Appendix at the end of the book.)

Regression – Foundations

Correlation gave one number, r = 0.82, for how strongly age and weight move together in the under-five clinic register (recall §11.2). Regression turns that into an equation: how much weight changes, on average, per month of age, and what weight to expect for a given child. This chapter builds simple linear regression from the least-squares principle, then extends it to several predictors, to curved relationships, and to the diagnostic checks that tell you whether a fitted model can be trusted. Regression inherits every caution of §11.5: a coefficient describes an association in the data, and calling it an effect needs the same extra evidence as any correlation. Chapter 13 (generalised linear models, regression for binary and count outcomes) extends the same framework.

12.1 From Correlation to Prediction

DEFINITION

Regression fits an equation that gives the expected value of an outcome variable for given values of one or more predictor variables. Correlation measures how strongly two variables are linearly related; regression describes *what* the relationship is, in the units of the variables.

SIMPLIFIED EXPLANATION

A correlation coefficient answers "how strongly are these two variables related, and in which direction?" It does not answer "given this value of the predictor, what value of the outcome should I expect?" Regression answers that second question by fitting a line (or curve) through the data.

ILLUSTRATION / EXAMPLE

Knowing that age and weight correlate at r = 0.82 tells a clinician the relationship is strong and positive, but not what weight to expect for an 18-month-old. The regression equation for the clinic register, predicted weight = 8.5 + 0.15 × age (months), gives 11.2 kg, and §12.2.4 adds how widely individual children scatter around that prediction.

Regression to the mean. The word "regression" itself comes from a pattern every screening programme meets. Suppose adults are screened for hypertension, and everyone whose first systolic reading is 160 mmHg or more is called back for a second reading. Blood pressure varies from day to day, so a very high first reading is partly a truly high pressure and partly a high day. At the second visit the high day is not repeated on average, and the group's mean falls, with no treatment at all.

Worked Calculation (a simulation; the values are illustrative, not survey data)

1. Assume true systolic pressure varies between adults with mean 130 and SD 15 mmHg, and each reading adds day-to-day variation with SD 10 mmHg.
2. Among 100,000 simulated adults, 4,825 have a first reading ≥ 160. Their mean first reading is 167.7 mmHg.
3. The same 4,825 adults, measured again, average 156.0 mmHg: 11.7 mmHg lower, with nothing done to them.
4. Why: the two readings share the true pressure but not the day-to-day variation, so their covariance is the true-pressure variance, $15^2 = 225$, while each reading has variance $15^2 + 10^2 = 325$ (variances of independent parts add, recall §6.5). Their correlation is $225 \div 325 = 0.692$. §12.2.2 shows that a regression prediction lies only r times as many SDs from the mean as the predictor does. A reading 30 mmHg above the mean predicts a second reading $0.692 \times 30 = 20.8$ mmHg above it: 150.8 mmHg, not 160.

MUST-KNOW NOTES (MD-LEVEL)

- Correlation is symmetric: r between X and Y equals r between Y and X. Regression is directional. Regressing weight on age gives a different line from regressing age on weight, because each minimises errors in a different variable.
- The weaker the correlation, the more individual points scatter around the fitted line, and the less precise predictions for individuals become. The equation can still be calculated.
- **Regression to the mean** occurs whenever people are selected because of an extreme first measurement and then measured again. It makes natural variation look like real change, and it is larger when measurement error is larger (Barnett, van der Pols & Dobson, 2005). A before–after study of "patients with high BP at screening" will show a fall even if the intervention does nothing. The defences are a randomised control group, which regresses equally, and basing selection on the mean of two or more readings.

12.2 Simple Linear Regression

DEFINITION

Simple linear regression fits a straight line between one continuous predictor x and one continuous outcome y: $\hat{y} = b_0 + b_1 x$, where $\hat{y}$ ("y-hat") is the predicted outcome, b_0 the **intercept** and b_1 the **slope**.

SIMPLIFIED EXPLANATION

This is the simplest regression model: the "best" straight line through a scatter diagram (recall §4.4), where "best" has the exact meaning given in §12.2.2. The equation has just two numbers to interpret: where the line crosses the y-axis, and how steeply it rises.

ILLUSTRATION / EXAMPLE

For the 100 children aged 12 to 36 months in the clinic register, the fitted line is predicted weight (kg) = 8.5 + 0.15 × age (months). Each extra month of age adds 0.15 kg to the predicted weight. At age 0 the equa-

tion gives 8.5 kg, but no child younger than 12 months was measured, so that value is an extrapolation with no biological meaning (§12.2.3). The figure therefore draws the line only over the observed ages.

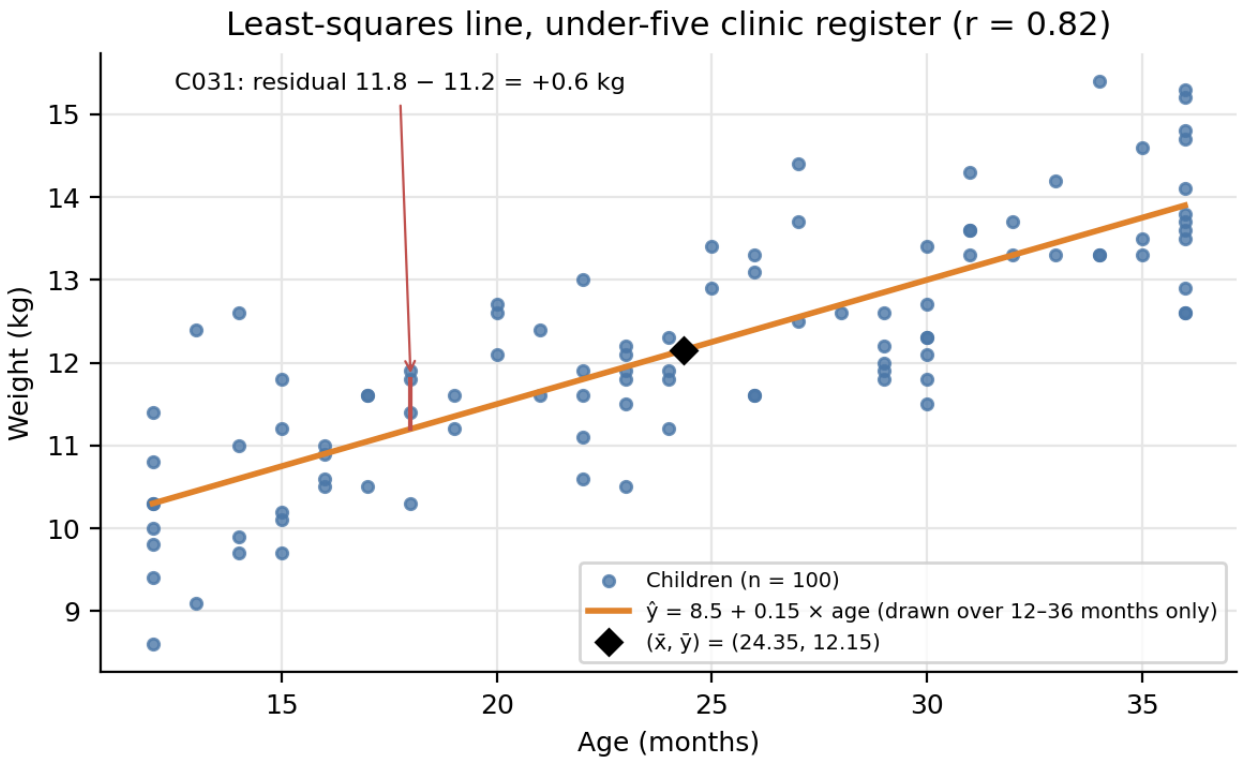


Figure 12.1. The least-squares line for weight on age in the clinic register, drawn over the observed ages (12–36 months) only. The line passes through the mean point ($\bar{x}$, $\bar{y}$); the red segment is child C031's residual (§12.2.1).

MUST-KNOW NOTES (MD-LEVEL)

- Look at the scatter diagram before fitting (recall §4.4). A straight line fitted to a curved relationship over-predicts in some parts of the range and under-predicts in others (§12.4.5 shows how to model curves).

12.2.1 The Regression Model – Population, Sample, and Error

DEFINITION

A regression equation is a **model** of how the outcome arises in the population: $Y = \beta_0 + \beta_1 X + \varepsilon$, where β_0 and β_1 are fixed, unknown population **parameters** and ε is a random **error term** holding everything that affects the outcome other than the predictor. The line fitted to a sample, $\hat{y} = b_0 + b_1 x$, gives **estimates** b_0 and b_1 of those parameters.

SIMPLIFIED EXPLANATION

For a straight line to describe children's weights perfectly, every child of a given age would have to weigh exactly the same. They don't: genetics, nutrition, illness and ordinary individual variation all matter. So the honest model is not "weight equals a line" but "weight equals a line plus everything else". That "everything else" is ε, and it is an expected part of the model, not a defect in it.

This is the parameter-versus-statistic distinction of Chapter 1 again (recall §1.3). There is a true intercept β_0 and a true slope β_1 for the whole population, fixed and unobservable. From one sample you calculate b_0 and b_1. Greek letters for what is true and unknown; Roman letters for what you calculated. Another sample of children would give a slightly different b_1, while β_1 never moves. Because b_1 is an *estimate of* β_1, it carries uncertainty, which §12.2.4 quantifies.

Two terms follow:

- The **fitted value** $\hat{y}$ is the model's prediction for an observation: the height of the line at that observation's x.
- The **residual** is observed y − fitted $\hat{y}$. It is the calculable stand-in for the unobservable ε.

ε belongs to the population model and can never be observed. A residual belongs to your sample and can be calculated for every data point, which is why residuals are the main diagnostic tool (§12.5).

ILLUSTRATION / EXAMPLE

Child C031 in the register is 18 months old and weighs 11.8 kg. The fitted model is $\hat{y} = 8.5 + 0.15x$.

- **Fitted value:** $\hat{y} = 8.5 + (0.15 \times 18) =$ **11.2 kg**, the model's prediction for any 18-month-old.
- **Observed value:** 11.8 kg.
- **Residual:** 11.8 − 11.2 = **+0.6 kg**. This child is 0.6 kg heavier than the model expected.

That 0.6 kg is not a mistake. It is this child being an individual. The model never claimed every 18-month-old weighs 11.2 kg; it claimed 11.2 kg is the average around which 18-month-olds vary.

MUST-KNOW NOTES (MD-LEVEL)

- **Population model:** $Y = \beta_0 + \beta_1 X + \varepsilon$. **Fitted sample model:** $\hat{y} = b_0 + b_1 x$.
- β_0 and β_1 are parameters (fixed, unknown). b_0 and b_1 are statistics (calculated, varying from sample to sample).
- **Residual** = observed − fitted. Every assumption check in §12.3 and §12.5 is done on residuals, because they are the observable counterpart of ε.

12.2.2 Least Squares Principle

DEFINITION

The **least squares principle** chooses the line that minimises the **sum of squared residuals**, SSE = $\Sigma(y_i - b_0 - b_1 x_i)^2$, over all possible values of b_0 and b_1.

SIMPLIFIED EXPLANATION

Infinitely many lines could be drawn through a scatter of points. For any candidate line, measure each point's vertical distance from it, square each distance, and add them up. The least-squares line is the one line that makes this total as small as possible. Squaring does what it did for the variance (recall §3.6): a point above the line and one equally far below both count, instead of cancelling, and large misses are penalised more than small ones.

Two sums do all the work. For n pairs (x_i, y_i) with means $\bar{x}$ and $\bar{y}$:

$Sxx = \Sigma(x_i - \bar{x})^2$ $Sxy = \Sigma(x_i - \bar{x})(y_i - \bar{y})$

Sxx is the numerator of the variance of x, and Sxy is the numerator of the covariance of x and y (recall §11.2).

DERIVATION

1. SSE is a smooth bowl-shaped function of b_0 and b_1. At its lowest point its slope in each direction is zero.
2. Slope with respect to b_0: $-2\Sigma(y_i - b_0 - b_1x_i) = 0$. Divide by $-2n$: $\bar{y} - b_0 - b_1\bar{x} = 0$, so $\mathbf{b_0 = \bar{y} - b_1\bar{x}}$.
3. Substitute this b_0 into SSE: each residual becomes $(y_i - \bar{y}) - b_1(x_i - \bar{x})$.
4. Slope with respect to b_1: $-2\Sigma(x_i - \bar{x})[(y_i - \bar{y}) - b_1(x_i - \bar{x})] = 0$, so $Sxy - b_1Sxx = 0$, and $\mathbf{b_1 = Sxy \div Sxx}$.
5. Pearson's r is $Sxy \div \sqrt{(Sxx \cdot Syy)}$ (recall §11.2), and the SDs are $s_x = \sqrt{(Sxx/(n-1))}$ and $s_y = \sqrt{(Syy/(n-1))}$. Therefore $\mathbf{b_1 = r \times s_y \div s_x}$.

Three properties follow at once:

- **The line passes through the mean point.** Step 2 says $\bar{y} = b_0 + b_1\bar{x}$: at $x = \bar{x}$ the line predicts exactly $\bar{y}$.
- **The slope has the sign of r**, because $s_y \div s_x$ is always positive. $b_1 = 0$ exactly when $r = 0$.
- **Regression to the mean.** Rewrite the line as $\hat{y} - \bar{y} = r(s_y/s_x)(x - \bar{x})$, then divide by s_y: $(\hat{y} - \bar{y})/s_y = r \times (x - \bar{x})/s_x$. In SD units, the prediction is r times as far from the mean as the predictor. Unless $|r| = 1$, predictions are pulled towards the mean (§12.1).

Step 2 also shows that the residuals sum to zero: $\Sigma(y_i - b_0 - b_1x_i) = 0$.

Illustration / Example: Worked Calculation (clinic register, n = 100)

1. $\bar{x} = 24.35$ months, $\bar{y} = 12.15$ kg; $Sxx = 6190.8$ months^2, $Sxy = 928.6$ kg·months.
2. $b_1 = 928.6 \div 6190.8 =$ **0.150 kg per month**.
3. Check through r: $s_x = 7.91$ months, $s_y = 1.45$ kg, $r = 0.820$, so $r \times s_y \div s_x = 0.820 \times 1.446 \div 7.908 = 0.150$. Same answer.
4. $b_0 = 12.15 - 0.150 \times 24.35 = 12.15 - 3.65 =$ **8.50 kg**.
5. Mean point: $8.50 + 0.150 \times 24.35 = 12.15$ kg $= \bar{y}$.
6. In SD units: a 36-month-old is $(36 - 24.35) \div 7.91 = 1.47$ SD above the mean age. The predicted weight, $8.50 + 0.150 \times 36 = 13.90$ kg, is $(13.90 - 12.15) \div 1.446 = 1.21$ SD above the mean weight, which is 0.82×1.47.

MUST-KNOW NOTES (MD-LEVEL)

- $\mathbf{b_1 = Sxy/Sxx = r \cdot s_y/s_x}$ and $\mathbf{b_0 = \bar{y} - b_1\bar{x}}$. Every software package computes these (for several predictors, the same principle solved with matrix algebra, which is beyond this book).
- The fitted line always passes through $(\bar{x}, \bar{y})$. This is a quick check on a hand-computed line.
- Least squares is a rule for choosing the line. It needs no assumption about the distribution of ε. The assumptions of §12.3 are needed for the standard errors, CIs and p-values of §12.2.4.

CHECKPOINT 12.1

1. For the clinic register, Sxy = 928.6, Sxx = 6190.8, $\bar{x}$ = 24.35 months and $\bar{y}$ = 12.15 kg. Compute b_1 and b_0. Then find the fitted weight and the residual for a 30-month-old child who weighs 12.0 kg.
2. Children in a school are weighed, and the 20 thinnest are enrolled in a feeding programme. Three months later their mean BMI z-score has risen. Name one reason, other than the programme, why their mean would rise, and one design feature that would remove it.
3. Two variables have r = 0.5. A person is 2 SD above the mean of x. How many SDs above the mean of y is the predicted value?

(Answers in the Appendix at the end of the book.)

12.2.3 Interpreting Slope and Intercept

DEFINITION

The **slope** b_1 is the expected change in the outcome for a one-unit increase in the predictor. The **intercept** b_0 is the predicted outcome when the predictor equals zero.

SIMPLIFIED EXPLANATION

These two numbers carry the real-world meaning of the line. Interpreting them in the units of the actual variables is the most practical skill in reading regression output.

ILLUSTRATION / EXAMPLE

In "predicted weight (kg) = 8.5 + 0.15 × age (months)", the slope means that, on average across the sampled children, each additional month of age goes with 0.15 kg more predicted weight (1.8 kg per year). The intercept, 8.5 kg, is the predicted weight at age 0. No child younger than 12 months was measured, so this number only positions the line; it is not an estimate of birth weight.

MUST-KNOW NOTES (MD-LEVEL)

- A slope's units are (units of outcome) per (unit of predictor). "0.15" alone means nothing; "0.15 kg per month" does.
- The intercept is meaningful only when x = 0 lies within, or close to, the observed range. Extrapolating beyond the data assumes the relationship stays linear where it was never observed. To give the intercept meaning, centre the predictor: with age − 24 months as x, b_0 becomes the predicted weight of a 24-month-old (8.50 + 0.150 × 24 = 12.1 kg).
- The slope has the sign of Pearson's r and is zero exactly when r is zero (§12.2.2). The slope carries units; r does not.

12.2.4 Coefficient Standard Error, Confidence Interval and p-value

DEFINITION

Because b_1 estimates an unknown parameter, it has a sampling distribution. Its **standard error** $SE(b_1)$ is the standard deviation of that distribution. Its **95% confidence interval** is $b_1 \pm t \times SE(b_1)$. Its **p-value** tests H_0: $\beta_1 = 0$, the hypothesis that the predictor has no linear association with the outcome. All three rest on the **standard error of the estimate** (the residual standard error), $s = \sqrt{[SSE/(n-2)]}$, which estimates σ, the SD of ε.

SIMPLIFIED EXPLANATION

Every idea here was built earlier. It is four familiar steps applied to b_1.

Step one: the scatter around the line. Before asking how precise the slope is, ask how far points typically fall from the line. SSE is the total squared miss; divide by its degrees of freedom and take the square root: $s = \sqrt{[SSE/(n-2)]}$. The divisor is $n - 2$, not $n - 1$, because two parameters (b_0 and b_1) were estimated from the data before the residuals could be computed (recall the degrees-of-freedom logic of §3.6). s is in the outcome's units. With roughly Normal residuals, about 95% of individuals lie within about ±2s of the line.

Step two: b_1 varies, and its spread is $SE(b_1)$. Another sample of children would give a slightly different b_1. The SD of those b_1 values across repeated samples is its standard error (recall §6.7). It answers: if I repeated the study, how much would the slope move?

Step three: the confidence interval. As in Chapter 7, the 95% CI is estimate ± (critical value × SE): $b_1 \pm t \times SE(b_1)$, with t from the t-distribution on $n - k - 1$ degrees of freedom (n observations, k predictors, one more for the intercept). The interval is the range of slopes compatible with the data. The 95% belongs to the procedure: intervals built this way capture β_1 in 95% of repeated samples. It is not a 95% probability that β_1 lies in this particular interval (recall §7.2).

Step four: the test. H_0: $\beta_1 = 0$ says the true line is flat. The test statistic counts how many SEs b_1 sits from zero, $t = b_1 \div SE(b_1)$, compared with the same t-distribution (recall §8.5). The CI and the p-value are two views of one calculation: a 95% CI that excludes zero goes with $p < 0.05$.

Derivation (why $SE(b_1) = s \div \sqrt{Sxx}$)

1. Because $\Sigma(x_i - \bar{x}) = 0$, $Sxy = \Sigma(x_i - \bar{x})y_i$. So $b_1 = \Sigma\, w_i y_i$ with weights $w_i = (x_i - \bar{x})/Sxx$: a weighted sum of the outcomes.
2. Treat the x values as fixed. If the y_i are independent, each with variance σ^2 about the line, then $Var(b_1) = \Sigma\, w_i^2 \sigma^2$ (recall §6.5: variances of independent quantities add, and a constant multiplier comes out squared).
3. $\Sigma\, w_i^2 = \Sigma(x_i - \bar{x})^2/Sxx^2 = Sxx/Sxx^2 = 1/Sxx$. So $Var(b_1) = \sigma^2/Sxx$ and $SD(b_1) = \sigma/\sqrt{Sxx}$.
4. σ is unknown, so it is estimated by s: **$SE(b_1) = s \div \sqrt{Sxx}$**. Because σ is estimated, the t-distribution replaces the Normal (recall §5.7.2.3).

The formula shows the three things that shrink $SE(b_1)$: a smaller s (less scatter around the line), a larger n, and a wider spread of x (both make Sxx larger). A study of children aged 12 to 36 months estimates the age–weight slope far more precisely than one of 17- to 19-month-olds, because a short lever is harder to angle accurately than a long one.

Illustration / Example: Worked Calculation (clinic register)

1. SSE = 67.81 kg², n − 2 = 98. s = √(67.81 ÷ 98) = √0.692 = **0.83 kg**. Individual children typically sit within about ±1.7 kg (2 × 0.83) of the line.
2. √Sxx = √6190.8 = 78.68. SE(b_1) = 0.832 ÷ 78.68 = **0.011 kg per month**.
3. t = 0.14999 ÷ 0.01057 = **14.19** on 98 df (with the rounded SE, 0.150 ÷ 0.011 ≈ 14). p < 0.001.
4. t_{98} for 95% = 1.984. CI = 0.150 ± 1.984 × 0.0106 = 0.150 ± 0.021 = **0.129 to 0.171 kg per month**.

Predictor	Coefficient	Std. Error	t	p-value	95% CI
Age (months)	0.150	0.011	14.19	<0.001	0.129 to 0.171

Reading it: for each additional month of age, predicted weight rises by 0.15 kg. With an SE of 0.011, slopes around 0.14 or 0.16 would be unremarkable in a repeat of the study. The estimate sits about 14 SEs from zero, so p < 0.001. The data are compatible with true slopes from about 0.13 to 0.17 kg per month, that is, a monthly gain of about 130 to 170 grams. "p < 0.001" says only that zero is incompatible with the data; the interval gives a range a paediatrician can judge as important or trivial.

Two different intervals at x = 18 months. The **confidence interval for the mean** weight of all 18-month-olds uses SE = s√[1/n + (x − x̄)²/Sxx] = 0.832 × √[0.01 + 0.0065] = 0.107 kg: 11.20 ± 1.984 × 0.107 = 10.99 to 11.41 kg. The **prediction interval** for one new 18-month-old adds the child's own scatter, SE = s√[1 + 1/n + (x − x̄)²/Sxx] = 0.839 kg: 11.20 ± 1.984 × 0.839 = 9.54 to 12.86 kg. The first narrows as n grows; the second can never be narrower than about ±2s.

IN R (OPTIONAL)

```
cl  ← read.csv("data/clinic_children.csv")
fit ← lm(weight_kg ~ age_months, data = cl)
summary(fit)
confint(fit)
predict(fit, data.frame(age_months = 18), interval = "confidence")
predict(fit, data.frame(age_months = 18), interval = "prediction")
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.49963    0.27053   31.42   <2e-16 ***
age_months   0.14999    0.01057   14.19   <2e-16 ***
...
Residual standard error: 0.8318 on 98 degrees of freedom
Multiple R-squared:  0.6726,    Adjusted R-squared:  0.6692
F-statistic: 201.3 on 1 and 98 DF,  p-value: < 2.2e-16

                2.5 %    97.5 %
(Intercept) 7.9627662 9.0364895
age_months  0.1290151 0.1709744

       fit      lwr      upr
1 11.19953 10.98741 11.41166
       fit      lwr      upr
1 11.19953 9.535254 12.86381
```

The `age_months` row gives b_1 = 0.150, SE 0.0106 and t = 14.19 (steps 2–3); "Residual standard error: 0.8318 on 98 degrees of freedom" is s from step 1. `confint` gives the 0.129 to 0.171 interval of step

4. The two `predict` lines are the CI for the mean and the prediction interval for one child at 18 months. The R^2 lines are read in §12.2.5.

MUST-KNOW NOTES (MD-LEVEL)

- **$s = \sqrt{[SSE/(n - k - 1)]}$** ($n - 2$ in simple regression): the typical distance of a point from the line, in outcome units. **$SE(b_1) = s/\sqrt{Sxx}$** in simple regression.
- **95% CI:** $b_1 \pm t(n-k-1) \times SE(b_1)$. **Test:** $t = b_1 \div SE(b_1)$ for H_0: $\beta_1 = 0$.
- Report the CI, not only the p-value: it shows the size of the effects compatible with the data.
- A statistically significant coefficient is not automatically an important one. With a large sample, a clinically negligible slope can reach a tiny p-value (recall §8.5).
- A non-significant coefficient does not show that the predictor is unrelated to the outcome. A wide CI spanning zero means the study could not distinguish a meaningful association from none.
- A prediction interval for an individual is always wider than the CI for the mean at the same x.

12.2.5 R-Squared

DEFINITION

R-squared (R^2) is the proportion of the total variation in the outcome that the regression model accounts for: $R^2 = SSR \div SST = 1 - SSE \div SST$, where $SST = \Sigma(y_i - \bar{y})^2$ is the total sum of squares, SSE the residual sum of squares, and $SSR = SST - SSE$ the regression (explained) sum of squares. It runs from 0 to 1.

SIMPLIFIED EXPLANATION

SST measures how much the children's weights vary around their mean. After fitting the line, SSE is the variation left over around the line. The part that disappeared, SSR, is what age accounts for. This is the same partition of variation into parts that ANOVA uses (recall §9.5.1).

In simple regression, $R^2 = r^2$. Proof: $SSR = \Sigma(\hat{y}_i - \bar{y})^2 = b_1^2 Sxx = (Sxy/Sxx)^2 Sxx = Sxy^2/Sxx$, so $SSR/SST = Sxy^2/(Sxx \cdot Syy) = r^2$.

ILLUSTRATION / EXAMPLE

For the clinic register, SST = 207.1, SSE = 67.8 and SSR = 139.3 kg^2. $R^2 = 139.3 \div 207.1 =$ **0.67**, and $r^2 = 0.820^2 = 0.67$ as well. About 67% of the variation in the children's weights is accounted for by differences in their age. The remaining 33% reflects everything else: feeding, illness, genetics and ordinary individual variation.

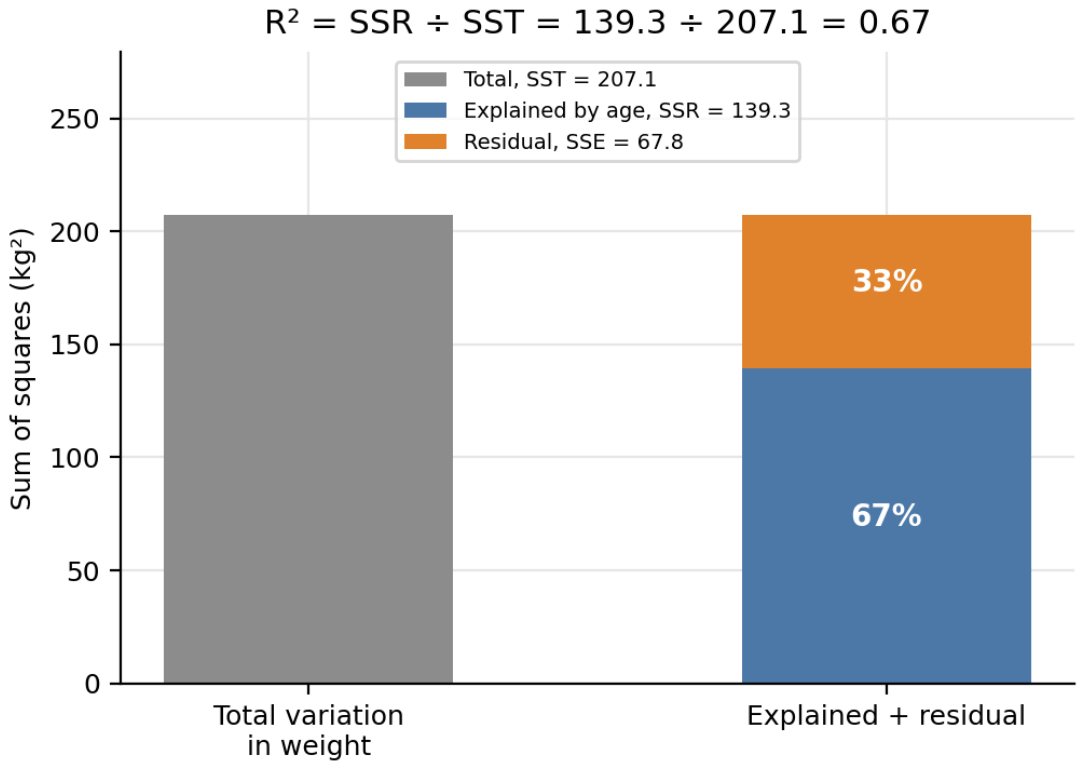


Figure 12.2. R^2 for the clinic regression of weight on age: of the total sum of squares (207.1 kg^2), age accounts for 139.3 (67%) and 67.8 (33%) remains as residual.

Adjusted R^2. Adding any predictor, even a column of random numbers, can only lower SSE, so plain R^2 can only rise. That makes plain R^2 useless for comparing models of different sizes. Adjusted R^2 replaces sums of squares with variance estimates:

adjusted $R^2 = 1 - [SSE/(n - k - 1)] \div [SST/(n - 1)] = 1 - (1 - R^2)(n - 1)/(n - k - 1)$

Derivation (when adjusted R^2 rises)

1. $SST/(n - 1)$ does not depend on the model, so adjusted R^2 rises exactly when $s^2 = SSE/(n - k - 1)$ falls.
2. Add one predictor. SSE falls by some amount Δ, and the residual df falls from $d + 1$ to d. Write SSE_1 for the new SSE.
3. s^2 falls when $SSE_1/d < (SSE_1 + \Delta)/(d + 1)$. Multiply out: $SSE_1 < d\Delta$, that is, $\Delta \div (SSE_1/d) > 1$.
4. $\Delta \div (SSE_1/d)$ is the partial F statistic for the new predictor. For a single added coefficient this statistic equals the square of that coefficient's t statistic (an algebraic identity whose proof is beyond this book; it matches the distributional fact $t^2 = F$ of §5.7.2.5). So **adjusted R^2 rises if and only if the new predictor's $|t| > 1$**, and falls when the predictor does not earn its place.

The bar $|t| > 1$ is much lower than significance at 5% ($|t| \approx 2$), so a rise in adjusted R^2 does not show that a predictor matters.

WORKED CALCULATION

1. Clinic model: $n = 100$, $k = 1$, $R^2 = 0.6726$. Adjusted $R^2 = 1 - 0.3274 \times 99/98 = 1 - 0.3308 =$ **0.669**.
2. Add a column of random numbers as a second predictor (simulated noise). Its $t = -0.68$. R^2 rises to 0.674, but adjusted R^2 falls to 0.667, as step 4 predicts for $|t| < 1$.

MUST-KNOW NOTES (MD-LEVEL)

- $R^2 = SSR/SST = 1 - SSE/SST$. In simple regression $R^2 = r^2$.
- R^2 measures statistical fit, not usefulness. A model with modest R^2 built on modifiable risk factors may matter more than a high-R^2 model built on age and sex alone. A modest R^2 can still go with a precisely estimated, important slope.
- Use adjusted R^2 for any model with more than one predictor, and for comparing models of different size. It rises only when the new predictor's $|t|$ exceeds 1.

CHECKPOINT 12.2

1. A colleague adds a fourth predictor to a model and R^2 rises from 0.41 to 0.43. They conclude the new predictor is clinically important. What is the flaw, and what should they check instead?
2. Explain, using the error term, why a fitted line almost never passes through every data point, even when the model is a good one.
3. A regression of haemoglobin (g/dL) on weeks of iron supplementation in 27 women gives b_1 = 0.12 g/dL per week, s = 0.9 g/dL and Sxx = 900 weeks^2. Compute SE(b_1), the t statistic and the 95% CI for β_1 ($t_{25} = 2.060$). Is the result compatible with no linear association?

(Answers in the Appendix at the end of the book.)

12.3 Assumptions of Linear Regression

DEFINITION

The CIs and p-values of linear regression rest on four assumptions, often remembered as **LINE**: **L**inearity, **I**ndependence of observations, **N**ormality of residuals, and **E**qual variance of residuals (homoscedasticity).

SIMPLIFIED EXPLANATION

Least squares will fit a line to any data (§12.2.2). Whether that line and its standard errors mean what they claim depends on these four conditions. Each protects a different part of the output.

ILLUSTRATION / EXAMPLE

If weight gain slowed in the older children but a straight line were fitted anyway, the line would over-predict at the ends of the age range and under-predict in the middle. The residual plot would show a curve (§12.5.1). For the clinic register the four standard plots show none of the warning patterns: the residuals scatter evenly around zero, follow the Normal line, keep a constant spread, and no child has a large Cook's distance (a measure of how much one child moves the fitted line, defined in §12.5.2).

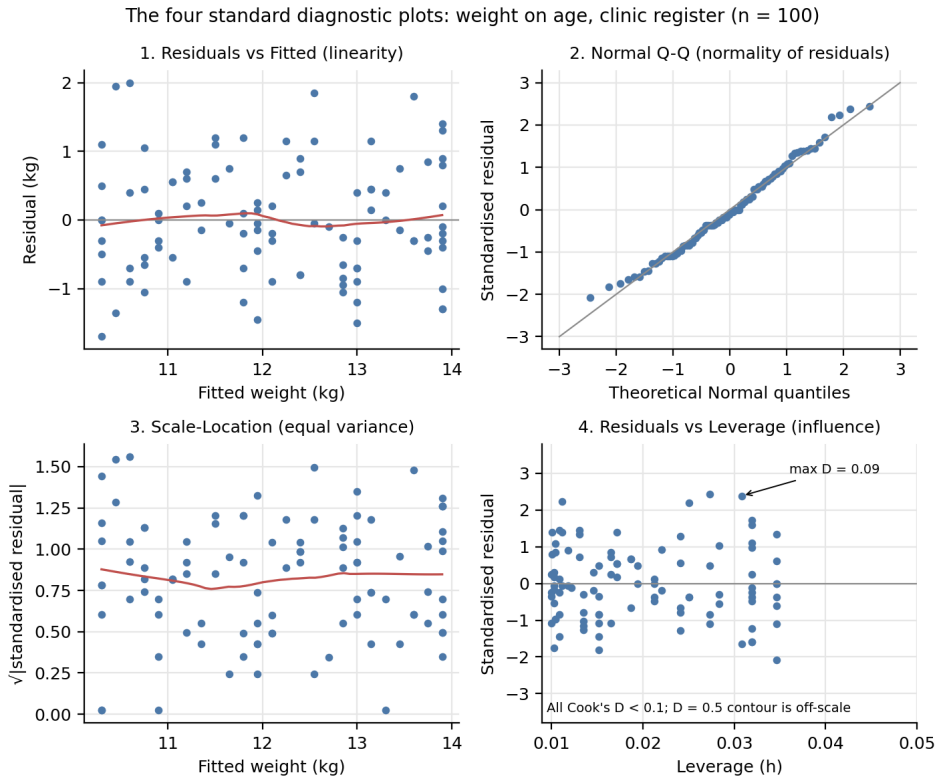


Figure 12.3. The four standard diagnostic plots (R's default set of regression plots) for weight on age in the clinic register. 1: residuals against fitted values, with a smoother, checks linearity. 2: Normal Q-Q plot of standardised residuals (residuals scaled to SD units, §12.5.1) checks normality. 3: Scale-Location, √|standardised residual| against fitted values, checks equal variance. 4: standardised residuals against leverage (how far a child's age lies from the mean age) check for influential points (§12.5.2). All four look healthy here.

MUST-KNOW NOTES (MD-LEVEL)

- **Linearity** protects the coefficient itself. If the true relationship is curved, a straight-line slope describes something that does not exist. Check with the scatter diagram and the residuals-vs-fitted plot; §12.4.5 shows how to model a curve.
- **Independence** protects the standard errors. Repeated measurements on one child, or children sampled within the same village, carry less information than independent observations. The model treats them as independent, so the SEs are too small, the CIs too narrow and the p-values too small: false positives become more likely (recall §6.2.6: the design effect, DEFF). Methods for clustered data are in §15.2.
- **Normality of residuals** protects CIs and p-values in small samples, not the coefficient. It concerns the residuals, not the raw outcome. In large samples linear regression is valid for almost any residual distribution, through the Central Limit Theorem (Lumley et al., 2002).
- **Equal variance** protects the standard errors. If residuals fan out as fitted values grow (heteroscedasticity), a single s is too small in part of the range and too large elsewhere.
- Linearity failures corrupt the coefficient; the other three corrupt the uncertainty around it.

CHECKPOINT 12.3

1. A residual-vs-fitted plot shows residuals fanning out into a cone as fitted values increase. Which assumption is violated, and what does it threaten: the coefficient, or something else?

2. Why does violating independence raise the risk of a false-positive result, rather than just making the estimate less precise?
3. The clinic regression (SE of the slope 0.011 kg per month) is refitted on children who were in fact sampled in village clusters, with a design effect of 2.44 for the slope. Roughly what is the correct SE, and what happens to the width of the 95% CI?

(Answers in the Appendix at the end of the book.)

12.4 Multiple Linear Regression

DEFINITION

Multiple linear regression includes two or more predictors: $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$. Each coefficient is the expected change in the outcome for a one-unit increase in that predictor, holding the other predictors in the model constant.

SIMPLIFIED EXPLANATION

Real questions involve several variables at once: age, sex, wealth and maternal education may all relate to a child's weight. Multiple regression fits one equation with a coefficient for each, chosen by the same least-squares principle (§12.2.2). Each coefficient describes a predictor's association with the outcome *after* allowing for the others. That is the "adjustment" that makes multiple regression the main statistical tool for handling measured confounding (§12.4.1).

ILLUSTRATION / EXAMPLE

In the clinic register, weight regressed on age and maternal education (none as reference) gives: $\hat{y}$ = 8.08 + 0.151 × age + 0.39 × (primary) + 0.77 × (secondary or higher). The age coefficient, 0.151 kg per month, now compares children of the same maternal-education group. It is almost the same as the simple-regression slope (0.150), so maternal education does not distort the age–weight association in these data.

MUST-KNOW NOTES (MD-LEVEL)

- The same LINE assumptions apply (§12.3), plus one new concern: predictors that are strongly correlated with each other (multicollinearity, §12.4.4).
- R^2 now gives the proportion of variation explained by all predictors together; use adjusted R^2 (§12.2.5).
- Degrees of freedom for each coefficient's t-test are $n - k - 1$.

12.4.1 Confounding and Adjustment

DEFINITION

Adjustment means including a potential confounder as an additional predictor, so that the exposure's coefficient estimates its association with the outcome among people with the same value of the confounder.

SIMPLIFIED EXPLANATION

Recall §11.6: a confounder is associated with the exposure, affects the outcome, and is not on the causal path between them. It can create or distort an apparent relationship. Including it in the model compares like with like: the exposure coefficient is estimated within levels of the confounder, as a stratified analysis does (recall §11.7), but for many strata and continuous confounders at once.

ILLUSTRATION / EXAMPLE

Recall the radio-ownership example of §11.6, with household wealth as the confounder. A regression of child nutritional status on radio ownership alone shows a strong association, because wealthier households both own radios and feed children better. In predicted nutritional status = $b_0 + b_1 \times$ radio $+ b_2 \times$ wealth, b_1 compares households of similar wealth, and the spurious part of the association is removed.

MUST-KNOW NOTES (MD-LEVEL)

- Regression adjusts only for confounders that were measured, measured well, and entered in the right form. An unmeasured confounder is untouched; a crudely measured or dichotomised one leaves **residual confounding** (recall §11.6, which separates it from unmeasured confounding; §12.4.5 shows an example). Design-stage alternatives (restriction, matching, stratification) are compared in §15.3.
- Adjustment can change a coefficient's size, significance or direction. A large change is *consistent with* confounding but does not prove it: chance and model specification can also shift a coefficient. In logistic regression, an odds ratio can change on adjustment even with no confounding at all, a property called non-collapsibility (§13.2.5, adjusted odds ratios). Which variables are confounders is decided from subject-matter knowledge of the causal structure (recall §11.6.1, causal diagrams), not from the size of the shift.
- Do not adjust for a variable on the causal path (a mediator): that removes part of the effect you want to estimate.

12.4.2 Interpreting Multivariable Coefficients

DEFINITION

In a multiple regression, each coefficient is the expected difference in the outcome per one-unit difference in that predictor, **holding the other predictors in the model constant**.

SIMPLIFIED EXPLANATION

The qualifier is part of the result. Dropping it makes an adjusted finding sound like a bigger, simpler claim than it is.

ILLUSTRATION / EXAMPLE

From §12.4: "a one-month increase in age is associated with 0.151 kg higher weight (95% CI 0.131 to 0.170), holding maternal education constant" is complete. "Age increases weight by 0.15 kg" drops the adjustment, drops the units of the predictor, and turns an association into an effect.

MUST-KNOW NOTES (MD-LEVEL)

- Always add "adjusting for ...", "holding ... constant" or "controlling for ..." when reporting a multivariable coefficient.
- A coefficient depends on which other predictors are in the model. There is no single "effect of age": there is the age coefficient adjusted for a stated set of variables.
- For a categorical predictor, the coefficient is the difference from the reference category, holding the other predictors constant; always name the reference (§12.4.3).

12.4.3 Categorical Predictors and Interaction

DEFINITION

A **categorical predictor** with k categories enters a model as k − 1 **dummy variables** (each coded 0/1), with the omitted category as the **reference**. An **interaction term**, the product of two predictors, lets the effect of one predictor differ by the level of the other; this is the regression form of **effect modification** (recall §11.7).

SIMPLIFIED EXPLANATION

Categorical predictors. A binary predictor is simple: code it 0/1, and its coefficient is the difference in mean outcome between the two groups, holding other predictors constant. With male = 0 and female = 1, the coefficient reads "average difference for girls compared with boys".

For more than two categories, create one dummy per category *except one*. With four socioeconomic classes you create three dummies; the fourth is the reference, and every coefficient compares with it. One category must be left out because a fourth dummy would be redundant: someone not in classes 1, 2 or 3 must be in class 4 (an extreme case of the redundancy problem of §12.4.4).

Changing the reference category changes every coefficient in that set but leaves the fitted values unchanged. It is a reporting choice. Choose a reference that makes comparisons interpretable, such as the largest group or a natural baseline.

Interaction. A model without interaction forces a predictor to have one effect for everybody. A nutrition supplement might help malnourished children substantially and well-nourished children hardly at all. The interaction model

$$\hat{y} = b_0 + b_1(\text{supplement}) + b_2(\text{malnourished}) + b_3(\text{supplement} \times \text{malnourished})$$

allows this. b_1 is the supplement's effect *in the reference group only* (well-nourished children), and b_3 is how much that effect *changes* in malnourished children, whose supplement effect is $b_1 + b_3$. Once an interaction is in the model, b_1 is no longer "the effect of the supplement"; reporting it as such misreads the output.

Interaction and confounding are distinct. Confounding is a distortion you adjust away to recover the association you want. Effect modification is a finding: the association really differs between groups, and you report it. One variable can be both. Interaction also depends on the scale: effects that add up exactly on the kilogram (additive) scale can show interaction on a ratio (multiplicative) scale, and vice versa, which is why interaction should be reported on both scales (Knol & VanderWeele, 2012). In linear regression the interaction term measures departure from additivity; §13.2.5 takes up the multiplicative scale of logistic regression.

ILLUSTRATION / EXAMPLE

A model of child weight with socioeconomic class (four categories, class 1 as reference) and sex (illustrative numbers):

Predictor	Coefficient	95% CI	Interpretation
Class 2	+0.42	0.10 to 0.74	0.42 kg heavier than class 1, adjusting for sex
Class 3	+0.88	0.55 to 1.21	0.88 kg heavier than class 1, adjusting for sex
Class 4	+1.35	0.98 to 1.72	1.35 kg heavier than class 1, adjusting for sex
Female	−0.31	−0.55 to −0.07	0.31 kg lighter than boys, adjusting for class

There is no "class 1" row, because class 1 is the baseline built into the intercept. To compare class 3 with class 2, subtract: 0.88 − 0.42 = 0.46 kg. Its CI and test need the model refitted with class 2 as reference (or the covariance of the two coefficients).

Now add an interaction between supplementation and baseline malnutrition, with b_1 (supplement) = +0.20 and b_3 (supplement × malnourished) = +0.75. The supplement raises weight by 0.20 kg in well-nourished children and by 0.20 + 0.75 = **0.95 kg** in malnourished children. Reporting only "the supplement increases weight by 0.20 kg" would understate its effect nearly fivefold in the group that most needs it.

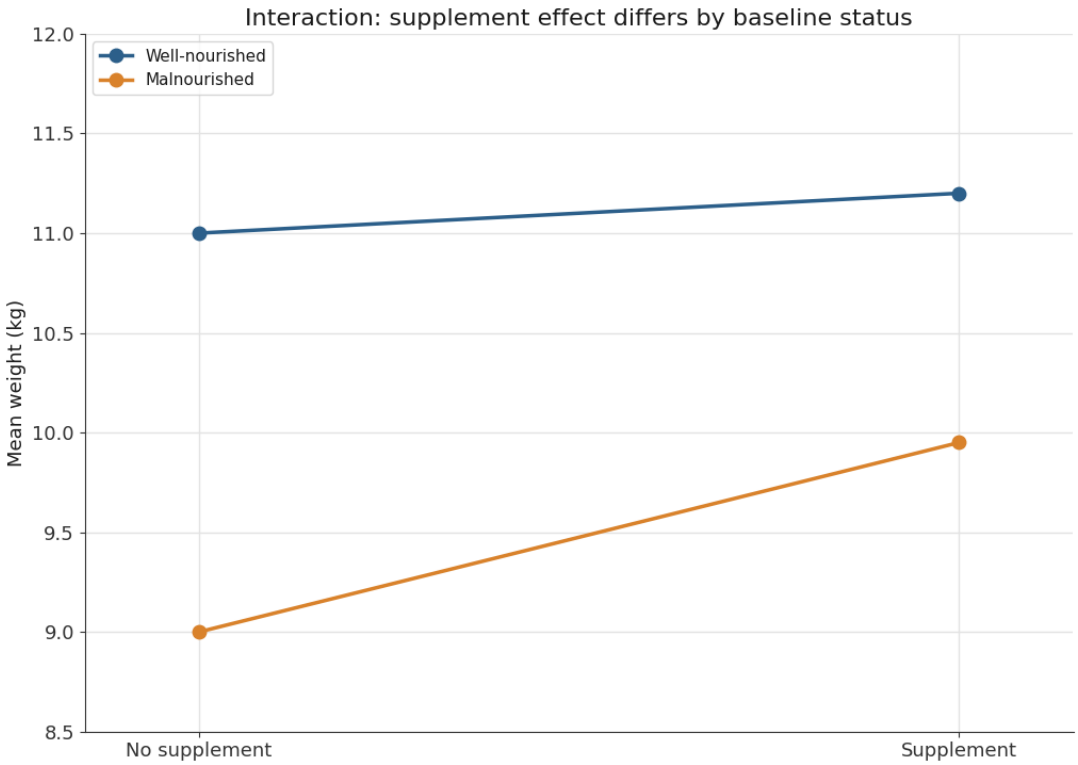


Figure 12.4. An interaction: mean weight rises by 0.20 kg with the supplement in well-nourished children and by 0.95 kg in malnourished children. Non-parallel lines are the picture of an interaction.

MUST-KNOW NOTES (MD-LEVEL)

- k categories need k − 1 dummies. The omitted category is the reference, absorbed into the intercept. State the reference whenever you report a categorical coefficient.
- Changing the reference changes coefficients, not the fit or the predictions.
- An interaction term is the product of two predictors; its coefficient is how much one predictor's effect changes across levels of the other.
- **With an interaction in the model, the main-effect coefficients apply only at the reference level of the other variable.** Reading them as overall effects is an easy and serious misreading.
- Specify interactions in advance, for a substantive reason. Testing every possible interaction until one is significant is a multiple-comparisons problem (recall §9.5.3).

12.4.4 Multicollinearity

DEFINITION

Multicollinearity is strong correlation among the predictors in a multiple regression, which makes it hard to separate each predictor's own contribution to the outcome.

SIMPLIFIED EXPLANATION

A multivariable coefficient uses only the part of a predictor that the other predictors do not already explain. If two predictors are highly correlated, like maternal height and maternal weight, little independ-

ent information is left in either once the other is in the model. The two compete to explain the same shared variation.

ILLUSTRATION / EXAMPLE

Entering both maternal height and maternal weight as predictors of birthweight, when height and weight are strongly correlated, makes it hard to say which is more strongly related to birthweight. Their coefficients have large SEs and may take unexpected signs, even though the model as a whole predicts birthweight well.

MUST-KNOW NOTES (MD-LEVEL)

- Symptoms: unexpectedly large SEs, counterintuitive or unstable signs, and coefficients that change a lot when a minor variable is added or removed.
- Multicollinearity does not bias predictions or R^2. It makes the individual coefficients of the correlated predictors imprecise.
- It is diagnosed with the variance inflation factor (§12.4.4.1).

12.4.4.1 Variance Inflation Factor

DEFINITION

The **variance inflation factor (VIF)** of a predictor is how many times larger the variance of its coefficient is than it would be if that predictor were uncorrelated with the other predictors: VIF = 1 ÷ (1 − R^2_j), where R^2_j comes from regressing predictor j on all the other predictors.

SIMPLIFIED EXPLANATION

If the other predictors explain none of predictor j ($R^2_j = 0$), VIF = 1: no inflation. If they explain 80% of it, only 20% of its variation is left to estimate its coefficient from, and the variance is 1 ÷ 0.2 = 5 times larger. The SE grows by √VIF.

ILLUSTRATION / EXAMPLE

With maternal height and weight in the same birthweight model, suppose the other predictors explain 78% of maternal weight: VIF = 1 ÷ 0.22 = 4.5. The variance of the maternal-weight coefficient is 4.5 times what it would be without that correlation, and its SE √4.5 = 2.1 times larger.

In the §12.6 clinic study (the analysis journey: 240 children, weight on wealth, age and sex), wealth and age correlate only weakly (r = 0.16), and every VIF is 1.03 or less.

IN R (OPTIONAL)

```
m  ← read.csv("data/journey_mlr.csv")
m$sex ← factor(m$sex, levels = c("M", "F"))
fm ← lm(weight_kg ~ wealth_index + age_months + sex, data = m)
library(car)
vif(fm)
cor(m$wealth_index, m$age_months, use = "complete.obs")
```

```
wealth_index   age_months          sex
    1.025452     1.027001     1.001543
[1] 0.1575423
```

With one other continuous predictor, R^2_j is close to $r^2 = 0.158^2 = 0.025$, so VIF ≈ 1 ÷ 0.975 = 1.03, as printed. The wealth coefficient's SE is inflated by only √1.03 ≈ 1.01.

MUST-KNOW NOTES (MD-LEVEL)

- VIF = 1 ÷ (1 − R^2_j). The minimum is 1; there is no upper limit.
- Cut-offs of 5 or 10 are conventions, not laws. The "rule of 10" is widely used, but a VIF of 10 or more does not by itself discredit a regression or require dropping a variable (O'Brien, 2007). What matters is whether the coefficient of interest is precise enough for the question, which its CI shows directly.
- Remedies, when the imprecision matters: drop one of two predictors that measure the same thing, combine them into a single index, or use principal components (§16.5: new uncorrelated variables built as weighted sums of the originals). High VIF in adjustment covariates that are not of interest usually needs no remedy.

CHECKPOINT 12.4

1. Household wealth and household electricity access are strongly correlated. In a model containing both, each has a wide CI and neither is significant, though R^2 is high. What is the likely explanation, and how would you confirm it?
2. A crude regression coefficient for a drug's effect is 0.90, and the coefficient adjusted for age is 0.35. What does this comparison show, and what does it not show?
3. In a model with an interaction, b_1 (supplement) = 0.30 kg and b_3 (supplement × malnourished) = 0.50 kg. What is the supplement's estimated effect in malnourished children? Separately: regressing one predictor on the others gives $R^2_j = 0.75$. What is its VIF, and by what factor is its SE inflated?

(Answers in the Appendix at the end of the book.)

12.4.5 Non-Linear Relationships – Categories, Trend Tests and Splines

DEFINITION

A continuous predictor need not enter a model as a straight line. It can enter as **categories** (dummy variables for ranges of values), as a **score** for ordered categories (a **test for linear trend**), or as a **spline**: a smooth curve built from pieces joined at chosen values called **knots**. A **restricted cubic spline** uses cubic pieces between the knots and is constrained to be a straight line beyond the outer knots.

SIMPLIFIED EXPLANATION

The linearity assumption (§12.3) says each extra unit of x adds the same amount to y. Three tools relax or test that:

- **Categories** make no assumption about shape: each group gets its own mean. The price is several coefficients, and the assumption that the outcome is flat within each group and jumps at the boundaries.
- **A trend test** gives ordered categories the scores 1, 2, 3, … and fits one slope across them: one coefficient, testing whether the outcome rises steadily across the ordered groups (recall the chi-square test for trend, §10.6.5, the same idea for proportions). Comparing the trend model with the category model tests *departure* from a straight-line trend.
- **A spline** follows a smooth curve using a few extra coefficients. With 4 knots, a restricted cubic spline uses 3 coefficients instead of 1. The knots are usually placed at fixed centiles of x; for 4 knots, Harrell's defaults are the 5th, 35th, 65th and 95th centiles (Harrell, 2015). Comparing the spline with the straight line tests non-linearity.

Do not dichotomise a continuous predictor. Splitting age at 24 months, or BMI at 25, throws information away. It assumes every child just below the cut-off is like the youngest child, and every child just above it is like the oldest. Dichotomising loses power, creates residual confounding when the variable is a confounder, and a cut-point chosen by looking at the data biases the results (Royston, Altman & Sauerbrei, 2006). Keep the variable continuous, and use a spline if the shape is in doubt. Report categories, if clinically useful, as a description alongside the continuous analysis.

Derivation (the F-test comparing nested models)

1. A smaller model (straight-line trend) is **nested** in a larger one (categories) when setting some of the larger model's coefficients to fixed values gives the smaller.
2. The larger model never has a larger SSE. The question is whether the drop, SSE_small − SSE_large, is bigger than chance would give for the extra q coefficients.
3. F = [(SSE_small − SSE_large) ÷ q] ÷ [SSE_large ÷ df_large]: the drop per extra coefficient, divided by the residual variance of the larger model.
4. Under H_0 (the smaller model is adequate), F follows the F-distribution on (q, df_large) degrees of freedom (recall §5.7.2.5). With q = 1, $F = t^2$, as in §12.2.5.

Illustration / Example: Worked Calculation (clinic register, wealth quintile 1 = poorest, adjusted for age)

1. **Categories.** With quintile 1 as reference, the age-adjusted coefficients are 0.36, 0.70, 1.11 and 1.62 kg for quintiles 2 to 5. The steps (0.36, 0.34, 0.41, 0.51 kg) are roughly equal, so the pattern looks like a straight-line trend.
2. **Trend.** Entering quintile as the score 1 to 5 gives 0.40 kg per quintile (SE 0.048, t = 8.35, p < 0.001): weight rises steadily with wealth.
3. **Departure from trend.** SSE is 39.46 for the trend model (97 df) and 39.26 for the category model (94 df), so q = 3. F = [(39.46 − 39.26) ÷ 3] ÷ (39.26 ÷ 94) = 0.067 ÷ 0.418 = **0.16**, p = 0.92. No evidence that a straight-line trend is inadequate, so the one-coefficient trend model is the better summary.
4. **Spline for age.** A 4-knot restricted cubic spline (knots at 12, 20.65, 29 and 36 months) lowers SSE from 67.81 to 66.61 at a cost of 2 coefficients. F = (1.20 ÷ 2) ÷ (66.61 ÷ 96) = 0.87, p = 0.42. The straight line for age is adequate in these data.

5. **Cost of dichotomising age.** Splitting age at 24 months drops R^2 from 0.67 to 0.49, and the residual SD rises from 0.83 to 1.03 kg. Four age groups recover $R^2 = 0.68$, but only by spending 3 coefficients on a step-shaped curve.
6. **Residual confounding.** In the §12.6 study, the wealth coefficient adjusted for age as a continuous variable is 0.62 kg. Adjusted for age split at 24 months it is 0.70 kg: the split leaves part of age's confounding in place.

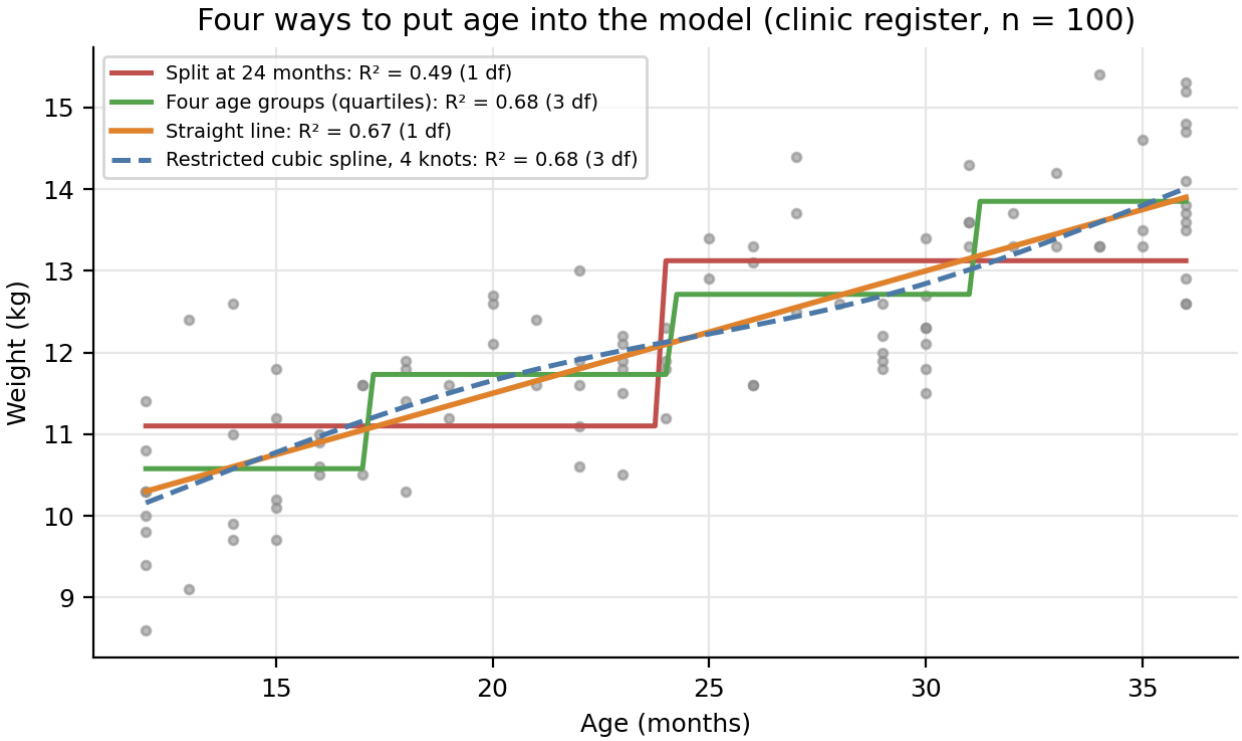


Figure 12.5. Weight on age in the clinic register, with age entered four ways. The split at 24 months ($R^2 = 0.49$) forces a jump at the cut-off and flat lines on either side. Quartile groups (0.68) and the spline (0.68) spend 3 coefficients each and follow the straight line (0.67), which is the adequate model here.

IN R (OPTIONAL)

```
suppressPackageStartupMessages(library(rms))    # for rcs()
cl ← read.csv("data/clinic_children.csv")
fcat ← lm(weight_kg ~ age_months + factor(wealth_quintile),
          data = cl)
ftr  ← lm(weight_kg ~ age_months + wealth_quintile, data = cl)
round(summary(fcat)$coef, 3)
round(summary(ftr)$coef, 3)
anova(ftr, fcat)                # departure from linear trend
flin ← lm(weight_kg ~ age_months, data = cl)
fspl ← lm(weight_kg ~ rcs(age_months, 4), data = cl)
anova(flin, fspl)               # is the spline better than the line?
```

```
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                7.560      0.269  28.081    0.000
age_months                 0.156      0.008  18.851    0.000
factor(wealth_quintile)2   0.362      0.220   1.643    0.104
factor(wealth_quintile)3   0.703      0.228   3.081    0.003
factor(wealth_quintile)4   1.111      0.230   4.826    0.000
factor(wealth_quintile)5   1.617      0.230   7.019    0.000
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       7.110      0.266  26.733        0
age_months        0.156      0.008  19.196        0
wealth_quintile   0.402      0.048   8.348        0

Model 1: weight_kg ~ age_months + wealth_quintile
Model 2: weight_kg ~ age_months + factor(wealth_quintile)
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     97 39.46
2     94 39.26  3   0.20003 0.1596 0.9232

Model 1: weight_kg ~ age_months
Model 2: weight_kg ~ rcs(age_months, 4)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     98 67.807
2     96 66.605  2     1.203 0.8669 0.4235
```

`factor()` makes the four dummies of step 1, each against quintile 1. Without `factor()`, the quintile enters as the score 1–5 and gives the single trend slope of step 2. The first `anova` is the F-test of step 3 (RSS is R's name for SSE); the second compares the spline with the straight line (step 4). `splines::ns(age_months, df = 3)` fits a similar natural spline without the `rms` package.

MUST-KNOW NOTES (MD-LEVEL)

- **Keep continuous predictors continuous.** Dichotomising loses power, leaves residual confounding, and a data-driven cut-point biases the estimate (Royston, Altman & Sauerbrei, 2006).
- **Categories** make no shape assumption but assume a flat outcome within each group. Fix the cut-points in advance (for example, quintiles or clinical bands), never by searching for the most significant split.
- **Test for trend:** code ordered categories 1, 2, 3, … and test the single slope. **Test for departure from trend:** F-test of the category model against the trend model.
- **Restricted cubic splines** model a smooth curve with k − 1 coefficients for k knots, linear beyond the outer knots; 3 to 5 knots at fixed centiles is the usual convention (Harrell, 2015). A spline coefficient has no simple interpretation: present the fitted curve as a plot.
- A non-significant test of non-linearity supports keeping the straight line; it does not prove the true relationship is linear. Choosing between linear, categorical and spline forms by repeated testing is itself a multiplicity problem; §15.1.5 (modelling strategy, how the aim of the model governs these choices) returns to this.

CHECKPOINT 12.5

1. A study reports "children with maternal BMI ≥ 25 had 0.4 kg higher birthweight (p = 0.04)", after the authors tried cut-offs of 22, 23, 24, 25 and 26. Give two reasons this analysis is weak, and say how you would reanalyse it.
2. Adding three dummy variables for four categories to a trend model lowers SSE from 52.0 to 50.5. The category model has 120 residual df. Compute F for departure from a linear trend and say whether it suggests the trend model is inadequate (the 5% critical value of F on 3 and 120 df is 2.68).

(Answers in the Appendix at the end of the book.)

12.5 Regression Diagnostics

Fitting a model is the first step. Diagnostics then check, from the model's own residuals and fitted values, whether the assumptions of §12.3 held in this dataset (§12.5.1) and whether one or two observations are driving the result (§12.5.2). Some problems, such as mild curvature, show far more clearly in a residual plot than in the raw scatter diagram.

12.5.1 Residual Analysis

DEFINITION

Residual analysis examines the residuals (observed − fitted) with plots to check linearity, normality and equal variance. It often uses **standardised residuals**, each residual divided by its own estimated SD, $s\sqrt{(1-h)}$, where h is the leverage of §12.5.2. About 95% of standardised residuals should lie between −2 and +2 if the model holds.

SIMPLIFIED EXPLANATION

The assumptions of §12.3 are assumptions about ε, and residuals are its observable stand-ins. If the model is right, residuals should look like random noise: centred on zero at every fitted value, with constant spread, and roughly Normal.

ILLUSTRATION / EXAMPLE

For the clinic regression, residuals plotted against fitted values scatter randomly around zero with constant spread (left panel below). The right panel shows what a problem looks like: a straight line fitted to simulated data in which weight gain slows with age. The residuals form an arch, positive in the middle and negative at both ends, because the line misses the curve systematically.

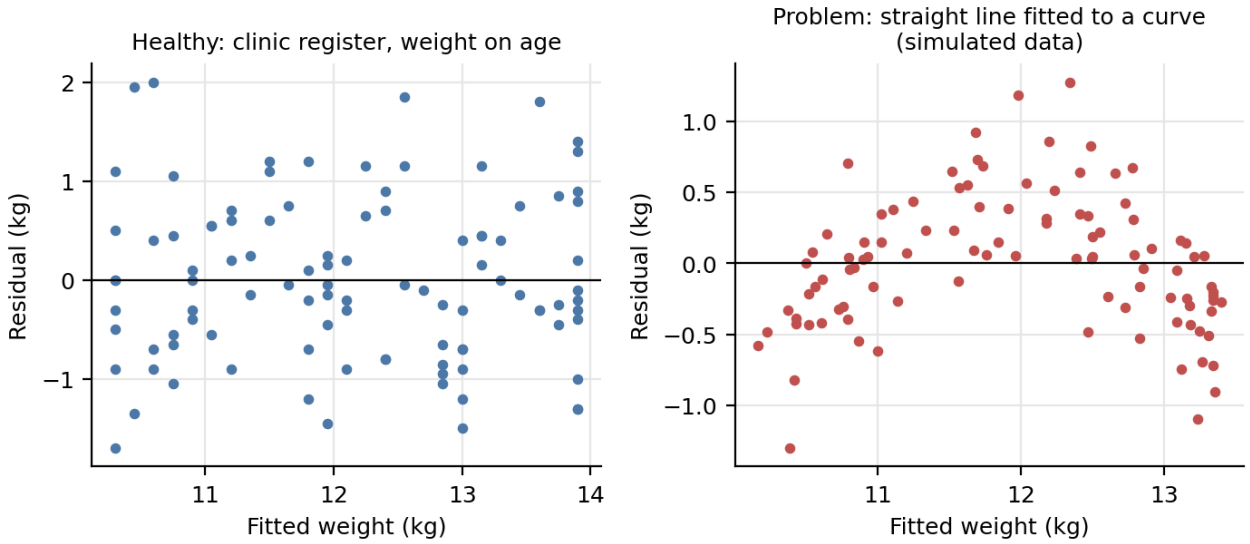


Figure 12.6. Residuals against fitted values. Left: the clinic regression of weight on age, a healthy random scatter. Right: simulated data in which weight gain slows with age, fitted with a straight line; the arch shows non-linearity.

IN R (OPTIONAL)

```
cl  ← read.csv("data/clinic_children.csv")
fit ← lm(weight_kg ~ age_months, data = cl)
par(mfrow = c(2, 2)); plot(fit)          # the four standard plots
range(rstandard(fit))                    # standardised residuals
max(hatvalues(fit)); max(cooks.distance(fit))
```

```
[1] -2.079537  2.438442
[1] 0.03463716
[1] 0.09016558
```

`plot(fit)` draws the four panels shown in §12.3: by default R gives residuals vs fitted, Normal Q-Q, Scale-Location and residuals vs leverage (R Core Team, `plot.lm`). The standardised residuals run from −2.08 to +2.44, as expected for 100 children. The largest leverage (0.035) and Cook's distance (0.09) are read in §12.5.2.

MUST-KNOW NOTES (MD-LEVEL)

- **Residuals vs fitted:** random scatter supports linearity and equal variance; a curve means non-linearity (§12.4.5); a funnel means heteroscedasticity.
- **Normal Q-Q plot** of residuals checks normality. The assumption concerns the residuals, not the raw outcome.
- **Scale-Location** plot: a rising trend in √|standardised residual| is the clearest sign of unequal variance.
- Look at the plots. A pattern in a plot says more than any single test statistic.

12.5.2 Influential Points and Leverage

DEFINITION

Leverage (h) measures how far a point's predictor values lie from the centre of the predictors. In simple regression, $h = 1/n + (x - \bar{x})^2/Sxx$; the leverages average $(k + 1)/n$. An **influential point** is one whose removal would substantially change the fitted coefficients. **Cook's distance** measures this: $D = (r_std^2/(k + 1)) \times h/(1 - h)$, where r_std is the standardised residual.

SIMPLIFIED EXPLANATION

A point with an extreme x has high leverage: like a long lever, it *can* move the line. Whether it does depends on its residual. If its y is where the line predicts, it changes nothing. A point with both high leverage and a large residual pulls the line towards itself. Cook's distance multiplies the two together.

ILLUSTRATION / EXAMPLE

Add one hypothetical, severely malnourished child weighing 8.0 kg to the 100 clinic children (`figs/ch12_influence.py`). If the child is 36 months old, the oldest age in the register, h = 0.031, the standardised residual is −5.7, Cook's D = 0.52, and the slope falls from 0.150 to 0.139 kg per month. The same child at 24 months, near the mean age, has h = 0.010 and D = 0.10, and the slope stays at 0.150. Same weight, same kind of outlier; only the leverage differs.

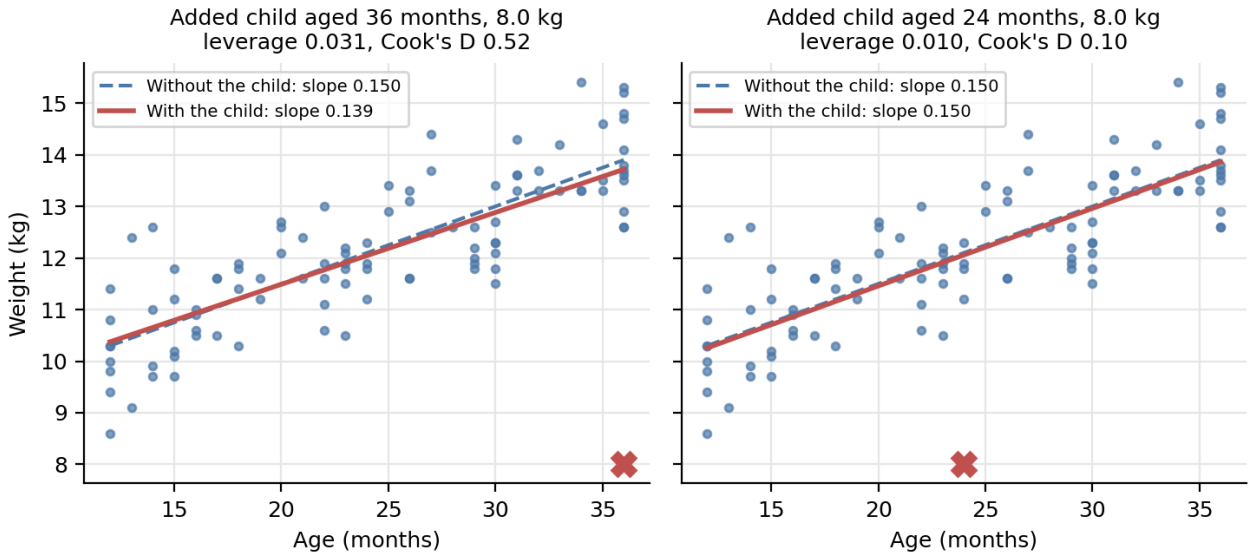


Figure 12.7. A hypothetical 8.0 kg child added to the 100 clinic children. Left: at 36 months (leverage 0.031, Cook's D 0.52) it tilts the line from a slope of 0.150 to 0.139 kg per month. Right: at 24 months (leverage 0.010, D 0.10) the slope is unchanged.

Worked Calculation (leverage of a 36-month-old in the original 100 children)

1. $h = 1/100 + (36 - 24.35)^2/6190.8 = 0.010 + 135.7/6190.8 = 0.010 + 0.022 =$ **0.032**.
2. The average leverage is $(k + 1)/n = 2/100 = 0.02$, so this child has 1.6 times the average. The 12-month-olds are furthest from the mean age and have the largest leverage in the register: $h = 0.010 + (12 - 24.35)^2/6190.8 =$ **0.035**. The largest Cook's distance among the real children, which software lists, is 0.09, against 0.52 for the hypothetical child: no real child pulls the line noticeably.

MUST-KNOW NOTES (MD-LEVEL)

- **Leverage** depends only on the predictor values, not the outcome. A high-leverage point can fit the line perfectly.
- **Influence** needs both leverage and a large residual; Cook's distance combines them. Compare each point's D with the rest. Refit without the point and see whether the conclusion changes: that is the direct test.
- An influential point is a prompt to investigate, not to delete. If it is a data-entry error or a child who does not belong to the study population, correct or exclude it and say so. If it is a genuine child, keep it and report how much the result depends on it.

CHECKPOINT 12.6

1. One unusual point has an extreme predictor value but lies on the regression line; another has an ordinary predictor value but a very unexpected outcome. Which has high leverage, and which a large residual? Which is more likely to be influential?
2. Why is an influential point a prompt to investigate rather than a reason to delete?
3. In a regression of 50 adults' systolic BP on age, $\bar{x}$ = 45 years and Sxx = 8,000 years^2. Compute the leverage of a 25-year-old and of a 75-year-old, and compare each with the average leverage.

(Answers in the Appendix at the end of the book.)

12.6 Analysis Journey – Multiple Linear Regression

This section runs one regression analysis from question to report, in the order a Community Medicine researcher would, using only tools built in this chapter.

Research question. Among children aged 12 to 36 months attending a district under-five clinic, is household wealth associated with weight (kg), independent of the child's age and sex?

Study design. A cross-sectional survey of 240 children attending the clinic over three months, with weight, age, sex and a household wealth index recorded at a single visit (`data/journey_mlr.csv`).

Outcome. Weight in kilograms: continuous, so linear regression is the natural first choice. (An alternative outcome is the weight-for-age z-score, recall §5.7.2.2.3, which is already standardised for age and sex; age would then not be entered as a covariate. This analysis keeps raw kilograms and adjusts for age and sex directly.)

Predictors and why each is there. Wealth index (continuous) is the exposure. Age in months enters for two separate reasons. First, it is a strong determinant of weight (r = 0.61 in these data), so including it removes a large part of the residual variation and makes the wealth coefficient more precise: a **precision covariate**. Second, age is a potential **confounder**: it cannot be caused by household wealth, and it would distort the crude wealth–weight comparison if wealthier and poorer households brought children of different ages to the clinic. That second condition was judged plausible before the analysis (care-seeking patterns differ by wealth); in this sample age and wealth correlate at r = 0.16. Sex is a determinant of weight with no plausible link to household wealth, so it enters as a precision covariate only.

Data preparation. Sex is coded 0/1 with boys as reference (§12.4.3). Three children (1.25%) have a missing wealth index. There is no evident reason why wealth data were missed differently for any group, so the analysis uses the 237 complete cases and says so; a larger or patterned loss would need the missing-data methods of §15.4 (such as multiple imputation).

Crude analysis. Weight on wealth alone: 0.90 kg per wealth-index unit (95% CI 0.52 to 1.29).

Model specification. Weight is regressed on wealth, age and sex together. The predictors were chosen by the subject-matter reasoning above, not by an automated search.

Assumptions. Linearity: residuals against fitted values show no curve, and a spline for wealth adds nothing (§12.4.5). Normality and equal variance: the Q-Q and Scale-Location plots look acceptable. Independence: each child attended once and no household contributed two children.

Diagnostics. Multicollinearity is negligible: every VIF is 1.03 or less (§12.4.4.1). One child, J090 (33 months, girl, wealth index −0.20, 17.6 kg), has a standardised residual of 3.73: a very heavy child from a below-average-wealth household. Her Cook's D is 0.077, the largest in the data. Nothing suggests a data-entry error, so she is retained. Refitting without her moves the wealth coefficient from 0.62 to 0.65 kg: the conclusion does not depend on her.

IN R (OPTIONAL)

```
m  ← read.csv("data/journey_mlr.csv")
m$sex ← factor(m$sex, levels = c("M", "F"))
cc ← na.omit(m)                          # 237 complete cases
crude ← lm(weight_kg ~ wealth_index, data = cc)
fm    ← lm(weight_kg ~ wealth_index + age_months + sex, data = cc)
round(cbind(coef(crude), confint(crude)), 2)
summary(fm)
round(confint(fm), 2)
i ← which.max(rstandard(fm)); cc[i, ]
c(rstandard(fm)[i], cooks.distance(fm)[i])
coef(update(fm, data = cc[-i, ]))["wealth_index"]
```

```
                  2.5 % 97.5 %
(Intercept)  11.53 11.33  11.74
wealth_index  0.90  0.52   1.29
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.28253    0.39858  18.271  < 2e-16 ***
wealth_index  0.62037    0.15819   3.922 0.000116 ***
age_months    0.18065    0.01589  11.368  < 2e-16 ***
sexF         -0.24944    0.16668  -1.496 0.135887
...
Residual standard error: 1.281 on 233 degrees of freedom
Multiple R-squared:  0.4127,    Adjusted R-squared:  0.4051

             2.5 % 97.5 %
(Intercept)   6.50   8.07
wealth_index  0.31   0.93
age_months    0.15   0.21
sexF         -0.58   0.08
   child_id age_months sex wealth_index weight_kg
90     J090         33   F         -0.2      17.6
        90         90
3.73109528 0.07652096
wealth_index
   0.6453372
```

The crude slope is 0.90 (0.52 to 1.29). In the adjusted model the wealth row gives 0.62 kg per unit, t = 3.92 on 233 df (237 children − 3 predictors − 1), p < 0.001. `sexF` is girls compared with boys. The last three results are the J090 checks described under Diagnostics.

Output.

Predictor	Coefficient	95% CI	p-value
Wealth index (per unit)	0.62	0.31 to 0.93	<0.001
Age (per month)	0.18	0.15 to 0.21	<0.001
Sex: girl (vs boy)	−0.25	−0.58 to 0.08	0.14

Adjusted R^2 = 0.41; residual SD 1.28 kg; n = 237.

Interpretation. Adjusting for age and sex, each one-unit higher wealth index is associated with 0.62 kg higher weight on average (95% CI 0.31 to 0.93 kg). Girls weighed 0.25 kg less than boys of the same age and wealth, but the interval (−0.58 to 0.08 kg) is compatible with anything from a moderate deficit to a small excess, so this study cannot tell whether sex matters. The model accounts for about 41% of the variation in weight, leaving room for factors not measured here (illness, feeding, birth weight).

Why did the wealth coefficient fall from 0.90 to 0.62? In a model with wealth and age only (sex barely matters here), least squares gives an exact identity: crude slope = adjusted slope + (age coefficient) × (slope of age on wealth). Here 0.621 + 0.180 × 1.569 = 0.903. Wealthier households' children were on average 1.57 months older per wealth unit, and each month adds 0.18 kg, so about 0.28 kg of the crude 0.90 was age. Age also did its precision job: without it the residual SD would be 1.59 kg instead of 1.28, and the SE of the wealth coefficient 0.194 instead of 0.158.

What the result does not mean. It does not show that raising a household's wealth would *cause* a 0.62 kg gain. The study is cross-sectional and observational, and the wealth index stands for a bundle of things (diet, housing, care-seeking) not measured separately; unmeasured confounding

by factors such as parental education or dietary diversity cannot be ruled out, and the wealth index itself may leave residual confounding (recall §11.6). Nor does p = 0.14 show that sex is unrelated to weight. And an average association says nothing certain about any one child.

Reporting. *"In a multiple linear regression of weight (kg) on household wealth index, adjusting for age and sex (n = 237 complete cases), wealth was positively associated with weight (adjusted β = 0.62 kg per unit, 95% CI 0.31 to 0.93, p < 0.001; crude β = 0.90). The association between sex and weight was imprecisely estimated (girls −0.25 kg, 95% CI −0.58 to 0.08). Adjusted R^2 = 0.41. Given the cross-sectional design, the association should not be interpreted causally."*

MUST-KNOW NOTES (MD-LEVEL)

- Order of reasoning: question and outcome; predictors, each with its reason (exposure, confounder, precision covariate); data preparation and crude analysis; adjusted model; assumptions and diagnostics; interpretation; limits.
- A **confounder** must be associated with the exposure and affect the outcome, off the causal path. A **precision covariate** affects the outcome but is unrelated to the exposure: it shrinks the SE without shifting the coefficient. Age here was both; sex was only the second.
- A crude-to-adjusted shift (0.90 → 0.62) is consistent with confounding and, in linear regression, can be traced exactly. It supports but does not replace the prior reasoning for adjusting.
- Every number in the output table earns a plain-language sentence, and the "what this does not mean" step keeps association and causation apart.

CHECKPOINT 12.7

1. In a two-predictor model, the adjusted coefficient for maternal education (years) on birthweight is 25 g per year, the coefficient for maternal height is 12 g per cm, and the slope of maternal height on education is 0.5 cm per year. What was the crude coefficient for education?
2. A colleague says "age is a confounder in every child-growth study". Using the confounder and precision-covariate definitions, explain when age confounds a wealth–weight association and when it is only a precision covariate.

(Answers in the Appendix at the end of the book.)

Generalised Linear Models

Linear regression (recall Chapter 12) models a continuous outcome whose errors are roughly Normal. Many outcomes in Community Medicine are not like that: a household joins CBHI or it does not, a village records so many malaria cases. This chapter shows that linear regression is one member of a larger family, the generalised linear models, and builds the members you will use most: logistic regression for binary outcomes and its variants (§13.2–13.3), Poisson and negative binomial regression for counts and rates (§13.4–13.6), and models that estimate prevalence and risk ratios directly (§13.7).

13.1 Why We Need a Generalised Framework

DEFINITION

A **generalised linear model (GLM)** has three parts:

1. A **distribution** for the outcome Y, given the predictors: Normal for a continuous outcome, binomial for a 0/1 outcome (recall §5.7.1.2), Poisson for a count (recall §5.7.1.3).
2. A **linear predictor**, $b_0 + b_1x_1 + b_2x_2 + \ldots$, exactly as in multiple regression (recall §12.4).
3. A **link function** g that connects the two: $g(\mu) = b_0 + b_1x_1 + \ldots$, where μ is the expected value (mean) of Y.

Ordinary linear regression is itself a GLM: Normal distribution, identity link ($g(\mu) = \mu$). The other GLMs change the distribution, the link, or both. GLMs can therefore also handle continuous outcomes that are not Normal, such as skewed costs, but this book uses them for binary and count outcomes.

SIMPLIFIED EXPLANATION

Fit a straight line to a 0/1 outcome and you meet a structural problem. A straight line runs to $\pm\infty$, but a probability must stay between 0 and 1. With enough spread in the predictor, the line will predict "probabilities" below 0 or above 1. Counts have the same problem at the bottom end: a line can predict −3 cases.

The GLM solution is to model a transformed mean that *can* take any value from $-\infty$ to $+\infty$, and transform back afterwards. For a probability the transformation is the logit (§13.2.2); for a count it is the natural logarithm (§13.5).

A GLM is not fitted by least squares. It is fitted by **maximum likelihood**. The **likelihood** of a set of parameter values is the probability of the data you actually observed, calculated as if those values were true. The **maximum likelihood estimate (MLE)** is the value that makes the observed data most probable. §15.1.1 develops likelihood further (model fit and comparison); here you need only the idea.

DERIVATION

Toss a coin 10 times and get 7 heads. If the probability of a head is p, and tosses are independent (recall §5.4), the probability of this particular sequence is

$L(p) = p^7 (1 - p)^3$.

The log-likelihood is $\ln L = 7 \ln p + 3 \ln(1 - p)$. Its slope, found by calculus (beyond this book; the table below shows the peak numerically), is $7/p - 3/(1 - p)$. Setting the slope to zero gives $7(1 - p) = 3p$, so $\hat{\mathbf{p}} = \mathbf{7/10 = 0.7}$. The MLE is the sample proportion, which is what intuition says. For linear regression with Normal errors, maximising the likelihood gives exactly the least-squares estimates (recall §12.2.2). This is why the two methods agree for the identity-link GLM.

ILLUSTRATION / EXAMPLE

Fit a straight line by least squares to CBHI membership (1 = member, 0 = not) against the wealth index in the 500-household survey: predicted membership = 0.315 + 0.077 × wealth. For the poorest household (wealth −5.27) the line predicts 0.315 − 0.077 × 5.27 = −0.09, and six households get a negative "probability". With a wider wealth range the line would also pass 1 at the top. These impossible predictions show that the straight line is the wrong model for this outcome.

WORKED CALCULATION

The likelihood of 7 heads in 10 tosses at five candidate values of p:

p	$L(p) = p^7(1 - p)^3$	ln L
0.5	0.000977	−6.931
0.6	0.001792	−6.325
0.7	**0.002224**	**−6.109**
0.8	0.001678	−6.390
0.9	0.000478	−7.645

The likelihood peaks at p = 0.7. Software does the same search for every coefficient of a GLM at once.

IN R (OPTIONAL)

```
cl ← read.csv("data/clinic_children.csv")
coef(lm(weight_kg ~ age_months, data = cl))
coef(glm(weight_kg ~ age_months, family = gaussian(link = "identity"),
         data = cl))
```

```
(Intercept)  age_months
  8.4996278   0.1499948
(Intercept)  age_months
  8.4996278   0.1499948
```

`glm` with a Gaussian (Normal) family and identity link returns the same intercept (8.50 kg) and slope (0.150 kg per month) as `lm` on the 100 clinic children: linear regression is the identity-link GLM. Changing `family` is all it takes to move to logistic (`binomial`) or Poisson (`poisson`) regression.

MUST-KNOW NOTES (MD-LEVEL)

- Three parts: distribution, linear predictor, link. Linear regression = Normal distribution + identity link.
- The coefficients act on the *link* scale. Each GLM needs its own back-transformation to give a readable measure: exp(b) is an odds ratio in logistic regression and a rate ratio in Poisson regression with an offset.
- GLMs are fitted by maximum likelihood, not least squares. The two coincide for the Normal, identity-link case.

13.2 Logistic Regression

Logistic regression models the log-odds of a binary outcome as a linear function of predictors. Odds and the odds ratio are recalled first (§13.2.1), then the logit (§13.2.2), the binary model (§13.2.3), its assumptions (§13.2.4), adjustment and interaction (§13.2.5), and how to read the output (§13.2.6). The section ends with the variants for ordered (§13.2.7), unordered (§13.2.8) and individually matched (§13.2.9) outcomes.

13.2.1 Odds and Odds Ratio

DEFINITION

The **odds** of an event are p ÷ (1 − p) (recall §5.5.3). The **odds ratio (OR)** is the odds of the event in one group divided by the odds in another (recall §10.6.4). It says how many times higher or lower the *odds* are in the first group.

SIMPLIFIED EXPLANATION

Odds and probability move in the same direction but are different numbers. An OR is a ratio of odds, not of probabilities. It approximates the risk ratio only when the outcome is rare in both groups (recall §10.6.4). Logistic regression works on odds because of the logit (§13.2.2).

ILLUSTRATION / EXAMPLE

If 20% of CBHI members in one group drop out, their odds of dropout are 0.2 ÷ 0.8 = 0.25 (“1 to 4”). If 10% drop out in a comparison group, their odds are 0.1 ÷ 0.9 = 0.111. OR = 0.25 ÷ 0.111 = **2.25**: the odds of dropout are 2.25 times as high in the first group. The risk ratio is 0.2 ÷ 0.1 = 2.0. With a 10–20% outcome the two already differ.

MUST-KNOW NOTES (MD-LEVEL)

- OR = 1: no difference in odds. OR > 1: higher odds in the first (usually exposed) group. OR < 1: lower odds.
- Do not read an OR as "times as likely". That is risk-ratio language, and it is wrong unless the outcome is rare.
- exp(coefficient) from a logistic regression is an OR for a one-unit increase in that predictor, holding the other predictors constant (§13.2.3).

13.2.2 The Logit Transformation

DEFINITION

The **logit** of a probability p is its log-odds: logit(p) = ln[p ÷ (1 − p)]. It maps probabilities between 0 and 1 onto the whole real line.

SIMPLIFIED EXPLANATION

This is the transformation §13.1 asked for. As p approaches 0 the logit heads to −∞; as p approaches 1 it heads to +∞. A straight line on the logit scale can therefore never give an impossible probability.

ILLUSTRATION / EXAMPLE

p = 0.5 has odds 1 and logit ln(1) = 0, the midpoint. p = 0.9 has odds 9 and logit ln 9 = 2.197. p = 0.1 has odds 1/9 and logit −2.197. The scale is symmetric about zero.

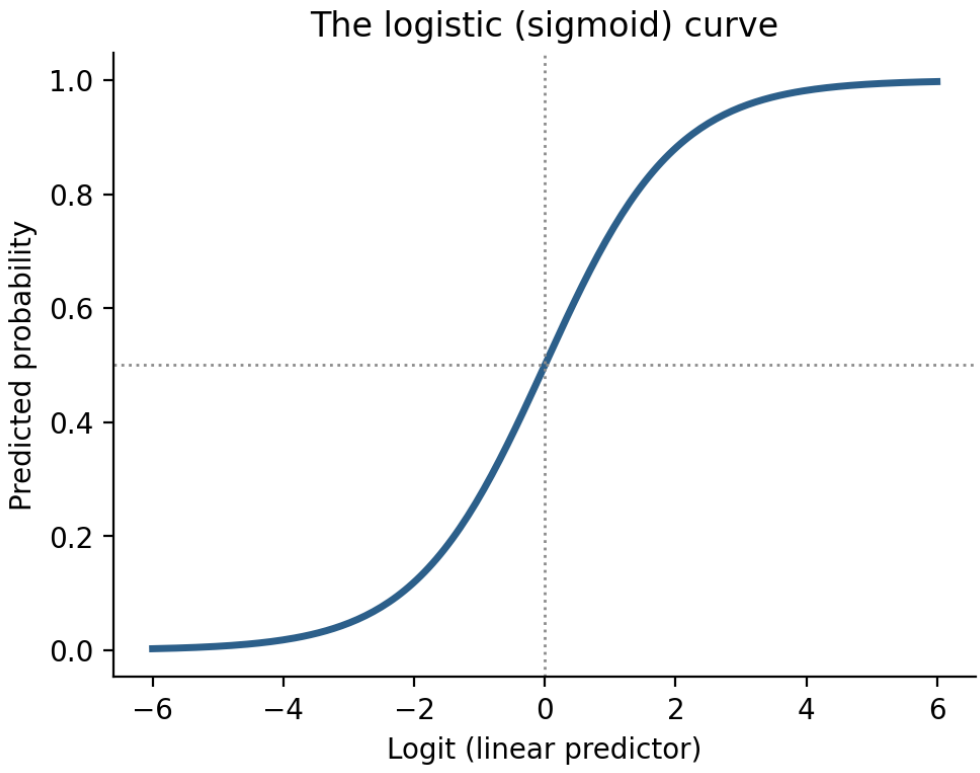


Figure 13.1. The logistic (sigmoid) curve linking the logit to a probability.

MUST-KNOW NOTES (MD-LEVEL)

- Model: logit(p) = $b_0 + b_1x_1 + b_2x_2 + \ldots$ The logit is the link function (recall §13.1) of logistic regression.
- Back-transform with the **inverse logit**: p = 1 ÷ (1 + e^(−logit)). Plotted against a predictor, p traces the S-shaped curve of Figure 13.1; the logit itself is a straight line.
- The model assumes linearity on the logit scale, not the probability scale. §13.2.4 shows how to check it.

13.2.3 Binary Logistic Regression

DEFINITION

Binary logistic regression models the log-odds of one category of a two-category outcome (coded 1, the "event") as a linear function of the predictors, fitted by maximum likelihood (recall §13.1).

SIMPLIFIED EXPLANATION

The simplest form of logistic regression: yes/no, member/non-member, died/survived. The likelihood extends the coin of §13.1. Household i has its own probability p_i, given by the inverse logit of its linear predictor. A member contributes p_i to the likelihood and a non-member contributes $1 - p_i$:

$L = \Pi\, p_i^{(y_i)}\, (1 - p_i)^{(1 - y_i)}$,

where Π (capital pi) means "multiply together over all households", as Σ means "add up", and a power of $y_i = 1$ or 0 keeps or drops that factor.

Software searches for the b's that make L largest. There is no closed-form answer as there is for least squares, so the search is iterative.

ILLUSTRATION / EXAMPLE

In the CBHI survey (500 households in 20 villages, 25 per village; 150 members, 30%), regress membership on distance to the nearest facility alone. The fitted coefficient for distance is −0.162 per km, so OR = e^(−0.162) = 0.85: each extra kilometre multiplies the odds of membership by 0.85.

IN R (OPTIONAL)

```
cb ← read.csv("data/cbhi_survey.csv")
m0 ← glm(cbhi_member ~ distance_km, family = binomial, data = cb)
summary(m0)$coefficients
exp(cbind(OR = coef(m0), confint.default(m0)))
```

```
             Estimate Std. Error   z value     Pr(>|z|)
(Intercept)  0.3367709 0.24028160  1.401568 1.610444e-01
distance_km -0.1622277 0.03098085 -5.236387 1.637501e-07
                   OR     2.5 %    97.5 %
(Intercept) 1.4004182 0.8744393 2.2427755
distance_km 0.8502476 0.8001556 0.9034754
```

The `Estimate` column is on the log-odds scale: −0.162 per km. The Wald z is the estimate ÷ its SE (−0.162/0.031 = −5.24). `exp()` turns the coefficient and its Wald CI (b ± 1.96 SE) into the OR: **0.85 (95% CI 0.80 to 0.90)**. This is the *crude* OR; §13.3 adjusts it. The intercept, 0.34, is the log-odds of membership at 0 km.

MUST-KNOW NOTES (MD-LEVEL)

- The outcome has exactly two categories, coded 0 and 1. The choice of which is 1 sets the direction of every OR, so state it.
- Estimation is by maximum likelihood, not least squares.
- Several predictors enter exactly as in multiple regression (§13.2.5).

13.2.4 Assumptions of Logistic Regression

DEFINITION

Logistic regression assumes: a linear relation between each continuous predictor and the log-odds; independent observations; no severe multicollinearity; and enough events for the number of coefficients, with no predictor that separates the outcome perfectly.

SIMPLIFIED EXPLANATION

Some conditions carry over from linear regression in modified form (recall §12.3): linearity, independence, multicollinearity. Others are specific to a binary outcome: the number of events and separation. Normality of residuals and constant variance do **not** apply. A 0/1 outcome has binomial variance p(1 − p), which the model already builds in.

ILLUSTRATION / EXAMPLE

A model of CBHI membership on distance assumes that the log-odds of membership fall in a straight line with distance. To check it, add a quadratic term (distance^2) and see whether the fit improves; or group distance into categories and see whether the log-odds of each category lie roughly on a line. §13.3 does the first.

MUST-KNOW NOTES (MD-LEVEL)

- **Linearity in the logit:** for each continuous predictor, on the log-odds scale, not the probability scale (which is S-shaped by design).
- **Independence:** broken when observations come in clusters (households in a village) or as repeated measures on one person (recall §5.4). Clustered binary data need the methods of §15.2 (models that allow for correlation within clusters).
- **Multicollinearity:** checked with the VIF, as in linear regression (recall §12.4.4).
- **Events per variable (EPV):** a common convention asks for at least 10 events per estimated coefficient. It comes from a simulation in which no major problems occurred at EPV ≥ 10 (Peduzzi et al., 1996). Later simulations found acceptable bias and coverage below 10 in many settings (Vittinghoff & McCulloch, 2007), and prediction-model work recommends calculating

the required size rather than using a rule of thumb (Riley et al., 2019). Treat EPV < 10 as a warning about **overfitting** (a model that fits the chance quirks of this sample and so predicts poorly in new data), not a bar.

- **Separation:** if a predictor (or combination) perfectly predicts the outcome, the MLE does not exist. The coefficient runs off towards ±∞ and software prints a huge OR with a huge SE (§13.2.6).

CHECKPOINT 13.1

1. A colleague fits a logistic regression and finds one predictor has an OR of 45,000 with an enormous standard error. What is the most likely explanation? Is it an error to fix with a different technique?
2. Explain what "linearity in the logit" means, and why checking it on a plot of raw probability against the predictor would be the wrong approach.
3. Of 20 children screened, 5 are anaemic. Write the likelihood for the prevalence p, find the MLE, and give its logit. What probability corresponds to a logit of 1.2?

(Answers in the Appendix at the end of the book.)

13.2.5 Multivariable Logistic Regression

DEFINITION

Multivariable logistic regression includes two or more predictors at once. Each exponentiated coefficient is an **adjusted OR**: the OR for that predictor, holding the others in the model constant.

SIMPLIFIED EXPLANATION

This is the step from simple to multiple regression (recall §12.4), taken on the logit scale. Adding a confounder to the model adjusts the exposure's OR for it (recall §12.4.1).

Two features are new.

Noncollapsibility. A linear regression coefficient does not change when you add a predictor that is unrelated to the exposure. An OR can. Even with no confounding, the adjusted (conditional) OR can differ from the crude (marginal) one. So a crude-versus-adjusted comparison of ORs mixes confounding with noncollapsibility (Pang et al., 2013). The worked calculation below shows the direction: with a non-confounding covariate, adjustment moves the OR **away from 1**. A shift **towards 1** after adjustment therefore points to confounding, though it does not prove it.

Interaction on two scales. Interaction (effect modification) means the effect of one factor differs by the level of another (recall §12.4.3, §11.7). Logistic and Poisson models test interaction on the **multiplicative** scale: is the joint ratio equal to the product of the separate ratios? Public health often needs the **additive** scale: do the excess risks add? Additive interaction shows where an intervention prevents the most cases. For public-health purposes the additive scale "is thus useful in targeting specific populations and in resource allocation" (Knol & VanderWeele, 2012). The two scales can give different answers on the same data.

The **relative excess risk due to interaction (RERI)** measures additive interaction from relative risks, with the group exposed to neither factor as the reference:

$\text{RERI} = \text{RR}_{11} - \text{RR}_{10} - \text{RR}_{01} + 1.$

RERI = 0 means the excess risks add exactly; RERI > 0 means the joint effect is more than additive.

DERIVATION

RERI. Let R_{00}, R_{10}, R_{01}, R_{11} be the risks with neither factor, only A, only B, and both. If the excess risks add, then $R_{11} - R_{00} = (R_{10} - R_{00}) + (R_{01} - R_{00})$. Divide by R_{00}: $\text{RR}_{11} - 1 = (\text{RR}_{10} - 1) + (\text{RR}_{01} - 1)$. The departure from additivity is $\text{RR}_{11} - 1 - (\text{RR}_{10} - 1) - (\text{RR}_{01} - 1) = \text{RR}_{11} - \text{RR}_{10} - \text{RR}_{01} + 1 = \text{RERI}$. Only relative risks are needed, which is why RERI can be computed from a regression model.

Noncollapsibility. Take two equal-sized strata of a covariate, exposure equally common in each (so the covariate is not a confounder), and an OR of exactly 2 in both. Unexposed risk: 0.1 in stratum 1, 0.5 in stratum 2. Exposed odds are twice the unexposed odds: $2 \times 0.111 = 0.222$ (risk 0.182) and $2 \times 1 = 2$ (risk 0.667). Pooled, the risks are 0.300 (unexposed) and 0.424 (exposed). Crude OR = $(0.424/0.576) \div (0.300/0.700) =$ **1.72**. The adjusted OR is 2.00, further from 1 than the crude 1.72, with no confounding at all.

ILLUSTRATION / EXAMPLE

In a hypothetical 5-year cohort of 4,000 adults, 1,000 in each group, the risk of type 2 diabetes (per 1,000) is 20 with neither obesity nor a family history, 40 with obesity only, 60 with a family history only, and 120 with both (Figure 13.2).

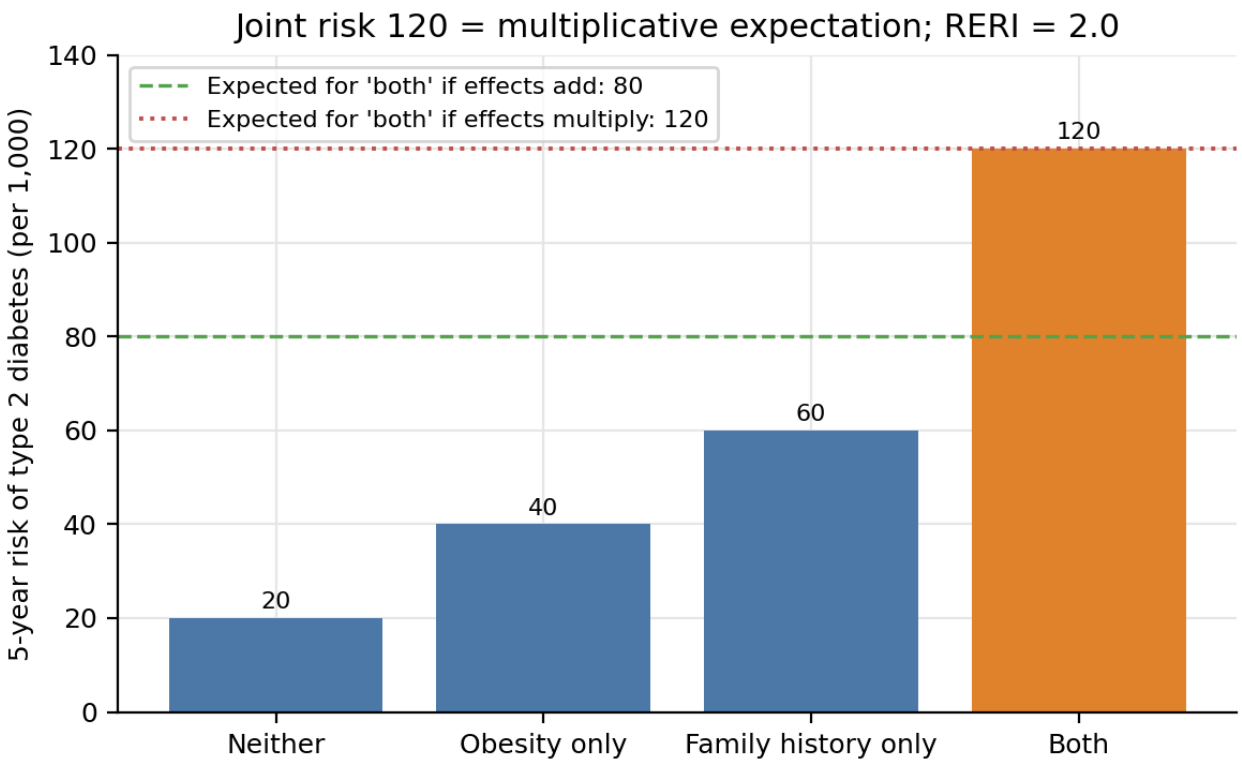


Figure 13.2. Five-year diabetes risk per 1,000 in four groups (hypothetical). If the excess risks added, the "both" group would have 20 + 20 + 40 = 80 per 1,000; if the risk ratios multiplied, 20 × 2 × 3 = 120. The observed 120 shows no multiplicative interaction but a clear additive one.

WORKED CALCULATION

1. Risk ratios against "neither": $RR_{10} = 40/20 = 2$, $RR_{01} = 60/20 = 3$, $RR_{11} = 120/20 = 6$.
2. Multiplicative scale: $RR_{11} \div (RR_{10} \times RR_{01}) = 6 \div 6 = \mathbf{1}$. No interaction.
3. Additive scale: RERI = 6 − 2 − 3 + 1 = **2**. Strong positive interaction.
4. In cases: obesity alone adds 20 cases per 1,000, family history alone adds 40. Together they add 100, which is 40 more than 20 + 40. If these associations are causal, weight-reduction effort among people with a family history prevents the most diabetes per person reached.

IN R (OPTIONAL)

```
d ← data.frame(obese = c(0, 1, 0, 1), famhx = c(0, 0, 1, 1),
               cases = c(20, 40, 60, 120), n = 1000)
m ← glm(cbind(cases, n - cases) ~ obese * famhx,
        family = binomial, data = d)
round(exp(cbind(OR = coef(m), confint.default(m))), 2)
b ← coef(m)
OR10 ← exp(b[2]); OR01 ← exp(b[3]); OR11 ← exp(b[2] + b[3] + b[4])
c(RERI_OR = unname(OR11 - OR10 - OR01 + 1))
```

```
              OR 2.5 % 97.5 %
(Intercept) 0.02  0.01   0.03
obese       2.04  1.18   3.52
famhx       3.13  1.87   5.23
obese:famhx 1.05  0.56   1.97
 RERI_OR
2.512492
```

`obese * famhx` fits both main effects and their product. The product term's OR (1.05, CI 0.56 to 1.97) is the multiplicative-scale interaction; it is close to 1, so a reader who looked only at this line would say "no interaction". RERI computed from the model's ORs is 2.51, larger than the risk-based 2.0 because the outcome reaches 12% in the "both" group. Knol & VanderWeele (2012) warn that "when the outcome is common... using ORs may exaggerate interaction measures"; the risk-ratio models of §13.7 avoid this.

MUST-KNOW NOTES (MD-LEVEL)

- An adjusted OR must be reported with what it was adjusted for.
- **Noncollapsibility:** crude and adjusted ORs can differ without confounding. When the covariate is not a confounder, adjustment tends to move the OR away from 1. A move towards 1 is more suggestive of confounding, but still not proof. Choose confounders on subject-matter grounds (recall §11.6), not by the change they produce.
- **Report interaction on both scales**, with the stratum-specific estimates and a single reference category (Knol & VanderWeele, 2012). A product term in a logistic model tests multiplicative interaction only.
- RERI = $RR_{11} - RR_{10} - RR_{01} + 1$; 0 = additivity. Computed from ORs it approximates the risk-based value only when the outcome is rare.

13.2.6 Reading Logistic Regression Output

DEFINITION

Reading a logistic regression table means answering five questions: what each OR means given the coding, how precise it is, what probability the model predicts for a given person, how well the model performs overall, and what the table does *not* show.

SIMPLIFIED EXPLANATION

Coding first. Every number depends on which outcome category was coded 1 and which level of each categorical predictor is the reference.

Continuous predictors. The OR is the multiplicative change in odds per *one-unit* increase, and it compounds. An OR of 1.05 per year means 10 years multiply the odds by $1.05^{10} = 1.63$, not 1.5.

CI and p-value are built on the log-odds scale, as in §12.2.4, then exponentiated. A 95% CI for an OR that excludes 1 goes with $p < 0.05$; 1 is the no-association value.

Predicted probability. For a given person, add up $b_0 + b_1x_1 + \ldots$ to get the logit, then apply the inverse logit (recall §13.2.2). This is the number a clinician can use.

Model performance. Discrimination asks whether the model gives higher predicted probabilities to people who had the event than to those who did not. **Calibration** asks whether the predicted probabilities are numerically right: among people predicted at 30%, do about 30% have the event? A model can have one without the other. §15.5.1 builds the ROC curve and its area (AUC), and §15.5.2 applies them to a model's predicted probabilities (the c-statistic) together with calibration checks.

Failure modes. Separation (recall §13.2.4) shows up as a huge OR with a huge SE. Do not report that OR. The data are real; the ordinary MLE is the problem. Firth's penalised likelihood gives finite estimates (Heinze & Schemper, 2002); exact logistic regression is another option. **Overfitting** threatens when the number of coefficients is large relative to the number of events (the EPV convention, recall §13.2.4).

ILLUSTRATION / EXAMPLE

A hypothetical study of 400 patients models in-hospital death (1 = died) on age (years), sex (reference: male) and diabetes (reference: no diabetes). The table is invented for teaching, and so is the intercept, $b_0 = -5.00$: an illustrative value, not an estimate from real data. An OR table like this one does not show the intercept; you need it from the full model output.

Predictor	OR	95% CI	p-value
Age (per year)	1.04	1.02 to 1.06	<0.001
Sex: Female	0.78	0.52 to 1.17	0.23
Diabetes: Yes	2.15	1.34 to 3.45	0.002

Each additional year multiplies the odds of death by 1.04, adjusted for sex and diabetes; a 20-year gap multiplies them by $1.04^{20} = 2.19$. Women have lower odds than men, but the CI crosses 1 ($p = 0.23$), so the data are compatible with no difference. Diabetes is associated with about twice the odds of death (OR 2.15), adjusted for age and sex.

WORKED CALCULATION

Predicted probability of death for a 70-year-old woman with diabetes:

1. Convert ORs to coefficients: ln 1.04 = 0.0392 per year; ln 0.78 = −0.248; ln 2.15 = 0.765.
2. Logit (with the illustrative $b_0 = -5.00$) = −5.00 + 70 × 0.0392 + (−0.248) + 0.765 = −5.00 + 2.745 − 0.248 + 0.765 = **−1.738**.
3. p = 1 ÷ (1 + e^(1.738)) = 1 ÷ (1 + 5.69) = **0.15**.

The model predicts a 15% chance of in-hospital death for her. A 70-year-old man without diabetes has logit −5.00 + 2.745 = −2.255 and p = 0.095. Neither number can be read off the OR column alone.

MUST-KNOW NOTES (MD-LEVEL)

- State the event coding and every reference category. An OR without them cannot be interpreted.
- A per-unit OR compounds: OR^k for k units.
- Discrimination and calibration are different properties; §15.5.2 measures both.
- **An OR is not a risk ratio.** When the outcome is common (more than about 10%), the OR overstates the risk ratio if it is above 1 and understates it if below 1 (Zhang & Yu, 1998). §13.7 gives the conversion formula and models that estimate the ratio directly.
- **Reporting checklist:** adjusted OR with 95% CI; outcome coding; reference categories and units; what was adjusted for.

CHECKPOINT 13.2

1. A logistic regression reports an OR of 1.08 per year of age. A colleague says this means a 30-year age gap corresponds to 30 × 1.08 = 32.4-fold higher odds. What is wrong, and what is the correct figure?
2. Two logistic models on the same data report the same OR for a predictor, but one discriminates well between cases and non-cases and the other does not. How is this possible?
3. In a cohort, the 1-year risk of TB is 1% in adults with neither diabetes nor undernutrition, 1.5% with undernutrition only, 2% with diabetes only and 4% with both. Compute the multiplicative interaction ratio and the RERI, and say what each means.

(Answers in the Appendix at the end of the book.)

13.2.7 Ordinal Logistic Regression

DEFINITION

Ordinal logistic regression (the **proportional-odds** or **cumulative-link** model) handles an outcome with three or more ordered categories. It models the odds of being in a higher rather than a lower category, across every cut-point between categories.

SIMPLIFIED EXPLANATION

Ordinal outcomes (recall §1.1.2), such as disease severity or a Likert score, have an order but not equal spacing. With categories mild/moderate/severe there are two cut-points: mild vs moderate-or-severe, and mild-or-moderate vs severe. The model fits one logistic equation per cut-point with a separate intercept for each, but a single common coefficient for each predictor.

ILLUSTRATION / EXAMPLE

Modelling disease severity (mild/moderate/severe) on age and treatment delay: an OR of 1.5 for each week of delay means that each week multiplies the odds of being in a more severe category, at either cut-point, by 1.5.

MUST-KNOW NOTES (MD-LEVEL)

- Collapsing to a binary outcome, or ignoring the order with a multinomial model (§13.2.8), throws information away.
- The **proportional-odds assumption**: each predictor has the same OR at every cut-point. Check it (for example by fitting separate binary models at each cut-point and comparing the ORs); if it fails, a multinomial or partial-proportional-odds model is an alternative.
- Statistical software fits it as an "ordered logistic" (proportional-odds) model. Declare the outcome as ordered first (recall §1.6).

13.2.8 Multinomial Logistic Regression

DEFINITION

Multinomial logistic regression handles an outcome with three or more unordered categories. It models the odds of each category relative to a chosen reference category.

SIMPLIFIED EXPLANATION

Blood group or cause of death has no natural order. With k categories, the model fits k − 1 logistic equations at once, each comparing one category with the reference.

ILLUSTRATION / EXAMPLE

Cause of under-five death (diarrhoeal disease, pneumonia, malnutrition-related, other; reference "other") modelled on household water source and nutritional status gives three sets of coefficients: diarrhoea vs other, pneumonia vs other, malnutrition vs other. Each exponentiated coefficient is an OR (strictly a relative-risk ratio between two categories) against "other".

MUST-KNOW NOTES (MD-LEVEL)

- Choose a substantively meaningful reference, often the largest category, and state it. Changing the reference changes every reported coefficient but not the model's fit.
- Do not use an ordinal model for unordered categories: it imposes an order that does not exist.

- Binary, ordinal and multinomial logistic regression assume independent observations. Individually matched data need conditional logistic regression (§13.2.9).

13.2.9 Conditional Logistic Regression for Matched Data

DEFINITION

Conditional logistic regression fits a logistic model to data in matched sets (a case and its matched controls). It compares members *within* each set, conditioning away each set's own baseline odds, and estimates the exposure OR from within-set contrasts only.

SIMPLIFIED EXPLANATION

In an individually matched case-control study, each case has one control (or m controls) chosen to be similar on, say, age, sex and village. Matching makes controls resemble their cases on the matching factors and, through them, on exposure. The matched sets must therefore be kept in the analysis (Pearce, 2016).

Why not ordinary logistic regression with one dummy variable per matched set? With pairs, there are almost as many dummy variables as people. This is the sparse-data problem. For 1:1 pairs, unconditional logistic regression with pair dummies gives an OR equal to the *square* of the correct conditional OR (Pearce, 2016). Conditional logistic regression removes the set effects from the likelihood instead of estimating them.

Only **discordant** pairs carry information: pairs where exactly one member was exposed. With one binary exposure, the conditional MLE is the matched-pair OR of §10.8 (recall: OR = b/c, the ratio of the two kinds of discordant pair). The regression form adds what §10.8 cannot do: adjustment for further covariates that were not matched on, and 1:m matching.

DERIVATION

Take one matched pair, one case and one control, where exactly one member is exposed. Within the pair, the exposed member has odds ψe^{β} and the unexposed member odds ψ, where ψ is the pair's own baseline odds and e^{β} is the OR. Given that exactly one of the two is a case, the probability that the case is the exposed member is

$$e^{\beta} \div (1 + e^{\beta}),$$

and ψ has cancelled. Concordant pairs (both or neither exposed) give a probability that does not involve β, so they drop out. With b pairs in which the case is exposed and c pairs in which the control is exposed, the conditional likelihood is

$$L(\beta) = [e^{\beta}/(1 + e^{\beta})]^{b} \times [1/(1 + e^{\beta})]^{c}.$$

This is the coin of §13.1 with "heads" = the case is the exposed member: the MLE of $e^{\beta}/(1 + e^{\beta})$ is $b/(b + c)$, which gives $\mathbf{e^{\hat{\beta}} = b/c}$. SE(ln OR) = $\sqrt{(1/b + 1/c)}$.

ILLUSTRATION / EXAMPLE

A district case-control study of infant diarrhoea matched each of 80 cases to one control of the same age and village. Exposure: not exclusively breastfed (not-EBF). Of the 80 pairs, 16 had both exposed, 30 neither, 24 had only the case exposed (b) and 10 only the control exposed (c).

WORKED CALCULATION

1. Matched OR = b/c = 24/10 = **2.40**.
2. SE(ln OR) = √(1/24 + 1/10) = 0.376.
3. 95% CI = exp(ln 2.40 ± 1.96 × 0.376) = **1.15 to 5.02**.
4. Ignoring the matching: pooled, 40 of 80 cases and 26 of 80 controls were exposed, so OR = (40 × 54)/(40 × 26) = 2.08. Breaking the pairs gives a smaller OR, closer to 1, here.

IN R (OPTIONAL)

```
library(survival)
mc ← read.csv("data/matched_casecontrol.csv")
cl ← clogit(case ~ not_ebf + strata(pair_id), data = mc)
summary(cl)
```

```
Call:
coxph(formula = Surv(rep(1, 160L), case) ~ not_ebf + strata(pair_id),
    data = mc, method = "exact")

  n= 160, number of events= 80

          coef exp(coef) se(coef)     z Pr(>|z|)
not_ebf 0.8755    2.4000   0.3764 2.326     0.02 *
...
        exp(coef) exp(-coef) lower .95 upper .95
not_ebf       2.4     0.4167     1.148     5.019
...
Likelihood ratio test= 5.94  on 1 df,   p=0.01
Wald test            = 5.41  on 1 df,   p=0.02
Score (logrank) test = 5.76  on 1 df,   p=0.02
```

`strata(pair_id)` tells `clogit` which rows form a matched set. `coef` is ln OR = 0.8755 = ln(24/10), with SE 0.376 = √(1/24 + 1/10), step 2. `exp(coef)` is the OR of 2.40 with CI 1.15 to 5.02, step 3. The score test statistic (5.76) is McNemar's χ^2 for these pairs (recall §10.8: $(b - c)^2/(b + c) = 196/34 = 5.76$). `clogit` is built on the Cox model of Chapter 14, which is why the call says `coxph`; you do not need survival analysis to use it. Further covariates are added after `not_ebf` in the formula.

MUST-KNOW NOTES (MD-LEVEL)

- **Individually matched sets (pairs, 1:m, siblings, neighbours): conditional logistic regression.** Frequency (group) matching on broad categories such as 5-year age groups: ordinary logistic regression that includes the matching variables (Pearce, 2016; Kuo et al., 2018).
- Matching does not remove confounding by the matching factors in a case-control study. It must still be accounted for in the analysis (Pearce, 2016).
- Only discordant sets contribute. 80 pairs with 34 discordant is effectively a study of 34 pairs.

- The matching variables themselves cannot be studied as risk factors: their effect is conditioned away.

CHECKPOINT 13.3

1. A multinomial model of cause of under-five death uses "other" as the reference. How would the reported results change if "pneumonia" were the reference, and what would stay the same?
2. Why is ordinal logistic regression a mistake for an unordered outcome such as blood group?
3. In a 1:1 matched case-control study of 100 pairs, 30 pairs have only the case exposed and 12 have only the control exposed. Compute the matched OR and its 95% CI. What would an unconditional logistic model with 99 pair dummies report?

(Answers in the Appendix at the end of the book.)

13.3 Analysis Journey – Logistic Regression

This journey introduces no new method. It follows one binary-outcome question from design to a reported odds ratio, using §13.2.

Research question. In a CBHI household survey, is distance to the nearest health facility associated with CBHI membership, independent of household wealth and maternal education?

Study design. Cross-sectional survey of 500 households in 20 villages (25 per village), recording membership, distance (km), a wealth index and maternal education.

Outcome. CBHI membership, coded 1 = member, 0 = non-member. 150 of 500 households (30%) are members.

Exposure and confounders. Distance (continuous, km) is the exposure. Wealth (continuous index) and maternal education (none / primary / secondary-or-higher; reference: none) are confounders, chosen before fitting because each plausibly affects both where a household lives relative to a facility and whether it enrols (recall §11.6).

Why this method? A binary outcome, a mix of continuous and categorical predictors, and a question about an adjusted association: binary logistic regression (§13.2.3).

Data preparation. Education becomes two dummy variables against "none" (recall §12.4.3). Three households are recorded at 0 km; on review this is the code for "facility in the village", so they are kept.

Crude analysis. Membership on distance alone: OR 0.85 per km (95% CI 0.80 to 0.90) (§13.2.3).

Model specification. Membership on distance, wealth and education together. 150 events for 4 coefficients gives an EPV of 37.5 (recall §13.2.4).

Assumptions. Linearity in the logit: adding quadratic terms for distance and wealth does not improve the fit (likelihood-ratio test, which compares the log-likelihoods of the two models, §15.1.4: $p = 0.84$). Separation: members and non-members occur across the whole range of distance and wealth and at every education level. Independence: households are clustered in villages. The ordinary model ignores this, so its SEs are likely too small; §15.2 refits the model with the clustering taken into account.

Diagnostics. Wealth and education are moderately correlated (Spearman ρ = 0.39), but every VIF is below 1.2 (recall §12.4.4.1), so each coefficient can be read on its own. Discrimination: the mean predicted probability is 0.39 among members and 0.26 among non-members, and the c-statistic is 0.72 (the probability that a randomly chosen member gets a higher predicted probability than a randomly chosen non-member; §15.5.2 develops it).

IN R (OPTIONAL)

```
cb ← read.csv("data/cbhi_survey.csv")
cb$maternal_education ← factor(cb$maternal_education,
  levels = c("none", "primary", "secondary+"))
m1 ← glm(cbhi_member ~ distance_km + wealth_index +
           maternal_education,
         family = binomial, data = cb)
round(exp(cbind(OR = coef(m1), confint.default(m1))), 2)
nd ← data.frame(distance_km = 5, wealth_index = 0,
  maternal_education = factor("secondary+",
                              levels(cb$maternal_education)))
predict(m1, nd, type = "response")
```

```
                              OR 2.5 % 97.5 %
(Intercept)                 0.62  0.32   1.19
distance_km                 0.91  0.85   0.97
wealth_index                1.34  1.18   1.53
maternal_educationprimary   1.29  0.76   2.20
maternal_educationsecondary+ 1.87  1.07   3.25
        1
0.4194204
```

Each row after the intercept is an adjusted OR with its Wald 95% CI; these are the numbers in the Output table. The intercept, 0.62, is the odds of membership at 0 km, wealth index 0 and no maternal education ($b_0 = -0.480$). The last line is a predicted probability, worked by hand below.

Output.

Predictor	OR	95% CI	p-value
Distance (per km)	0.91	0.85 to 0.97	0.005
Wealth index (per unit)	1.34	1.18 to 1.53	<0.001
Education: Primary	1.29	0.76 to 2.20	0.35
Education: Secondary+	1.87	1.07 to 3.25	0.027

Worked Calculation (predicted probability)

A household 5 km from the facility, wealth index 0, mother with secondary education:

1. Coefficients: $b_0 = -0.480$; distance −0.094 per km; secondary+ 0.625.
2. Logit = −0.480 + 5 × (−0.094) + 0.625 = **−0.325**.
3. p = 1 ÷ (1 + e^(0.325)) = **0.42**.

Interpretation. Adjusting for wealth and education, each additional kilometre is associated with 9% lower odds of membership (OR 0.91, 95% CI 0.85 to 0.97); 5 km more multiplies the odds by $0.91^5 = 0.62$. The adjusted OR (0.91) is closer to 1 than the crude OR (0.85). Noncollapsibility alone would tend to move it the other way (recall §13.2.5), so the shift suggests that wealth and education

confounded part of the crude association. Households whose mothers have secondary or higher education have 1.87 times the odds of membership of households with no maternal education (95% CI 1.07 to 3.25). The primary-education comparison is imprecise, and its CI includes 1.

What the result does not mean. Membership is common (30%), so these ORs are not risk ratios (recall §13.2.6). The same model fitted as a prevalence-ratio model (§13.7) gives 0.95 per km instead of 0.91, and 1.52 instead of 1.87 for secondary education. Nor does a cross-sectional association show that better roads would raise membership by a predictable amount: distance may stand for unmeasured remoteness (market access, outreach visits).

Reporting. *"In a multivariable logistic regression adjusting for household wealth and maternal education, greater distance to the nearest health facility was associated with lower odds of CBHI membership (adjusted OR 0.91 per km, 95% CI 0.85–0.97, p = 0.005). Secondary or higher maternal education was associated with higher odds of membership than no education (adjusted OR 1.87, 95% CI 1.07–3.25, p = 0.027). Membership was common (30%), so the odds ratios overstate the corresponding prevalence ratios, and the cross-sectional design does not allow causal interpretation."*

MUST-KNOW NOTES (MD-LEVEL)

- Report the crude-to-adjusted change (0.85 → 0.91), but do not use it to choose confounders. Wealth and education were adjusted for because subject-matter reasoning marked them as confounders before the model was fitted; distance is the exposure.
- Outcome coding and reference categories go into the reporting sentence.
- State what the result does not mean in the same paragraph as the result.

13.4 Counts, Person-Time, Rate, and the Offset

DEFINITION

Recall §1.5 and §7.4.1: a **rate** divides a count of events by **person-time** at risk, while an **incidence proportion** (risk) divides by the number of people at risk over a fixed period. An **offset** is a term, log(person-time) or log(population), entered into a Poisson model with its coefficient fixed at 1. It makes the model describe rates instead of raw counts.

SIMPLIFIED EXPLANATION

Two villages report malaria cases this year: Village A had 40 cases among 4,000 people, Village B had 20 among 500. Village B's incidence proportion (20/500 = 4%) is **four times** Village A's (40/4,000 = 1%), despite half the count. A count is comparable across nothing until its denominator is known.

A Poisson model (§13.5) models the log of an expected count. Without an offset it compares raw counts, as if every village had the same population and follow-up. The offset tells the model how much opportunity each unit had.

DERIVATION

For unit i with person-time T_i and rate λ_i, the expected count is $\mu_i = \lambda_i T_i$. Model the log rate linearly: $\log \lambda_i = b_0 + b_1 x_i$. Then

$\log \mu_i = \log T_i + b_0 + b_1 x_i$.

The term $\log T_i$ has no coefficient to estimate (it is fixed at 1). That is the offset. Subtracting it gives $\log(\mu_i/T_i) = b_0 + b_1 x_i$: the model is for the log *rate*, although the outcome fed to the software is the raw count. For a binary x, $e^{(b_1)} = \lambda(x = 1)/\lambda(x = 0)$, the **incidence rate ratio (IRR)**.

Without the offset, $e^{(b_1)}$ is the ratio of expected counts. It equals the rate ratio only if every unit had the same T_i.

ILLUSTRATION / EXAMPLE

Ten villages report malaria cases for one year (`data/malaria_villages.csv`). Five lie near standing water (191 cases among 9,500 people) and five do not (118 cases among 13,500 people). The villages near water are also smaller on average.

WORKED CALCULATION

1. Rate near water = 191/9,500 = 20.1 per 1,000 per year. Rate elsewhere = 118/13,500 = 8.74 per 1,000.
2. IRR = 20.1 ÷ 8.74 = **2.30**. SE(ln IRR) = √(1/191 + 1/118) = 0.117 (recall §7.4.4), so 95% CI = **1.83 to 2.89**.
3. Without the population offset, the model compares mean cases per village: 191/5 = 38.2 against 118/5 = 23.6, a count ratio of **1.62**. Ignoring the populations understates the difference in rates, because the villages near water have fewer people.

MUST-KNOW NOTES (MD-LEVEL)

- **The offset is log(person-time) or log(population), with its coefficient fixed at 1.** With it, exp(b) is a rate ratio; without it, a count ratio. Check which before you read a Poisson coefficient. (§13.5 and §13.6 rely on this and do not repeat it.)
- The unit of person-time sets the unit of the rate, not the rate ratio.

13.5 Poisson Regression

DEFINITION

Poisson regression models a count outcome that follows a Poisson distribution (recall §5.7.1.3) with a log link: log(expected count) = $b_0 + b_1 x_1 + \ldots$ (+ offset). exp(b) is an IRR when a person-time offset is present (recall §13.4).

SIMPLIFIED EXPLANATION

Counts cannot be negative; the log link keeps predicted counts positive, as the logit kept probabilities inside 0 to 1 (recall §13.1). The Poisson distribution has one parameter: its variance equals its mean. That single fact makes Poisson regression simple, and it is also its weak point (§13.6).

ILLUSTRATION / EXAMPLE

The malaria data of §13.4: log(expected cases) = log(population) + b_0 + b_1 × near_water. The number of antenatal visits a woman makes, modelled on distance and wealth, is another Poisson-type outcome (here without an offset, since every woman has one pregnancy).

IN R (OPTIONAL)

```
ml ← read.csv("data/malaria_villages.csv")
p1 ← glm(cases ~ near_water + offset(log(population)),
         family = poisson, data = ml)
summary(p1)$coefficients
exp(cbind(IRR = coef(p1), confint.default(p1)))
# no offset
p0 ← glm(cases ~ near_water, family = poisson, data = ml)
exp(coef(p0))
```

```
             Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -4.7397603 0.09205745 -51.486985 0.000000e+00
near_water   0.8329867 0.11709045   7.114045 1.126908e-12
                   IRR       2.5 %     97.5 %
(Intercept) 0.008740741 0.007297747 0.01046906
near_water  2.300178412 1.828495725 2.89353738
(Intercept)  near_water
  23.600000    1.618644
```

With the offset, exp(intercept) = 0.00874 is the rate away from water (8.74 per 1,000), and exp(b_1) = 2.30 (1.83 to 2.89) is the IRR. The SE of b_1, 0.117, is exactly √(1/191 + 1/118) from §13.4: with one binary predictor, Poisson regression reproduces the hand calculation. Without the offset, the intercept is the mean count per village (23.6) and exp(b_1) = 1.62 is the count ratio.

MUST-KNOW NOTES (MD-LEVEL)

- Poisson regression assumes that events occur independently and that the variance equals the mean. Check both.
- **Does the model apply here?** Malaria is vector-borne. Cases in one village share mosquitoes, breeding sites and each other's parasites, so they are not independent events. Expect the counts to vary more than Poisson allows; §13.6 checks this.
- When the assumption fails, the IRR estimate may still be reasonable, but its SE and CI are too small.

13.6 Negative Binomial Regression

DEFINITION

Overdispersion means the variance of a count exceeds its mean, beyond what the Poisson model allows. **Negative binomial (NB) regression** adds one parameter, $\alpha \geq 0$, so that

$\mathrm{Var}(Y) = \mu + \alpha\mu^2$.

When $\alpha = 0$ this is the Poisson model. Some software, R among it, reports $\theta = 1/\alpha$ instead and writes the variance as $\mu + \mu^2/\theta$ (Zeileis et al., 2008).

SIMPLIFIED EXPLANATION

Counts often cluster more than a pure Poisson process predicts, because units differ in unmeasured ways or because cases cause cases. The Poisson model then treats the data as more informative than they are: SEs too small, CIs too narrow, p-values too small. NB regression lets the data estimate the extra variation. The exponentiated coefficient keeps its meaning (IRR with an offset), but its value and SE are re-estimated.

Checking for overdispersion. A rough check is the **dispersion statistic**: the Pearson χ^2 of the Poisson fit, $\Sigma(\text{observed} - \text{fitted})^2/\text{fitted}$ over the units (recall Pearson's statistic, §10.6), divided by its residual degrees of freedom. Values near 1 are consistent with Poisson variation; values well above 1 suggest overdispersion. The residual deviance (a likelihood-based measure of lack of fit, defined in §15.1.1) ÷ df is used the same way. These are rules of thumb, not tests. A formal comparison is the likelihood-ratio test of the Poisson against the NB model ($\alpha = 0$ against $\alpha > 0$), which compares their log-likelihoods (Zeileis et al., 2008).

Excess zeros. Sometimes there are more zeros than either model expects, because part of the population cannot have the event at all. Cigarettes smoked per day is 0 for every non-smoker, whatever their other characteristics. Two families of models deal with this (Zeileis et al., 2008):

- A **zero-inflated** model is a mixture: a "structural zero" group (always 0) plus a Poisson or NB count group that can also produce zeros. A logistic part models membership of the zero group.
- A **hurdle** model has two parts: a binary model for zero vs any positive count, and a truncated count model (one that allows only counts of 1 or more) for the positive counts.

Both models have two sets of predictors: one for the zeros and one for the counts.

ILLUSTRATION / EXAMPLE

The malaria model of §13.5 has a Pearson χ^2 of 23.7 on 8 residual df: dispersion 2.97. The residual deviance, 24.37 on 8 df, gives 3.05. The village counts vary about three times as much as the Poisson model assumes, as the contagion argument of §13.5 predicted.

WORKED CALCULATION

1. Dispersion statistic = 2.97 (Pearson), 3.05 (deviance). Both well above 1.
2. The NB fit gives $\theta = 18.7$, so $\alpha = 1/18.7 = 0.053$.
3. For a village near water with mean 38.2 cases, Poisson variance = 38.2; NB variance = $38.2 + 0.053 \times 38.2^2$ = **116.2**, about three times as large.

4. Likelihood-ratio test of $\alpha = 0$: LR = 4.67; against χ^2_1, p = 0.031. Because α cannot be negative, $\alpha = 0$ lies on the boundary of its possible values, and the correct null distribution is a 50:50 mixture of a point mass at 0 and χ^2_1 (StataCorp, 2023). The p-value is therefore half: **p = 0.015**.
5. NB IRR = **2.47 (95% CI 1.70 to 3.60)**, against the Poisson 2.30 (1.83 to 2.89). On the log scale the interval is about 1.6 times as wide.

IN R (OPTIONAL)

```
# dispersion statistic
sum(residuals(p1, "pearson")^2) / p1$df.residual
library(MASS)
nb ← glm.nb(cases ~ near_water + offset(log(population)), data = ml)
exp(cbind(IRR = coef(nb), confint.default(nb)))
c(theta = nb$theta, alpha = 1 / nb$theta)
lr ← as.numeric(2 * (logLik(nb) - logLik(p1)))
c(LR = lr, p = pchisq(lr, 1, lower.tail = FALSE))
```

```
[1] 2.968425
                  IRR       2.5 %     97.5 %
(Intercept) 0.0083963 0.006361756 0.01108151
near_water  2.4748383 1.701596933 3.59945683
      theta       alpha
18.70345592  0.05346605
        LR           p
4.67230655 0.03065273
```

The first line is the dispersion statistic of step 1. `glm.nb` returns the NB IRR, 2.47 (1.70 to 3.60), and $\theta = 18.7$ ($\alpha = 0.053$). The LR of 4.67 compares the two log-likelihoods; the printed p (0.031) uses χ^2_1 and is halved to 0.015 for the boundary, step 4.

MUST-KNOW NOTES (MD-LEVEL)

- Poisson: Var = μ. NB: Var = $\mu + \alpha\mu^2$ (some software reports $\theta = 1/\alpha$). Overdispersion ignored → SEs too small.
- Dispersion statistic (Pearson χ^2/df, or deviance/df) is a rough check; the LR test of $\alpha = 0$ is the formal one (halve its χ^2_1 p-value, because $\alpha = 0$ is a boundary).
- Switching to NB changes the coefficient values and SEs, not the meaning of exp(b).
- An alternative is the quasi-Poisson model, which keeps the Poisson estimates and inflates the SEs by the estimated dispersion (Zeileis et al., 2008).
- NB absorbs extra variation; it does not model transmission. With 10 villages, α is poorly estimated, and contagion is a reason to look at designs and models built for infectious disease.
- More zeros than expected, from a group that cannot have the event: consider a zero-inflated or hurdle model.

CHECKPOINT 13.4

1. A Poisson regression of malaria cases on "near standing water", with log(village population) as an offset, gives exp(b) = 2.30. Is this a rate ratio or a count ratio? What would exp(b) be without the offset in these data, and why does it differ?
2. A Poisson model shows clear overdispersion. What does switching to negative binomial regression change, and what does it leave unchanged?
3. A Poisson model of monthly diarrhoea episodes has Pearson $\chi^2 = 40$ on 10 residual df. Compute the dispersion statistic. If an NB fit gives $\alpha = 0.1$, what variance does it imply for a child whose expected count is 2 episodes? For an expected count of 20?

(Answers in the Appendix at the end of the book.)

13.7 Estimating Prevalence and Risk Ratios Directly

DEFINITION

A **risk-ratio (or prevalence-ratio) regression** models a binary outcome with a **log link**: $\log(p) = b_0 + b_1x_1 + \ldots$ Then exp(b) is an adjusted **risk ratio (RR)** in a cohort, or **prevalence ratio (PR)** in a cross-sectional survey. There are two ways to fit it:

- **Log-binomial regression:** binomial distribution, log link.
- **Modified Poisson regression:** Poisson "family", log link, with a **robust (sandwich) standard error** (Zou, 2004).

A **robust (sandwich) standard error** is an SE computed from the observed spread of the residuals instead of from the variance the model assumes. It stays valid when the model's variance assumption is wrong, provided the model for the mean is right.

SIMPLIFIED EXPLANATION

When an outcome is common, the OR exaggerates the RR (recall §13.2.6). The more common the outcome, the more the OR overstates an RR above 1 or understates one below 1 (Zhang & Yu, 1998). CBHI membership (30%), anaemia in women, or full immunisation coverage are all common. A cross-sectional community survey should usually report prevalence ratios.

Log-binomial regression is the natural model, but the log link does not keep p below 1. The fitting algorithm must stay inside the region where every predicted $p < 1$, and it often fails to converge (McNutt et al., 2003).

Modified Poisson regression fits the same log-link model for the mean, using the Poisson likelihood as a working device. A 0/1 outcome is not a Poisson count: its variance is $p(1 - p)$, smaller than the Poisson's p. The coefficients are still valid estimates of log PRs; only the model-based SEs are wrong (too large). The sandwich SE fixes them (Zou, 2004).

A third option is **marginal standardisation**: fit the logistic model, predict every household's probability as if all were exposed, then as if none were, and divide the two averages. This gives an adjusted RR (or, by subtraction, a risk difference) on the probability scale.

DERIVATION

Converting an OR to an RR. Let p_0 be the risk in the reference group and p_1 in the exposed. OR = $[p_1/(1 - p_1)] \div [p_0/(1 - p_0)]$. Solve for p_1 and divide by p_0:

RR = OR ÷ $(1 - p_0 + p_0 \times \text{OR})$.

When p_0 is small, the denominator is close to 1 and RR ≈ OR. This is exact for a crude 2×2 table. Applied to an adjusted OR it is only approximate, and it can mislead in the CI (McNutt et al., 2003); use a model instead.

The sandwich, in the simplest case. Fit a Poisson model with only an intercept to n binary observations with x events. The MLE is $\hat{b}_0 = \ln(x/n) = \ln \hat{p}$. The Poisson model says Var(count) = mean, which gives $SE(\hat{b}_0) = 1/\sqrt{x}$. But the observations are 0/1, so the true variance of $\hat{p}$ is $\hat{p}(1 - \hat{p})/n$ (recall §7.2.2), and on the log scale $SE(\ln \hat{p}) \approx \sqrt{[(1 - \hat{p})/(n\hat{p})]} = \sqrt{[(1 - \hat{p})/x]}$. The ratio of the two is $\sqrt{(1 - \hat{p})}$: the Poisson SE is too big by the factor $1/\sqrt{(1 - \hat{p})}$. The general sandwich SE replaces the model's assumed variance with the squared residuals actually observed, $(y - \hat{\mu})^2$, and "sandwiches" them between two copies of the inverse of the model's information matrix (the curvature of the log-likelihood; the matrix algebra is beyond this book), which gives the method its name.

ILLUSTRATION / EXAMPLE

In the CBHI survey, membership is 44.5% (69/155) among households whose mothers have secondary or higher education, and 18.5% (31/168) with no maternal education.

WORKED CALCULATION

1. Crude PR = 0.445 ÷ 0.185 = **2.41**. Crude OR = (69/86) ÷ (31/137) = **3.55**. With a common outcome, the OR is almost 50% larger than the PR.
2. Adjusted for distance and wealth (§13.3): OR = 1.87. Converted with the formula at $p_0 = 0.185$: 1.87 ÷ (1 − 0.185 + 0.185 × 1.87) = **1.61**. The modified Poisson model gives **PR 1.52** (95% CI 1.03 to 2.25).
3. Sandwich SE for the overall prevalence (150/500): Poisson SE = $1/\sqrt{150} = 0.0816$; robust SE = $\sqrt{(0.70/150)}$ = **0.0683**, smaller by the factor $\sqrt{0.70} = 0.84$.

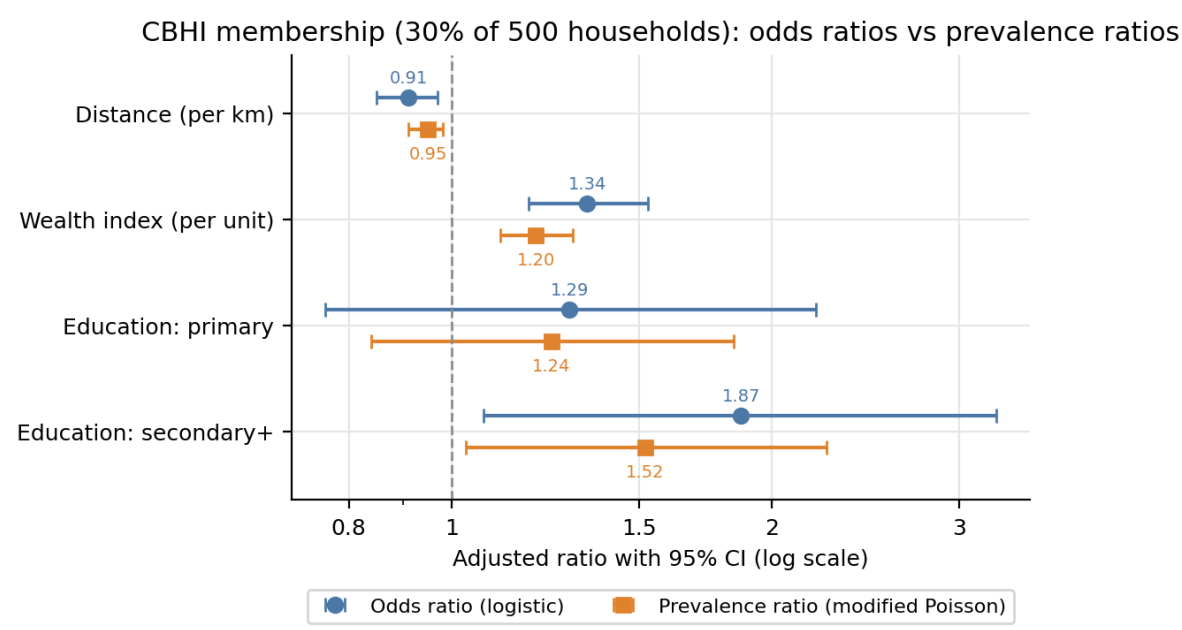


Figure 13.3. The §13.3 model fitted twice to the same 500 households: logistic regression (odds ratios) and modified Poisson regression (prevalence ratios). Every OR lies further from 1 than the matching PR, because 30% of households are members.

IN R (OPTIONAL)

```
library(sandwich); library(lmtest)
# cb as in §13.3; modified Poisson
mp ← glm(cbhi_member ~ distance_km + wealth_index +
           maternal_education, family = poisson, data = cb)
# robust SEs
coeftest(mp, vcov = vcovHC(mp, type = "HC0"))
round(exp(cbind(PR = coef(mp),
                coefci(mp, vcov = vcovHC(mp, type = "HC0")))), 2)
# log-binomial
lb ← glm(cbhi_member ~ distance_km + wealth_index +
           maternal_education, family = binomial(link = "log"),
         data = cb)
```

```
z test of coefficients:

                             Estimate Std. Error z value  Pr(>|z|)
(Intercept)                -1.097410   0.220379 -4.9796 6.370e-07 ***
distance_km                -0.054573   0.019710 -2.7688  0.005627 **
wealth_index                0.184159   0.039432  4.6703 3.007e-06 ***
maternal_educationprimary   0.215789   0.200784  1.0747  0.282495
maternal_educationsecondary+  0.418499   0.199397  2.0988  0.035833 *
...
                              PR 2.5 % 97.5 %
(Intercept)                 0.33  0.22   0.51
distance_km                 0.95  0.91   0.98
wealth_index                1.20  1.11   1.30
maternal_educationprimary   1.24  0.84   1.84
maternal_educationsecondary+ 1.52  1.03   2.25
Error: no valid set of coefficients has been found: please supply starting values
```

`vcovHC(type = "HC0")` computes the sandwich variance; `coeftest` and `coefci` use it for the SEs, z-tests and CIs. Exponentiated, the coefficients are adjusted PRs: 0.95 per km (0.91 to 0.98), 1.20 per

wealth unit, and 1.52 for secondary education (1.03 to 2.25). These are the orange points in Figure 13.3. The log-binomial model fails outright on these data. Supplied with starting values (`start = c(log(0.3), 0, 0, 0, 0)`), it runs but stops with "algorithm did not converge", which is the practical problem McNutt et al. (2003) describe.

Marginal standardisation from the §13.3 logistic model:

```
m1 ← glm(cbhi_member ~ distance_km + wealth_index +
           maternal_education, family = binomial, data = cb)
lev ← levels(cb$maternal_education)
d_sec ← transform(cb, maternal_education = factor("secondary+", lev))
d_non ← transform(cb, maternal_education = factor("none", lev))
p_sec ← mean(predict(m1, d_sec, type = "response"))
p_non ← mean(predict(m1, d_non, type = "response"))
round(c(p_secondary = p_sec, p_none = p_non,
        RR = p_sec / p_non, RD = p_sec - p_non), 3)
```

```
p_secondary      p_none          RR          RD
      0.361       0.243       1.488       0.118
```

If every household had a mother with secondary education, the model predicts 36.1% membership; if none did, 24.3%. The standardised RR is 1.49, close to the modified-Poisson 1.52, and the risk difference is 11.8 percentage points.

MUST-KNOW NOTES (MD-LEVEL)

- **Common outcome (more than about 10%) → report RR or PR, not only the OR.** Cross-sectional → prevalence ratio; cohort → risk ratio.
- **Modified Poisson = Poisson family + log link + robust SE** (Zou, 2004). Do not report its model-based SEs for binary data: they are too large.
- Modified Poisson can predict probabilities above 1 for extreme households (one of 500 here, at 1.05). This does not bias the PRs much, but do not use the model to predict individual probabilities.
- Log-binomial is the "proper" likelihood model but often fails to converge.
- The OR → RR formula is exact only for a crude 2×2 table; for adjusted ORs use a model or marginal standardisation.
- The robust SE here corrects the variance assumption, not clustering. Households in the same village are still correlated; §15.2.5 uses a clustered version of the same sandwich idea (GEE).

CHECKPOINT 13.5

1. In a cohort, 40% of unexposed people develop the outcome, and a crude OR is 2.0. Convert it to an RR. What would the RR be if the unexposed risk were 2%?
2. Why does a Poisson model with a log link give valid prevalence ratios for a 0/1 outcome, and why must its SEs be replaced by robust ones?
3. A colleague fits a log-binomial model and gets "algorithm did not converge". Name two alternatives that still give an adjusted prevalence ratio.

(Answers in the Appendix at the end of the book.)

Survival Analysis

Survival analysis studies time-to-event data: how long it takes for an event (death, relapse, recovery, dropout) to happen, when some subjects have not yet had the event by the time their follow-up ends. This chapter first shows why ordinary regression fails on such data (§14.1–14.3), then builds the survival curve (§14.4), the test that compares curves (§14.5), and the regression model for the hazard (§14.6). A worked analysis of a CBHI dropout cohort (§14.7) and the problem of competing events (§14.8) close it. Any research question phrased as "time until X", such as time to treatment default or time to scheme dropout, belongs here.

14.1 Why Ordinary Regression Fails for Time-to-Event Data

DEFINITION

Time-to-event data cannot be analysed by ordinary linear regression on the time, or logistic regression on whether the event happened, when some subjects are **censored**: their follow-up ends before the event, so their true event time is known only to exceed their observed time.

SIMPLIFIED EXPLANATION

A "time until dropout" looks like a continuous outcome for linear regression. The problem is the subjects who have not dropped out when the study ends. Their true time is longer than what you saw, but you do not know by how much. Excluding them wastes real information and biases the sample towards early events. Treating their last observed time as their event time shortens every one of their times. Neither is valid.

ILLUSTRATION / EXAMPLE

Take a conceptual study (not the §14.7 cohort) that follows 100 CBHI members for 3 years. Some drop out during the study; their dropout time is known. Others are still enrolled at 3 years. For them, "3 years" is not a dropout time: many would have stayed enrolled much longer had the study continued. Entering 3 years as their outcome in a regression would understate how long members typically stay.

MUST-KNOW NOTES (MD-LEVEL)

- Dropping censored subjects is a selection problem: those left in the analysis are no longer like the cohort that started (§17.2.1 treats selection bias). A censored subject's event-free time up to the last contact is genuine, usable data.
- Treating a censoring time as an event time biases the average time to event downwards.

- Survival methods use each subject's data exactly up to the point where it is known, and no further.
- Poisson regression with log person-time as the offset (recall §13.4) also uses censored follow-up correctly, but it assumes a constant rate over the follow-up. Survival methods drop that assumption.

14.2 Censoring

DEFINITION

Censoring occurs when a subject's exact event time is not known. In **right censoring**, follow-up ends before the event: the study ends, or the subject is lost to follow-up. The true event time is known only to be greater than the observed time.

SIMPLIFIED EXPLANATION

A censored subject gives you partial information: event-free for at least this long. This differs from ordinary missing data (a value absent altogether, §15.4), because the censored value has a known lower bound. Right censoring is the form this book uses. Left censoring (the event happened before observation began, at an unknown time) and interval censoring (the event is known only to lie between two visits) need other methods.

Two features of real follow-up show in Figure 14.1. First, subjects enter at different calendar times (**staggered entry**). Analysis uses time since each subject's own start, so every line is shifted back to zero (panel b). A late entrant is therefore censored earlier on the analysis scale even though the study ended on the same calendar day for everyone. Second, censoring has two routes: **administrative censoring** at the planned end of the study, and **loss to follow-up** partway through.

ILLUSTRATION / EXAMPLE

In the conceptual 3-year CBHI study (§14.1), subjects fall into three groups: members who drop out during the study (exact time known); members still enrolled at study end (administratively censored); and members who move away and cannot be traced (censored at the last date they were known to be enrolled). In Figure 14.1, Household 4 moves away at 14 months. Households 5 and 6 join 10 and 18 months into the study, so the study's end censors them at 26 and 18 months of follow-up.

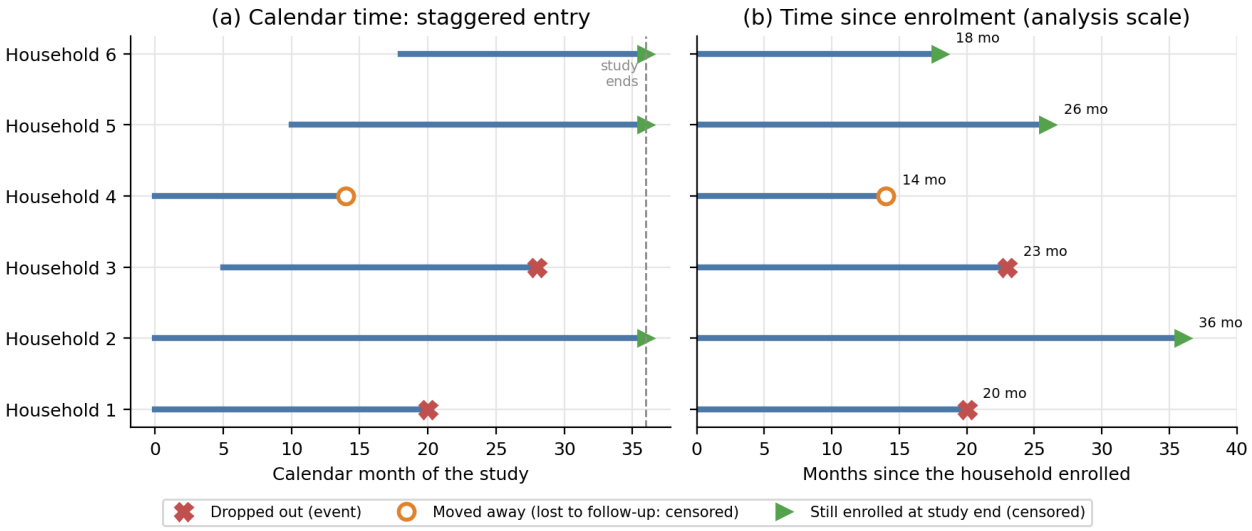


Figure 14.1. Six households. (a) Calendar time: households join the scheme at different times, and the study ends at month 36 for all. (b) The same follow-up re-aligned to time since enrolment, the scale survival analysis uses. Crosses are dropouts; the open circle is a household lost to follow-up; arrows are households still enrolled at the end.

MUST-KNOW NOTES (MD-LEVEL)

- Standard methods assume **non-informative censoring**: those censored "should be as likely to have a subsequent event as those individuals who remain in the study" (Clark et al., 2003). It fails if, say, households on the verge of leaving the scheme are also the ones who stop answering the door.
- A **competing event** is not ordinary censoring. An event that makes the outcome impossible (death, when the outcome is TB culture conversion) must not be treated as censored when you estimate the probability of the outcome; §14.8 shows why and what to do instead.
- Moving away counts as censoring only if the household could still have dropped out afterwards, unseen. If moving ends membership by another route, it is a competing event (§14.8 revisits this).

14.3 Survival Function, Hazard, and the At-Risk Set

DEFINITION

The **survival function**, S(t), is the probability of remaining event-free beyond time t. The **hazard**, h(t), is the instantaneous *rate* at which the event occurs at time t among those still event-free just before t. The **at-risk set** at time t is exactly that group: everyone who has neither had the event nor been censored before t.

SIMPLIFIED EXPLANATION

Two questions can be asked of time-to-event data. The first: what fraction is still event-free by time t? That is S(t), a probability between 0 and 1 that can only fall as t increases. It describes the *original* group, cumulatively.

The second: of those still event-free right now, how fast are they having the event? That is h(t). **Hazard is a rate, not a probability** (recall the rate as events per unit of person-time, §7.4.1 and §13.4). It has a time unit and is not bounded by 1: 3 events per person-year is a sensible rate for a repeatable event, while a probability of 3 is nonsense. It is also conditional on having survived this far.

That conditioning is why hazard is not risk. Risk ($1 - S(t)$) is what has happened to the original group by time t. Hazard is what is happening to the survivors now. A hazard can keep rising (mortality climbing with age) when few of the original cohort remain to experience it. Using "hazard" and "risk" as synonyms is an error you will meet in survival papers.

The at-risk set is the bookkeeping behind both. A censored subject leaves it at the moment of censoring, having contributed all their event-free time. Unlike a missing value, which contributes nothing, a censored subject contributes real information up to the last contact.

ILLUSTRATION / EXAMPLE

At month 12 of the conceptual 100-member study (§14.1), suppose 70 members are still enrolled and observed, 25 have dropped out and 5 have been censored earlier. The at-risk set is 70, not 100 and not 95. If 3 members drop out that month, the dropout rate for the month is $3/70 = 0.043$ per member-month, measured among those who could actually drop out. The survival curve takes the same step: S(t) is multiplied by $1 - 3/70 = 67/70 = 0.957$. S(12) itself is the running product of such steps for every month so far, which is exactly what the Kaplan–Meier method in §14.4 computes.

MUST-KNOW NOTES (MD-LEVEL)

- **S(t):** a probability, cumulative from the start, never increasing.
- **h(t):** a rate among those still event-free at t; not a probability and not bounded by 1.
- **Risk and hazard are not interchangeable.** Risk describes the original cohort; hazard describes the current survivors.
- **At-risk set at t:** no event and not censored before t. Both S(t) and h(t) are computed from how this set shrinks.

14.4 Kaplan-Meier Survival Curve

DEFINITION

The **Kaplan–Meier (product-limit) estimate** of the survival function is

$S(t) = \Pi$ over event times $t_j \le t$ of $(1 - d_j/n_j)$,

where d_j is the number of events at time t_j and n_j the number at risk just before t_j. Plotted against time, it is a step function that starts at 1 and drops only at event times.

SIMPLIFIED EXPLANATION

The curve is built one event time at a time. At each, ask: of the n_j still at risk, what fraction got through? Multiply these fractions together. A censored subject never causes a step. They simply leave the at-risk set, so later fractions have a smaller denominator. This is the mechanism that lets censored subjects count up to their last contact and no further.

The demographic life table does the same job for a population (recall §7.5, where $l_{x+n} = l_x \times (1 - {}_nq_x)$). Kaplan–Meier builds the survival curve from individual follow-up times instead of age bands.

DERIVATION

To be event-free beyond t_j you must be event-free beyond t_{j-1} *and then* get past t_j. By the multiplication rule for conditional probability (recall §5.3):

$P(T > t_j) = P(T > t_{j-1}) \times P(T > t_j \mid T > t_{j-1})$.

The conditional factor is estimated by the fraction of the at-risk set that did not have the event: $(n_j - d_j)/n_j = 1 - d_j/n_j$. Between event times nobody has the event, so the factor is 1 and the curve stays flat. Chaining the factors from the start gives the product.

Greenwood's standard error. Write $\log S(t) = \Sigma \log(1 - q_j)$, with $q_j = d_j/n_j$, a proportion out of n_j with variance $q_j(1 - q_j)/n_j$. The delta method (a first-order Taylor approximation, $\mathrm{Var}[g(X)] \approx g'(\mu)^2 \mathrm{Var}(X)$, where g' is the slope, or derivative, of g; the calculus is beyond this book) gives $\mathrm{Var}[\log(1 - q_j)] \approx q_j/[n_j(1 - q_j)] = d_j/[n_j(n_j - d_j)]$. Adding the terms, and applying the delta method once more to go from $\log S$ to S:

$\mathrm{Var}[S(t)] \approx S(t)^2 \times \Sigma\, d_j/[n_j(n_j - d_j)]$.

WORKED CALCULATION

`data/km_eight.csv` holds eight CBHI members (months to dropout; + = censored): 2, 4, 4+, 6, 7+, 9, 11, 12+.

Month t_j	At risk n_j	Dropouts d_j	Censored after t_j	$1 - d_j/n_j$	$S(t_j)$	$\Sigma\, d_j/[n_j(n_j - d_j)]$	SE of $S(t_j)$
2	8	1	0	7/8 = 0.875	0.875	0.0179	0.117
4	7	1	1 (4+)	6/7 = 0.857	0.750	0.0417	0.153
6	5	1	1 (7+)	4/5 = 0.800	0.600	0.0917	0.182
9	3	1	0	2/3 = 0.667	0.400	0.2583	0.203
11	2	1	1 (12+)	1/2 = 0.500	0.200	0.7583	0.174

1. Month 2: all 8 at risk, 1 drops out. $S = 7/8 = 0.875$.
2. Month 4: 7 at risk (the tie 4 and 4+ is broken the usual way: the censored member counts as at risk at 4). $S = 0.875 \times 6/7 = 0.750$. Member M3 (4+) then leaves.
3. Month 6: 5 at risk. $S = 0.750 \times 4/5 = 0.600$. M5 is censored at 7 and leaves without a step.
4. Month 9: 3 at risk. $S = 0.600 \times 2/3 = 0.400$. Month 11: $S = 0.400 \times 1/2 = 0.200$. The curve stays at 0.200 to 12 months, when the last member is censored.
5. Greenwood at 9 months: $1/(8 \times 7) + 1/(7 \times 6) + 1/(5 \times 4) + 1/(3 \times 2) = 0.0179 + 0.0238 + 0.0500 + 0.1667 = 0.2583$. SE $= 0.400 \times \sqrt{0.2583} = 0.400 \times 0.508 =$ **0.203**. A plain 95% CI, $0.400 \pm 1.96 \times 0.203$, runs from 0.002 to 0.798. With eight members it is wide, and near 0 or 1 a plain interval can cross the limits; software therefore builds the interval on the log or log(−log) scale and back-transforms it.

6. **Median survival** is the first time S(t) falls to 0.5 or below. S is 0.600 up to 9 months and 0.400 from 9 months, so the median is **9 months**.

If the curve never falls to 0.5 within follow-up, the median is **not reached**. Report it as "NR" and give S(t) at fixed times instead. Never read a median off an extrapolated curve.

ILLUSTRATION / EXAMPLE

In the §14.7 cohort of 300 CBHI households followed for up to 24 months, 120 dropped out, 18 moved away and 162 were still enrolled at 24 months. The Kaplan–Meier retention is S(12) = 0.75 and S(24) = 0.58 (Figure 14.2b). The curve never reaches 0.5, so the median time to dropout is **not reached**: more than half the households were still enrolled when follow-up ended. The summary is "median NR; 58% still enrolled at 24 months".

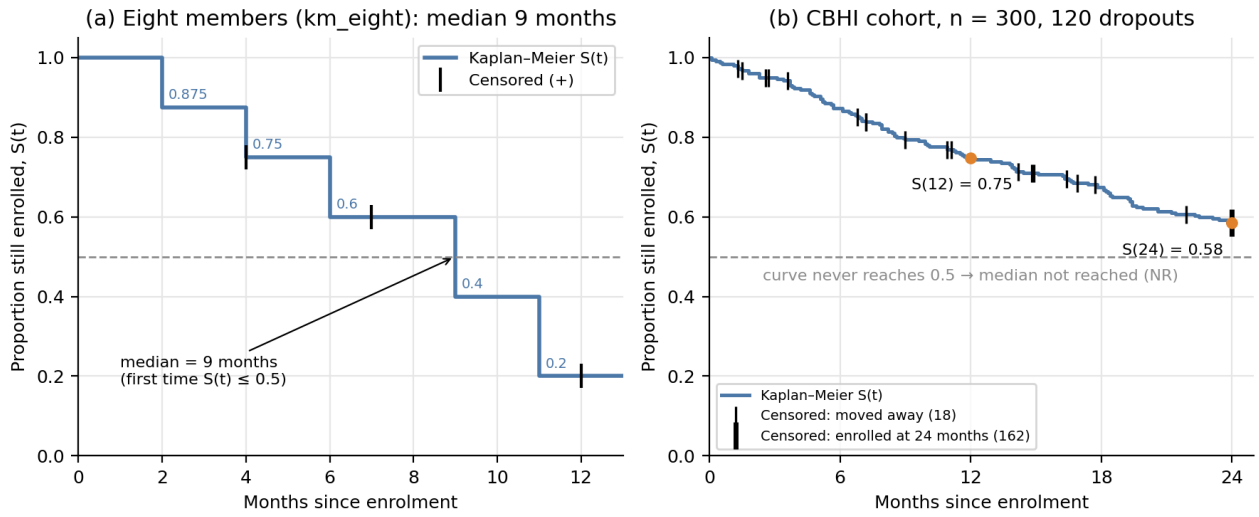


Figure 14.2. (a) The eight-member table above: steps at each dropout, ticks at censored times, median 9 months. (b) The 300-household cohort: short ticks are the 18 households that moved away; the thick tick at 24 months stands for the 162 still enrolled. The curve ends at 0.58, so the median is not reached.

The actuarial (clinical life-table) method. Before computers, follow-up was grouped into fixed intervals, as in the demographic life table (§7.5). Within each interval, the number at risk is adjusted as $n' = n - w/2$, where w is the number withdrawn (censored) in the interval: censored subjects contribute "half to the number of patients at risk" (Held et al., 2026), on the assumption that they left, on average, halfway through. The interval's probability of the event is $q = d/n'$, and survival is the running product of $(1 - q)$.

Interval (months)	At start n	Dropouts d	Withdrawn w	n′ = n − w/2	q = d/n′	1 − q	S at end
0–3	8	1	0	8.0	0.125	0.875	0.875
3–6	7	1	1	6.5	0.154	0.846	0.740
6–9	5	1	1	4.5	0.222	0.778	0.576
9–12	3	2	0	3.0	0.667	0.333	0.192

Kaplan–Meier gives 0.750, 0.600 and 0.200 just before 6, 9 and 12 months; the actuarial table gives 0.740, 0.576 and 0.192. The actuarial values are slightly lower because each withdrawal is counted as only half at risk. Kaplan–Meier uses exact times and needs no assumption about when within an interval people

left. The actuarial method suits large cohorts recorded only by interval, such as routine programme registers.

IN R (OPTIONAL)

```
library(survival)
k8 ← read.csv("data/km_eight.csv")
f8 ← survfit(Surv(time_months, dropout) ~ 1, data = k8)
summary(f8)
f8
co ← read.csv("data/cohort_dropout.csv")
fa ← survfit(Surv(time_months, dropout) ~ 1, data = co)
summary(fa, times = c(12, 24))
fa
```

```
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2      8       1    0.875   0.117       0.6734            1
    4      7       1    0.750   0.153       0.5027            1
    6      5       1    0.600   0.182       0.3315            1
    9      3       1    0.400   0.203       0.1477            1
   11      2       1    0.200   0.174       0.0363            1

     n events median 0.95LCL 0.95UCL
[1,] 8      5      9       6      NA
...
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   12    215      74    0.748  0.0253         0.70        0.800
   24    162      46    0.585  0.0291         0.53        0.645

       n events median 0.95LCL 0.95UCL
[1,] 300    120     NA      NA      NA
```

`Surv(time, event)` pairs each time with its event indicator. For the eight members, `n.risk`, `survival` and `std.err` reproduce the hand table (8, 7, 5, 3, 2; 0.875 to 0.200; Greenwood SE 0.203 at 9 months), and the median is 9. The confidence limits differ from the plain ones because R builds them on the log scale. For the cohort, `n.event` counts dropouts since the previous listed time (74 by 12 months, 46 more by 24); S(12) = 0.75 and S(24) = 0.58; and `median` is `NA`, R's way of saying not reached.

MUST-KNOW NOTES (MD-LEVEL)

- $S(t) = \Pi(1 - d_j/n_j)$. The curve is a step function, flat between events.
- Censored subjects leave the at-risk set without causing a step. Show them as tick marks on the curve.
- The median is the first time $S(t) \le 0.5$. If the curve stays above 0.5, the median is not reached (NR); report S(t) at stated times instead.
- The estimate is unreliable in the tail, where few remain at risk. Greenwood SEs widen there (0.203 at 9 months with 3 at risk); a common practice is to show the number at risk beneath the plot.
- Actuarial method: fixed intervals, withdrawals count as half at risk. Kaplan–Meier: exact event times.

CHECKPOINT 14.1

1. In a cohort, S(t) just before month 12 is 0.80. At month 12, 70 members are at risk and 3 drop out. What is S(12)? Is the hazard in that month 3/70 or 3/100 of the original cohort, and why?
2. In a two-year follow-up of TB patients, the Kaplan–Meier curve ends at 0.64. A colleague reports "median survival 24 months". What should be reported instead?
3. Why is a member who moves away at month 5 not simply deleted from a Kaplan–Meier analysis, and what assumption makes it safe to treat them as censored?

(Answers in the Appendix at the end of the book.)

14.5 Log-Rank Test

DEFINITION

The **log-rank test** tests H_0: two (or more) groups have the same survival experience over the whole follow-up. At each event time it compares the observed number of events in each group (O) with the number expected (E) if H_0 were true, sums over event times, and forms

$\chi^2 = \Sigma$ over groups $(O - E)^2/E$, on (number of groups − 1) degrees of freedom.

SIMPLIFIED EXPLANATION

Kaplan–Meier curves for two groups let you *see* a difference. The log-rank test asks whether it is larger than chance would produce. At every moment someone has the event, it asks: given who was at risk in each group right then, how many of these events should each group have had if the groups were alike? Too many events in one group, time after time, adds up to a large χ^2.

DERIVATION

At event time t_j, n_j subjects are at risk, n_{0j} of them in group 0, and d_j events occur. Under H_0 each at-risk subject is equally likely to be among the d_j, so the expected number in group 0 is

$E_{0j} = d_j \times n_{0j}/n_j$.

Sum over all event times: $E_0 = \Sigma E_{0j}$; O_0 is the total observed in group 0. For two groups, E_1 = total events − E_0. The statistic $\Sigma(O - E)^2/E$ has the same form as the goodness-of-fit χ^2 (recall §10.6.2), with df = 2 − 1 = 1 (Bland & Altman, 2004).

Software uses a slightly more exact version. The count in group 0 at t_j follows a hypergeometric distribution with variance $V_j = d_j(n_{0j}/n_j)(n_{1j}/n_j)(n_j - d_j)/(n_j - 1)$. With $V = \Sigma V_j$, the statistic is $\chi^2 = (O_0 - E_0)^2/V$, also on 1 df. The two versions are usually close.

WORKED CALCULATION

The §14.7 cohort: 147 households without the community education intervention, 153 with it; 120 dropouts at 86 distinct event times. The first three event times:

Month t_j	At risk, no intervention n_{0j}	At risk, intervention n_{1j}	Dropouts d_j (no int + int)	$E_{0j} = d_j \times n_{0j}/n_j$	$O_{0j} - E_{0j}$
0.1	147	153	2 (1 + 1)	2 × 147/300 = 0.980	+0.020
0.3	146	152	1 (0 + 1)	1 × 146/298 = 0.490	−0.490
0.5	146	151	1 (1 + 0)	1 × 146/297 = 0.492	+0.508
…	…	…	…	…	…

1. Continuing through all 86 event times: no intervention $O_0 = 70$, $E_0 = 55.95$; intervention $O_1 = 50$, $E_1 = 120 - 55.95 = 64.05$.
2. The no-intervention group had 14.05 more dropouts than expected; the intervention group 14.05 fewer.
3. $\chi^2 = 14.05^2/55.95 + 14.05^2/64.05 = 3.53 + 3.08 =$ **6.61**, df = 1, p = 0.010.
4. Variance version: V = 29.75, $\chi^2 = 14.05^2/29.75 =$ **6.63**, p = 0.010.

The log-rank test shows the groups differ (p = 0.01). It does not say by how much: that needs a hazard ratio (§14.6.1).

ILLUSTRATION / EXAMPLE

In the §14.7 cohort, 0.66 of intervention households and 0.51 of the others were still enrolled at 24 months (the curves are Figure 14.4 in §14.7). The worked calculation above shows that a gap this size would be unlikely if the groups truly had the same survival (p = 0.01).

IN R (OPTIONAL)

```
library(survival)
co ← read.csv("data/cohort_dropout.csv")
survdiff(Surv(time_months, dropout) ~ intervention, data = co)
```

```
                N Observed Expected (O-E)^2/E (O-E)^2/V
intervention=0 147       70       56      3.53      6.63
intervention=1 153       50       64      3.08      6.63

 Chisq= 6.6  on 1 degrees of freedom, p= 0.01
```

`Observed` and `Expected` are O and E from the worked calculation (E rounded to 56 and 64). The `(O-E)^2/E` column gives the two terms of the hand formula, 3.53 + 3.08 = 6.61. The `(O-E)^2/V` column is the variance version, 6.63, which is the `Chisq` R reports, with p = 0.01.

MUST-KNOW NOTES (MD-LEVEL)

- H_0: identical survival across the whole follow-up. Like ANOVA's F (recall §9.5), it is an overall test.
- It weights every event time equally. It "is most likely to detect a difference between groups when the risk of an event is consistently greater for one group than another" and is unlikely to do so "when survival curves cross" (Bland & Altman, 2004). Look at the curves as well as the p-value.

- It gives a p-value only, no effect size and no confidence interval. It cannot adjust for covariates; Cox regression (§14.6) does both.
- It plays the role of the t-test or Mann–Whitney test (recall §9.2.2 and §10.2) for censored time-to-event data, which neither of those can handle.

CHECKPOINT 14.2

1. A Kaplan–Meier curve for Group A lies above Group B's throughout follow-up, and the log-rank test gives p = 0.03. A colleague concludes: "Group A has a lower hazard at every single time point, guaranteed." What is the overreach?
2. A two-group log-rank analysis gives O = 30, E = 24 in Group A and O = 20, E = 26 in Group B. Compute $\Sigma(O - E)^2/E$, its df and approximate p-value.
3. Why can the log-rank test not tell you whether the intervention's association with dropout remains after adjusting for household income?

(Answers in the Appendix at the end of the book.)

14.6 Cox Proportional Hazards Regression

DEFINITION

Cox proportional hazards regression models the hazard (recall §14.3) as a baseline hazard multiplied by an exponential function of the predictors:

$h(t \mid x_1, \ldots, x_k) = h_0(t) \times \exp(b_1x_1 + b_2x_2 + \ldots + b_kx_k)$.

The baseline hazard $h_0(t)$, the hazard when every x is 0, is left unspecified.

SIMPLIFIED EXPLANATION

The log-rank test compares groups but cannot adjust for several predictors at once, as multiple linear or logistic regression can (recall §12.4 and §13.2.5). Cox regression fills that gap for censored time-to-event data. Each predictor multiplies the hazard by a fixed factor, exp(b). The shape of $h_0(t)$ over time can be anything: rising, falling or irregular. That is why the model is called **semi-parametric**: the predictor part is a parametric formula, the time part is not.

The coefficients are estimated by maximising a *partial likelihood* (likelihood recalled from §13.1). At each event time it asks: given that one of the at-risk subjects had the event now, how probable was it to be this one, given everyone's x values? $h_0(t)$ cancels from that ratio, so it never needs to be estimated (the full theory is beyond this book).

ILLUSTRATION / EXAMPLE

To model time to CBHI dropout on household income, distance to the nearest facility and receipt of the community education intervention, Cox regression estimates each predictor's association with the dropout **hazard**, adjusted for the others, as multiple regression adjusts for confounders of a continuous outcome.

MUST-KNOW NOTES (MD-LEVEL)

- exp(b) is a hazard ratio (§14.6.1): the multiplicative change in the hazard per one-unit increase in a predictor, others held constant.
- The model relies on proportional hazards (§14.6.2): each HR is constant over follow-up. Check it.
- Events, not subjects, limit the model. A common convention is at least 10 events per predictor: in simulations, "the value of EPV = 10 seems most prudent" (Peduzzi et al., 1995). As for logistic regression (recall §13.2.4), the rule is a guide; smaller ratios can work, larger ones are safer.

14.6.1 Hazard Ratio

DEFINITION

A **hazard ratio (HR)** is the ratio of the hazard in one group (or at one predictor level) to that in another. In Cox regression it is HR = exp(b).

SIMPLIFIED EXPLANATION

Exponentiating a coefficient gives an odds ratio in logistic regression and a rate ratio in Poisson regression (recall §13.2.6 and §13.5). In Cox regression it gives a hazard ratio: how many times higher or lower the instantaneous event rate is in one group than in another, at any moment, among those still event-free, provided proportional hazards holds.

ILLUSTRATION / EXAMPLE

The §14.7 cohort's adjusted HR for the intervention is 0.58. Inverted, households *without* the intervention have 1/0.58 = **1.72** times the dropout hazard of households with it, at any moment during follow-up, holding income and distance constant.

COMMON MISREADING

"HR 0.58 means 42% lower risk of dropout." It means 42% lower *hazard*: 1 − 0.58 = 0.42, a statement about the instantaneous rate among households still enrolled. The cumulative risk comparison is a different number. In the cohort, the crude HR is 0.62, a 38% lower hazard. The crude cumulative dropout by 24 months (1 − S(24), §14.4) is 1 − 0.661 = 0.339 with the intervention and 1 − 0.505 = 0.495 without, a risk ratio of 0.339/0.495 = **0.69** (0.686 from unrounded survival), a 31% lower risk. The HR and the risk ratio differ because cumulative risk depends on the baseline hazard as well as on the HR.

MUST-KNOW NOTES (MD-LEVEL)

- HR = e^b. HR = 1: no difference in hazard; HR > 1: higher hazard (shorter event-free times); HR < 1: lower hazard.

- **An HR is not a risk ratio.** A risk ratio compares cumulative probabilities by a fixed time. A constant HR does not mean the risk ratio, or the risk difference, stays constant over follow-up; those change with the baseline hazard.

14.6.2 Proportional Hazards Assumption

DEFINITION

The **proportional hazards (PH) assumption** states that the hazard ratio between any two groups or predictor levels is constant over follow-up. The hazards themselves may rise or fall; their ratio may not change.

SIMPLIFIED EXPLANATION

Cox regression can leave $h_0(t)$ unspecified only because every subject's hazard is the same curve $h_0(t)$ scaled by a fixed factor. If the intervention helps early but not later, no single factor describes it, and the one HR the model reports is a weighted average of different effects.

DERIVATION

Why log(−log S) curves are parallel. The cumulative hazard is $H(t) = -\log S(t)$ (the total hazard accumulated from 0 to t). Under PH, $H_1(t) = HR \times H_0(t)$. Taking logs:

$\log H_1(t) = \log HR + \log H_0(t)$, that is, $\log(-\log S_1(t)) = \log(-\log S_0(t)) + \log HR$.

So plots of $\log(-\log S(t))$ against $\log t$ for the groups differ by a constant vertical distance, $\log HR$, at every t. Curves that converge, diverge or cross signal non-proportional hazards.

Schoenfeld residuals. At each event time, take the covariate value of the subject who had the event and subtract the average covariate value of the at-risk set, weighted by each member's fitted hazard. That difference is the **Schoenfeld residual**. If the HR is constant, the residuals show no trend over time; a trend means the coefficient is drifting. The standard test rescales the residuals and runs "a formal score test for slope=0" against (by default, Kaplan–Meier-transformed) time for each predictor and globally (R survival package documentation, after Grambsch & Therneau, 1994).

ILLUSTRATION / EXAMPLE

If households without the intervention always have 1.72 times the hazard of those with it, at month 6 and at month 20 alike, PH holds. If the protective effect fades (large early, none by month 20), PH fails, and a single HR averages a strong early effect with no late effect.

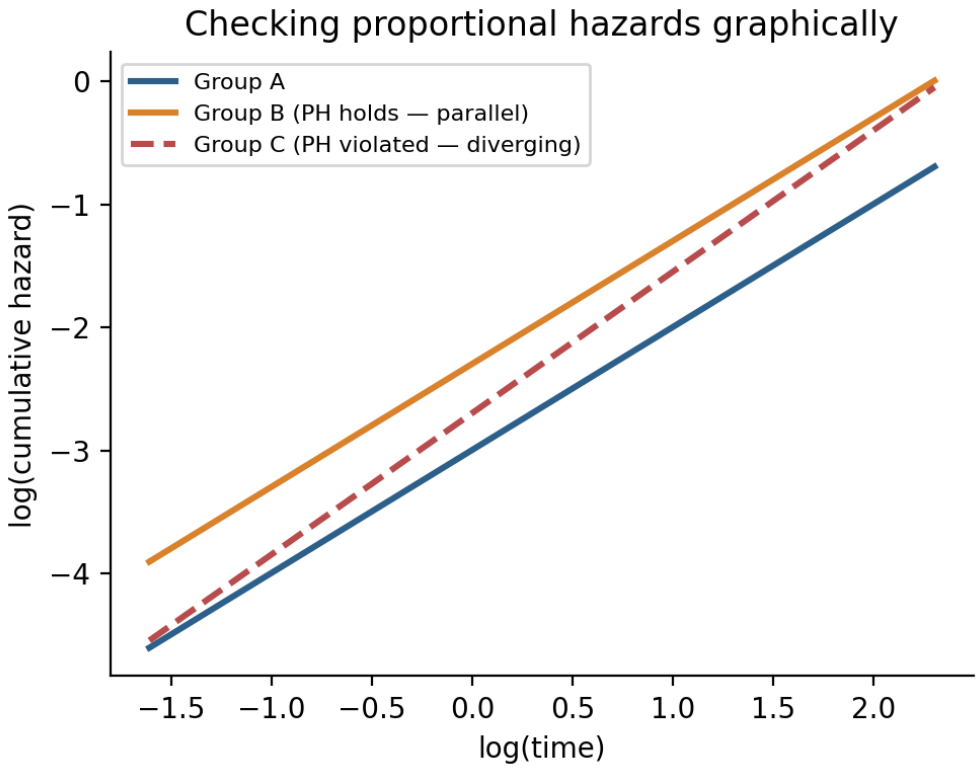


Figure 14.3. Schematic PH check: log cumulative hazard, log(−log S(t)), against log(time). Groups A and B run parallel (PH holds); Group C starts next to Group A and diverges from it over time (PH fails). The §14.7 cohort's version is Figure 14.4b.

IN R (OPTIONAL)

```
library(survival)
co ← read.csv("data/cohort_dropout.csv")
fc ← coxph(Surv(time_months, dropout) ~ intervention + income_k +
             distance_km, data = co)
cox.zph(fc)
```

```
             chisq df    p
intervention 0.273  1 0.60
income_k     1.602  1 0.21
distance_km  1.596  1 0.21
GLOBAL       3.282  3 0.35
```

Each row tests whether that predictor's Schoenfeld residuals trend with time. Intervention p = 0.60, income and distance p = 0.21, global p = 0.35 (3 df): no evidence against proportional hazards. A large p-value does not prove PH (recall §8.1: absence of evidence is not evidence of absence); read it with the plot. `plot(cox.zph(fc))` draws the residuals with a smooth line, which should be flat.

MUST-KNOW NOTES (MD-LEVEL)

- Check graphically with log(−log S(t)) against log t (parallel curves) and formally with the Schoenfeld residual test.

- If PH fails: stratify on the offending variable, let its effect vary with time, or report separate HRs for separate periods (§14.6.3).

CHECKPOINT 14.3

1. A Cox model reports HR = 0.6 for a treatment, but the two groups' log(−log S) curves clearly diverge over log time, and the Schoenfeld residual test gives p = 0.002 for treatment. What does this mean for the single HR of 0.6?
2. A Cox coefficient is b = −0.35 with SE 0.12. Give the HR and its 95% CI, and state the result in hazard language.
3. Explain why proportional hazards concerns the *ratio* of hazards between groups, not the hazard level itself.

(Answers in the Appendix at the end of the book.)

14.6.3 Stratified Cox, Time-Varying Covariates and Immortal-Time Bias

DEFINITION

A **stratified Cox model** fits a separate baseline hazard in each stratum of a variable, with the same HRs for the other predictors across strata. A **time-varying covariate** is a predictor whose value changes during follow-up, entered in **counting-process** (start, stop] format so that each subject contributes one row per period with a constant value. **Immortal-time bias** arises when a stretch of follow-up during which, "because of exposure definition, the outcome under study could not occur" is credited to the exposed group (Suissa, 2008).

SIMPLIFIED EXPLANATION

Stratification. If a variable violates PH, or you only need to control for it and not estimate its effect, stratify on it (a stratified Cox model). Each stratum gets its own $h_0(t)$, of any shape, so PH need not hold for that variable. The price: no HR for the stratifying variable.

Time-varying covariates. Some exposures start partway through follow-up: an education visit at month 3, a job loss at month 10. Coding them as fixed at baseline uses information from the future. Instead, split each subject's follow-up at the change: a row for (0, 3] unexposed and a row for (3, 12] exposed. Software builds these "start,stop data sets which have multiple intervals for each subject" (R survival package documentation). A related tool, a time-transform term, lets a *coefficient* change with time, such as exposure × log(t), which is also a direct test of PH.

Immortal time. If "ever visited" is fixed at baseline, every visited household had to stay enrolled until its visit. Those months before the visit are guaranteed dropout-free and are credited to the visited group. The exposure looks protective even when it does nothing. The bias grows with the length of the immortal period (Suissa, 2008).

ILLUSTRATION / EXAMPLE

A hypothetical CBHI programme visits households at month 3 after enrolment; households that drop out before month 3 are never visited. The programme reaches only some households; 100 are visited. The true dropout rate is 2.0 per 100 household-months in everyone, visited or not, so the visit has no effect.

WORKED CALCULATION (IMMORTAL-TIME BIAS)

1. Months 0–3: 600 household-months in total (300 from households later visited, 300 from households never visited). 12 dropouts, all among never-visited households, since a household that drops out cannot be visited later.
2. After month 3: visited households contribute 900 household-months with 18 dropouts (2.0 per 100); never-visited households contribute 400 household-months with 8 dropouts (2.0 per 100).
3. **Naive analysis ("ever visited" fixed at baseline).** Visited: 18 dropouts in 300 + 900 = 1,200 household-months, **1.50** per 100. Never visited: 12 + 8 = 20 dropouts in 300 + 400 = 700 household-months, **2.86** per 100. Rate ratio 1.50/2.86 = **0.53**: the visit appears to halve dropout.
4. **Correct analysis (visit as a time-varying covariate).** Each visited household's first 3 months are unexposed time. Exposed: 18 in 900, 2.0 per 100. Unexposed: 20 in 700 + 300 = 1,000, 2.0 per 100. Rate ratio **1.00**.

In counting-process format a visited household that drops out at month 8 has two rows:

id	tstart	tstop	visited	dropout
H017	0	3	0	0
H017	3	8	1	1

Its first 3 months are counted as unexposed, removing the immortal time.

IN R (OPTIONAL)

```
library(survival)
co ← read.csv("data/cohort_dropout.csv")
# above the median distance
co$far ← as.integer(co$distance_km > 5.1)
fs ← coxph(Surv(time_months, dropout) ~ intervention + income_k +
             strata(far), data = co)
round(summary(fs)$conf.int, 2)
ft ← coxph(Surv(time_months, dropout) ~ intervention + income_k +
             distance_km +
             tt(intervention), data = co,
           tt = function(x, t, ...) x * log(t))
round(summary(ft)$coefficients, 3)
```

```
             exp(coef) exp(-coef) lower .95 upper .95
intervention      0.61       1.65      0.42      0.88
income_k          0.95       1.06      0.91      0.99
                  coef exp(coef) se(coef)      z Pr(>|z|)
intervention    -0.484     0.617    0.406 -1.191    0.234
income_k        -0.062     0.940    0.022 -2.854    0.004
distance_km      0.108     1.114    0.027  3.928    0.000
tt(intervention) -0.029     0.972    0.179 -0.161    0.872
```

The stratified model gives near and far households (split at the 5.1 km median) separate baseline hazards; "separate baseline hazard functions are fit for each strata" (survival package documentation). There is no HR for distance, and the intervention HR is 0.61 (0.42–0.88), close to the 0.58 of the continuous-distance model in §14.7 but coarser, since a two-level split adjusts less finely. In the second model, `tt(intervention)` lets the intervention's log-HR change by −0.029 per unit of log(months), p = 0.87: no sign that the effect fades or grows. The `intervention` row is then the log-HR at t = 1 month (log t = 0), not an overall effect.

MUST-KNOW NOTES (MD-LEVEL)

- Stratify on a variable that violates PH or that you need to control for but not estimate. You lose its HR.
- Any exposure that begins after time zero must be a time-varying covariate. Classifying people by what happens later ("ever treated", "ever visited") creates immortal time and a spurious protective effect.
- An alternative is a **landmark analysis**: fix a time (say month 3), keep only those still event-free then, classify exposure by that time, and start the clock there.

CHECKPOINT 14.4

1. A cohort study reports that TB patients who "ever received" a nutrition kit had 20 deaths in 1,000 person-months, against 30 deaths in 1,000 person-months in those who never did. Kits were given at the end of month 2, and 200 of the kit group's person-months fell before the kit. Recompute the rates with kit receipt as a time-varying exposure, and the rate ratio before and after.
2. In the CBHI cohort, you suspect the effect of distance changes over follow-up and you only need to adjust for it. Name a way to do this and what you give up.

(Answers in the Appendix at the end of the book.)

14.7 Analysis Journey – Cox Regression

This journey, like §12.6 and §13.3, introduces no new method. It carries one time-to-event question to a reported hazard ratio with the tools of §14.1–14.6, in the order a real survival analysis takes: describe follow-up, plot, test crudely, model, check assumptions, interpret and report.

Research question. Among households newly enrolled in a CBHI scheme, is receiving a community education intervention associated with the hazard of dropout, independent of household income and distance to the nearest facility?

Study design. A cohort of 300 newly enrolled CBHI households (`data/cohort_dropout.csv`), 153 of whom received the intervention, followed for up to 24 months. **120 dropped out; 18 moved away; 162 were still enrolled at 24 months.** Total follow-up was 5,329.4 household-months, a crude dropout rate of 2.25 per 100 household-months.

Outcome. Time to dropout in months, with an event indicator (1 = dropped out, 0 = censored). Households that moved away or were still enrolled at 24 months are censored (recall §14.2). Treating moving away as censoring assumes a household that moved could still have dropped out afterwards; §14.8 checks what happens if it is instead a competing event.

Predictors/exposure. Intervention (1 = received, 0 = not) is the exposure. Income (Rs 1,000 per month; SD 4.42) and distance (km; SD 3.0) are covariates, chosen on substantive grounds as plausible confounders of both intervention uptake and dropout.

Why this method? A censored time-to-event outcome with several predictors. Ordinary regression cannot handle censoring (recall §14.1); the log-rank test cannot adjust for continuous covariates (recall §14.5). With 120 events and 3 predictors, there are 40 events per predictor, well above the convention of 10 (§14.6).

Crude analysis. The Kaplan–Meier curves (Figure 14.4a) separate early and stay apart: at 24 months, 0.66 of intervention households and 0.51 of the others were still enrolled. Neither curve reaches 0.5, so the **median time to dropout is not reached in either group**; retention at 24 months is the summary. The log-rank test gives $\chi^2 = 6.63$, 1 df, $p = 0.01$ (§14.5). A Cox model with the intervention alone gives the **crude HR 0.62 (95% CI 0.43–0.90)**.

Model specification. A Cox model with intervention, income and distance, all retained on substantive grounds rather than by automated selection.

Assumptions. Proportional hazards was checked two ways. The log(−log S) curves of the two groups run roughly parallel against log time (Figure 14.4b); the early steps are erratic because few events have occurred. The Schoenfeld residual test gives intervention $p = 0.60$ and global $p = 0.35$ (§14.6.2). Censoring is assumed non-informative: the 18 households that moved away are assumed no more or less likely to drop out than similar households that stayed.

Diagnostics. Dropping any one household changes the intervention HR only within 0.57 to 0.59 (dfbeta residuals: the change in a coefficient when that subject is removed). Income and distance are almost uncorrelated ($r = 0.08$), so collinearity is not a concern.

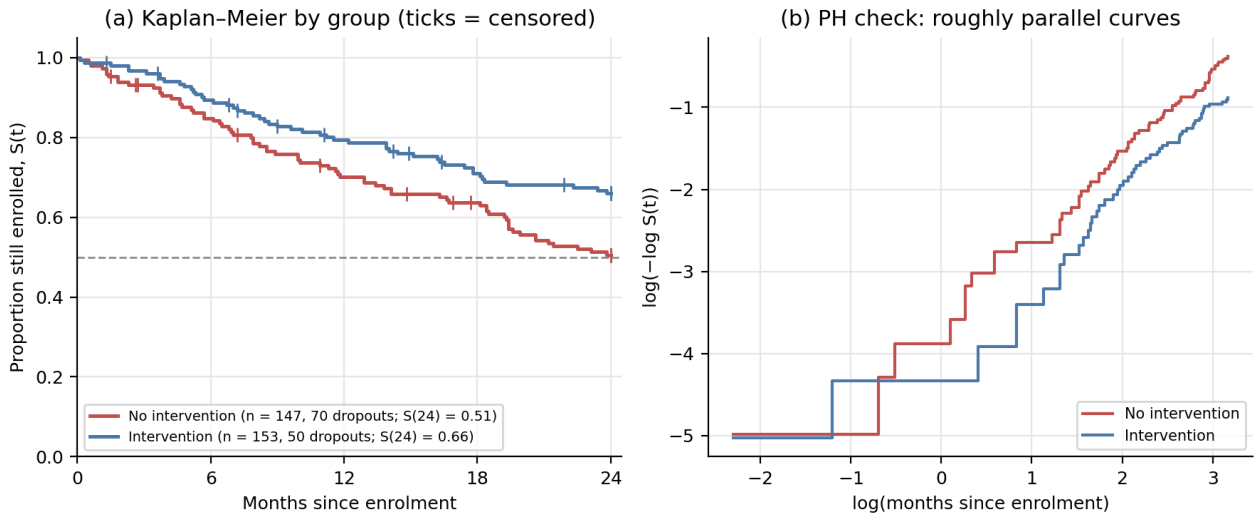


Figure 14.4. CBHI cohort by intervention group. (a) Kaplan–Meier retention with censoring ticks; neither curve falls to 0.5, so both medians are not reached. (b) log(−log S(t)) against log(months): the roughly constant vertical gap supports proportional hazards.

Output.

Predictor	HR	95% CI	p-value
Community education (yes vs no), **crude**	0.62	0.43 to 0.90	0.01
Community education (yes vs no), adjusted	0.58	0.40 to 0.84	0.004
Income (per Rs 1,000/month), adjusted	0.94	0.90 to 0.98	0.004
Distance (per km), adjusted	1.11	1.06 to 1.18	< 0.001

IN R (OPTIONAL)

```
library(survival)
co ← read.csv("data/cohort_dropout.csv")
crude ← coxph(Surv(time_months, dropout) ~ intervention, data = co)
fc ← coxph(Surv(time_months, dropout) ~ intervention + income_k +
             distance_km, data = co)
round(summary(crude)$conf.int, 2)
summary(fc)
```

```
             exp(coef) exp(-coef) lower .95 upper .95
intervention      0.62        1.6      0.43       0.9
...
  n= 300, number of events= 120

                 coef exp(coef) se(coef)      z Pr(>|z|)
intervention -0.54161   0.58181  0.18773 -2.885  0.00391 **
income_k     -0.06224   0.93965  0.02177 -2.858  0.00426 **
distance_km   0.10793   1.11397  0.02749  3.927 8.61e-05 ***
...
             exp(coef) exp(-coef) lower .95 upper .95
intervention    0.5818     1.7188    0.4027    0.8406
income_k        0.9397     1.0642    0.9004    0.9806
distance_km     1.1140     0.8977    1.0555    1.1756
...
Likelihood ratio test= 26.45  on 3 df,   p=8e-06
Wald test            = 26.16  on 3 df,   p=9e-06
Score (logrank) test = 27  on 3 df,   p=6e-06
```

`coef` is the log-HR; `exp(coef)` is the HR; z = coef/SE (−0.5416/0.1877 = −2.885); the CI is exp(b ± 1.96 SE) = exp(−0.542 ± 0.368) = 0.40 to 0.84. `exp(-coef)` is the inverted HR, 1.72 (§14.6.1). The crude model's score test equals the log-rank test (6.64 against 6.63, differing only in how tied times are handled).

Interpretation. Adjusting for income and distance, households that received the intervention had a **42% lower hazard of dropout** at any point during follow-up than those that did not (HR 0.58, 95% CI 0.40–0.84). This is a hazard: the instantaneous dropout rate among households still enrolled (recall §14.3), not the cumulative risk by 24 months. The HR moved from 0.62 (crude) to 0.58 (adjusted), a small shift. Each extra Rs 1,000 of monthly income went with a 6% lower hazard, and each extra km with an 11% higher hazard. Since PH was supported, one HR summarises the whole follow-up.

What the result does not mean. HR 0.58 is not "42% lower risk of dropping out by 24 months". The cumulative dropout gap is a separate quantity: the crude risk ratio at 24 months is 0.69 (§14.6.1), because cumulative risk also depends on the baseline hazard, which this model leaves unspecified. Nor does this observational association show that the intervention *causes* lower dropout. House-

holds that accepted it may differ in unmeasured ways (motivation, trust in the health system) that this analysis could not adjust for.

Reporting. *"Of 300 households followed for up to 24 months, 120 dropped out; median time to dropout was not reached (58% still enrolled at 24 months). In a Cox proportional hazards model adjusting for household income and distance to the nearest facility, receipt of the community education intervention was associated with a lower hazard of CBHI dropout (crude HR 0.62, 95% CI 0.43–0.90; adjusted HR 0.58, 95% CI 0.40–0.84; p = 0.004). Proportional hazards was supported graphically and by Schoenfeld residuals (global p = 0.35). Given the observational design, a causal effect of the intervention cannot be established from this analysis alone."*

MUST-KNOW NOTES (MD-LEVEL)

- Report the crude comparison (Kaplan–Meier, log-rank, crude HR) and the adjusted HR. The first shows the unadjusted difference; the second shows what remains after adjustment.
- State the median, or say it is not reached and give S(t) at fixed times.
- Check PH before leading with a single HR.
- Keep "hazard" and "risk" apart throughout, and state the causal caveat in the same paragraph as the main result.

14.8 Competing Risks

DEFINITION

A **competing event** is an event whose occurrence prevents the event of interest from happening. The **cumulative incidence function (CIF)**, F_k(t), is the probability of having event type k by time t *before* any competing event. It is estimated by the **Aalen–Johansen** estimator. Two regression approaches exist: the **cause-specific hazard** model (a Cox model for the event of interest, with competing events censored) and the **Fine–Gray subdistribution hazard** model (which models the CIF directly).

SIMPLIFIED EXPLANATION

Suppose the event of interest is sputum culture conversion among patients on treatment for drug-resistant TB, and some patients die first. A dead patient cannot convert. If you censor deaths and compute 1 − Kaplan–Meier for conversion, the method assumes the dead patients would have gone on to convert at the same rate as the survivors. That describes an imaginary world without death. In the real cohort, 1 − KM overstates the proportion who convert. Using the Kaplan–Meier function "results in estimates of incidence that are biased upward, regardless of whether the competing events are independent of one another" (Austin et al., 2016).

The CIF fixes this. At each conversion time, the probability of converting *now* is the probability of still being event-free of *both* kinds just before, times the fraction of the at-risk set converting now.

DERIVATION

Let S(t) be the Kaplan–Meier estimate treating *any* event (conversion or death) as the event: the probability of being free of both. At a conversion time t_j, with n_j at risk and d_{1j} conversions,

$P(\text{convert at } t_j) = S(t_{j-}) \times d_{1j}/n_j$, where $S(t_{j-})$ is S just before t_j.

Summing: $F_1(t) = \Sigma$ over $t_j \le t$ of $S(t_{j-}) \times d_{1j}/n_j$. By construction, $F_1(t) + F_2(t) + S(t) = 1$: everyone has converted, died first, or neither. 1 − KM for conversion instead multiplies $(1 - d_{1j}/n_j)$ factors with deaths censored, so each dead patient's share is carried forward as if they could still convert. Its value therefore exceeds $F_1(t)$ once a death has occurred, and 1 − KM for conversion plus 1 − KM for death can add to more than 1.

WORKED CALCULATION

Ten patients, weeks on treatment (C = converted, D = died before conversion, + = censored alive and unconverted): 2C, 3D, 4C, 5D, 6C, 7+, 8C, 9D, 10C, 12+.

Week	Event	At risk n_j	$S(t_{j-})$, free of both	CIF increment $S(t_{j-}) \times 1/n_j$	CIF of conversion	1 − KM, deaths censored
2	C	10	1.000	0.100	0.100	0.100
3	D	9	0.900	—	0.100	0.100
4	C	8	0.800	0.100	0.200	0.213
5	D	7	0.700	—	0.200	0.213
6	C	6	0.600	0.100	0.300	0.344
7	+	5	0.500	—	0.300	0.344
8	C	4	0.500	0.125	0.425	0.508
9	D	3	0.375	—	0.425	0.508
10	C	2	0.250	0.125	0.550	0.754

1. S(t) for "free of both" steps down at every C and D: 1 × 9/10 = 0.900; × 8/9 = 0.800; × 7/8 = 0.700; × 6/7 = 0.600; × 5/6 = 0.500; (censoring at 7: no step); × 3/4 = 0.375; × 2/3 = 0.250; × 1/2 = 0.125.
2. CIF of conversion at 12 weeks = 0.100 + 0.100 + 0.100 + 0.125 + 0.125 = **0.550**.
3. CIF of death before conversion, by the same rule at weeks 3, 5 and 9: 0.900/9 + 0.700/7 + 0.375/3 = 0.100 + 0.100 + 0.125 = **0.325**.
4. Check: 0.550 + 0.325 + 0.125 (neither yet) = 1.000.
5. 1 − KM with deaths censored: 1 − (9/10 × 7/8 × 5/6 × 3/4 × 1/2) = 1 − 0.246 = **0.754**. It overstates conversion by 0.754 − 0.550 = 0.20, because it treats the three dead patients as if they could still convert.

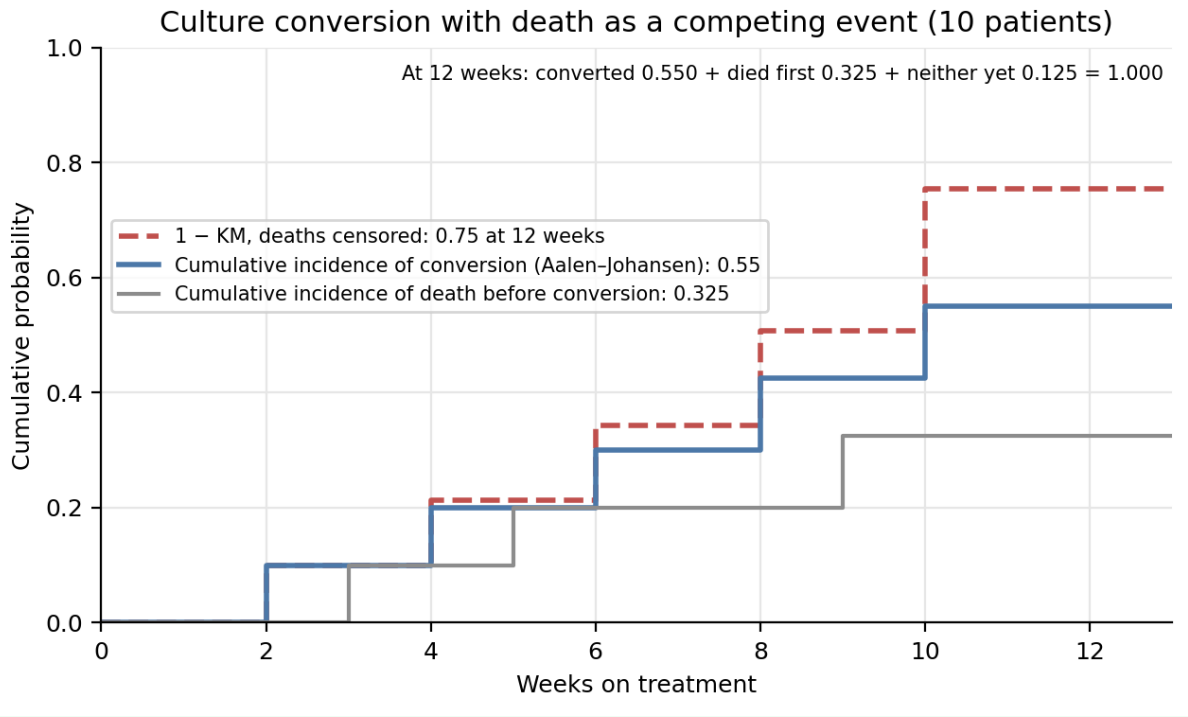


Figure 14.5. The ten-patient TB example. The dashed line (1 − KM, deaths censored) rises above the Aalen–Johansen cumulative incidence of conversion after the first death, and ends at 0.75 against 0.55.

ILLUSTRATION / EXAMPLE

Which regression model? The **cause-specific hazard** model is an ordinary Cox model for conversion with deaths censored. It estimates the effect of a covariate on the conversion *rate* among patients still alive and unconverted: an aetiological question. The **Fine–Gray** model estimates the effect of a covariate on the cumulative incidence of conversion, that is, on the actual proportion who convert in a population where people also die. Its HR (the **subdistribution HR**) is harder to read as a rate, because people who have already died stay in its risk set. Austin et al. (2016) summarise: the cause-specific model "may be better suited for addressing etiologic questions", the CIF model "for estimating a patient's clinical prognosis". Lau et al. (2009) add that the choice should be "guided by the scientific question". The two HRs can differ: a treatment that lowers only the death hazard leaves the cause-specific HR for conversion at 1, yet raises the cumulative incidence of conversion, because it keeps more patients alive to convert.

The CBHI cohort as a sensitivity analysis. Suppose moving out of the district ends CBHI membership by a route the scheme does not count as dropout. Moving is then a competing event, not censoring. The §14.7 Cox model already treats movers as censored, so its HR of 0.58 is the cause-specific HR, and it stands for the aetiological question. What changes is the proportion who drop out: 1 − KM gives 0.42 by 24 months (1 − 0.58), the CIF gives 0.40. With only 18 movers the difference is small; when competing events are common (deaths in an elderly cohort), it is large.

IN R (OPTIONAL)

```
library(cmprsk)
# status: 1 converted, 2 died, 0 censored
tb ← data.frame(week   = c(2, 3, 4, 5, 6, 7, 8, 9, 10, 12),
                status = c(1, 2, 1, 2, 1, 0, 1, 2, 1, 0))
timepoints(cuminc(tb$week, tb$status, cencode = 0), c(4, 8, 12))$est
co ← read.csv("data/cohort_dropout.csv")
co$status3 ← ifelse(co$dropout == 1, 1,
                    ifelse(co$censor_reason == "moved away", 2, 0))
ci ← cuminc(co$time_months, co$status3, group = co$intervention,
            cencode = 0)
round(timepoints(ci, 24)$est, 3)
ci$Tests
fg ← crr(co$time_months, co$status3,
         cov1 = as.matrix(co[, c("intervention", "income_k",
                                 "distance_km")]),
         failcode = 1, cencode = 0)
round(summary(fg)$conf.int, 2)
```

```
      4     8    12
1 1 0.2 0.425 0.550
1 2 0.1 0.200 0.325
       24
0 1 0.476
1 1 0.327
0 2 0.054
1 2 0.065
       stat         pv df
1 6.5294374 0.01061035  1
2 0.1539458 0.69479281  1
             exp(coef) exp(-coef) 2.5% 97.5%
intervention      0.59       1.70 0.41  0.85
income_k          0.94       1.06 0.90  0.98
distance_km       1.11       0.90 1.06  1.17
```

`cuminc` rows are labelled "group event". For the ten TB patients (one group, labelled 1), the conversion CIF is 0.550 and the death CIF 0.325 at 12 weeks, as in the hand table. For the cohort, the CIF of dropout by 24 months is 0.476 without and 0.327 with the intervention (1 − KM gives 0.495 and 0.339); moving away reached 0.054 and 0.065. `Tests` gives Gray's test comparing the groups' CIFs: p = 0.011 for dropout, 0.69 for moving. `crr` fits the Fine–Gray model: the subdistribution HR for the intervention is 0.59 (0.41–0.85), close to the cause-specific 0.58 because moving is rare and shows no clear link to the intervention (Gray's p = 0.69).

MUST-KNOW NOTES (MD-LEVEL)

- With competing events, 1 − KM overestimates the cumulative incidence. Use the CIF (Aalen–Johansen); $F_1 + F_2 + S = 1$.
- Cause-specific HR (Cox, competing events censored): effect on the rate among those still at risk; suits aetiology. Subdistribution HR (Fine–Gray): effect on cumulative incidence; suits prognosis and planning. Report which one you used.
- Decide at the design stage which events are censoring (the outcome could still happen, unseen) and which are competing (the outcome can no longer happen).

CHECKPOINT 14.5

1. Five patients: event of interest at week 1, competing event at week 2, event of interest at week 3, censored at week 4, event of interest at week 5. Compute the CIF of the event of interest at week 3, and 1 − KM at week 3 with the competing event censored.
2. A district study of elderly patients after hip fracture asks (a) whether diabetes raises the rate of re-fracture among those still alive, and (b) what proportion of patients will re-fracture within two years, for service planning. Death is common. Which model suits each question?

(Answers in the Appendix at the end of the book.)

Advanced Topics for the Community Medicine Researcher

This chapter applies the regression and testing toolkit of Chapters 8–14 to problems a community-medicine researcher meets again and again: choosing a model (§15.1), clustered and survey data (§15.2), confounding in observational comparisons (§15.3), missing values (§15.4), diagnostic tests and prediction models (§15.5), sample size (§15.6), meta-analysis (§15.7), Bayesian reasoning (§15.8) and randomised trials (§15.9). The sections are largely independent; each names the earlier sections it builds on. The running example is again the CBHI survey: 500 households in 20 villages, 25 per village, with 30% (150/500) CBHI members.

15.1 Model Selection

DEFINITION

Model selection is choosing which predictors, and in what form, enter a final regression model. **Overfitting** is what happens when a model has too many parameters for the data: it fits the noise of this sample as well as the signal, so it describes this sample better than it will describe the next one.

SIMPLIFIED EXPLANATION

A model with too few predictors misses real relationships. A model with too many learns the quirks of one sample. Figure 15.1 shows the pattern: as predictors are added, the error in the data used to fit the model keeps falling, but the error in new data eventually rises again. The tools in this section, likelihood (§15.1.1), AIC and BIC (§15.1.2), stepwise search (§15.1.3) and the likelihood ratio test (§15.1.4), all try to find the turning point. §15.1.5 then explains why the right tool depends on what the model is for.

ILLUSTRATION / EXAMPLE

With 30 candidate predictors and 100 observations, a model containing all 30 can reproduce the sample closely and still predict new patients badly: it has spent its information on idiosyncrasies of these 100 people.

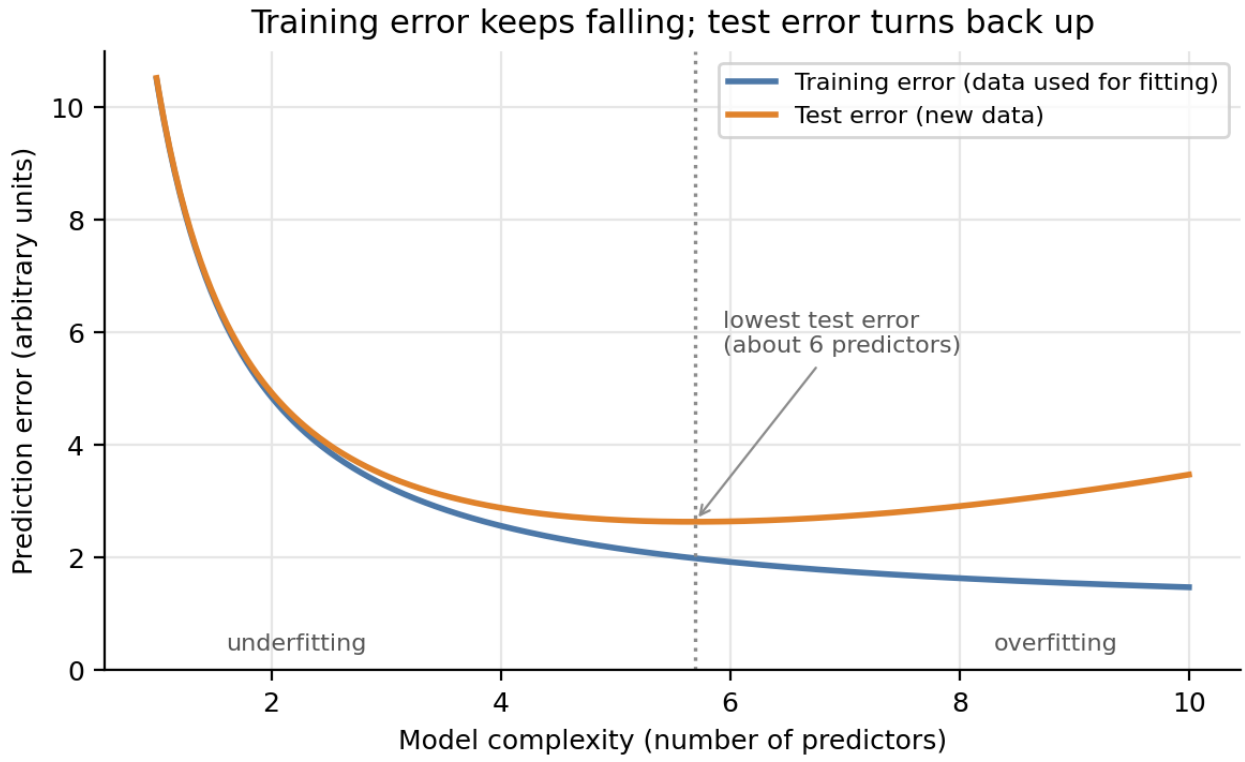


Figure 15.1. Schematic: error in the data used for fitting (training error) falls steadily as predictors are added; error in new data (test error) falls, then rises once the model starts fitting noise.

MUST-KNOW NOTES (MD-LEVEL)

- Underfitting (too few predictors) and overfitting (too many) are the two failure modes; selection tools balance them.
- Statistical criteria never replace subject knowledge about which variables are plausible confounders or predictors (§15.1.5).

15.1.1 Model Fit and the Idea of Likelihood

DEFINITION

Recall from §13.1 that the **likelihood** L of a set of parameter values is the probability of the observed data calculated as if those values were true, and that maximum likelihood chooses the values that make L as large as possible. The maximised **log-likelihood**, ln L, is a measure of **model fit**: the less negative it is, the better the model fits the data it was fitted to. The **deviance** is −2 ln L (up to a constant that is the same for every model fitted to the same data): lower deviance means better fit.

SIMPLIFIED EXPLANATION

R^2 (recall §12.2.5) measures fit by squared residuals, and it can only rise when a predictor is added. Log-likelihood works for every model fitted by maximum likelihood (logistic, Poisson; Cox regression uses a closely related partial likelihood) and has exactly the same weakness: adding a predictor, even pure

noise, never lowers the maximised likelihood. So raw fit cannot choose between models of different sizes. A fair comparison needs a penalty that grows with the number of parameters (§15.1.2) or a formal test of whether the extra parameters earn their place (§15.1.4).

DERIVATION

Why an extra predictor can never lower the maximised likelihood. Let model A have coefficients b_0, b_1, and model B add b_2.

1. Model B with b_2 set to 0 is exactly model A, so it can reach A's maximum likelihood, L_A.
2. Maximum likelihood for B searches over all values of b_2, including 0.
3. So L_B ≥ L_A, with equality only if the best b_2 is exactly 0.

In a real sample the best b_2 for a noise variable is almost never exactly 0, so ln L rises a little.

ILLUSTRATION / EXAMPLE

In the CBHI survey (recall the membership model of §13.3), membership modelled on distance and wealth gives ln L = −276.13. Adding maternal education (two more coefficients) raises it to −273.58. Adding household size as well, a variable with no detectable association with membership (crude p = 0.56), still nudges it up, to −273.57.

MUST-KNOW NOTES (MD-LEVEL)

- ln L measures fit for any maximum-likelihood model; deviance = −2 ln L.
- A bigger model never fits its own sample worse (the Derivation). So fit alone cannot choose model size.

15.1.2 AIC and BIC

DEFINITION

For a model with k estimated parameters (including the intercept), fitted to n observations:

$$AIC = -2 \ln L + 2k$$

$$BIC = -2 \ln L + k \times \ln n$$

These are the **Akaike** and **Bayesian (Schwarz) information criteria** (formulas as documented in R; R Core Team). The penalty is **added** to the deviance, so **lower is better**. The derivations of the two penalties (from information theory and from an approximation to Bayesian model comparison) are beyond this book.

SIMPLIFIED EXPLANATION

The first term, −2 ln L, falls as fit improves. The second term rises by 2 (AIC) or by ln n (BIC) for every parameter added. A new parameter lowers AIC only if it lowers the deviance by more than 2. BIC's penalty per parameter, ln n, exceeds AIC's 2 whenever ln n > 2, that is, n ≥ 8 ($e^2 = 7.39$). In any realistic study BIC therefore penalises size more heavily and tends to choose smaller models.

ILLUSTRATION / EXAMPLE

Three CBHI membership models, all fitted to the same 500 households: A (distance, wealth), B (A + maternal education, 2 parameters), C (B + household size).

WORKED CALCULATION

1. Model A: k = 3, ln L = −276.13, deviance = 552.25. AIC = 552.25 + 2 × 3 = **558.25**. BIC = 552.25 + 3 × ln 500 = 552.25 + 3 × 6.215 = **570.90**.
2. Model B: k = 5, deviance = 547.16. AIC = 547.16 + 10 = **557.16**. BIC = 547.16 + 5 × 6.215 = **578.24**.
3. Model C: k = 6, deviance = 547.15. AIC = 547.15 + 12 = **559.15**. BIC = 547.15 + 6 × 6.215 = **584.43**.
4. AIC prefers B (lowest, 557.16). BIC prefers A (570.90). Education lowers the deviance by 5.09 for 2 parameters: more than AIC's price of 4, less than BIC's price of 12.43.
5. Both reject C: household size lowers the deviance by 0.02, far less than either price.

IN R (OPTIONAL)

```
cb ← read.csv("data/cbhi_survey.csv")
cb$maternal_education ← factor(cb$maternal_education,
  levels = c("none", "primary", "secondary+"))
mA ← glm(cbhi_member ~ distance_km + wealth_index,
         family = binomial, data = cb)
mB ← update(mA, . ~ . + maternal_education)
mC ← update(mB, . ~ . + household_size)
sapply(list(A = mA, B = mB, C = mC), function(m)
  round(c(k = attr(logLik(m), "df"), lnL = as.numeric(logLik(m)),
          AIC = AIC(m), BIC = BIC(m)), 2))
anova(mA, mB, test = "LRT")
```

```
          A       B       C
k      3.00    5.00    6.00
lnL -276.13 -273.58 -273.57
AIC  558.25  557.16  559.15
BIC  570.90  578.24  584.43
Analysis of Deviance Table

Model 1: cbhi_member ~ distance_km + wealth_index
Model 2: cbhi_member ~ distance_km + wealth_index + maternal_education
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       497     552.25
2       495     547.16  2   5.0895  0.07849 .
```

`logLik` gives ln L and its `df` attribute gives k; `AIC` and `BIC` reproduce the hand calculations. The second block is the likelihood ratio test of §15.1.4: the drop in deviance (5.09) on 2 df.

MUST-KNOW NOTES (MD-LEVEL)

- AIC = −2 ln L + 2k; BIC = −2 ln L + k ln n. Penalty added, lower is better.
- BIC penalises more than AIC once n ≥ 8, so it favours smaller models. When they disagree, as here, the data do not strongly favour either model.
- A lower AIC means a better fit-versus-complexity trade-off *among the candidates compared*, not a true or even a good model.

- Compare AIC or BIC only between models fitted to the same outcome and exactly the same observations. A variable with missing values silently drops rows and makes the comparison invalid.
- k counts parameters, not variables: a three-level categorical predictor adds 2.

15.1.3 Stepwise Selection

DEFINITION

Stepwise selection adds or removes predictors one at a time by a fixed rule, a p-value threshold or a fall in AIC, until no single change improves the criterion. **Forward** selection starts empty and adds; **backward elimination** starts full and removes; **bidirectional** does both.

SIMPLIFIED EXPLANATION

Stepwise selection is a search procedure; AIC, BIC or a p-value is the scoring rule it follows. It is quick and automatic, and that is the problem: it has no subject knowledge, it answers a different question in every sample, and the p-values and confidence intervals it reports for the survivors ignore the search that produced them. A review of variable selection concludes that selection, "in particular if used in explanatory modeling where effect estimates are of central interest, can compromise stability of a final model, unbiasedness of regression coefficients, and validity of p-values or confidence intervals" (Heinze, Wallisch & Dunkler, 2018). Treat stepwise output as a warning sign, not a method.

ILLUSTRATION / EXAMPLE

Backward elimination from model C (§15.1.2) with the AIC penalty keeps distance, wealth and education (model B). With the BIC penalty it keeps only distance and wealth (model A). Same data, same procedure, different answer, because the scoring rule changed.

To see the instability, repeat the AIC search in 200 bootstrap samples (samples of 500 households drawn with replacement from the 500; the resampling idea of §7.2.4). Distance is kept in 96.5% of samples and wealth in 100%, but education only in 69.5%, and household size, which shows no detectable association with membership, still enters 15.5% of the final models.

IN R (OPTIONAL)

```
# cb as in §15.1.2
full ← glm(cbhi_member ~ distance_km + wealth_index +
             maternal_education + household_size,
           family = binomial, data = cb)
# AIC penalty (k = 2)
formula(step(full, trace = 0))
# BIC penalty (k = ln n)
formula(step(full, trace = 0, k = log(500)))
# how stable is the AIC choice?
set.seed(2026)
keep ← replicate(200, {
  b ← cb[sample(500, replace = TRUE), ]
  s ← step(update(full, data = b), trace = 0)
  vars ← c("distance_km", "wealth_index", "maternal_education",
           "household_size")
  vars %in% attr(terms(s), "term.labels")
})
setNames(rowMeans(keep),
         c("distance", "wealth", "education", "household size"))
```

```
cbhi_member ~ distance_km + wealth_index + maternal_education
cbhi_member ~ distance_km + wealth_index
      distance         wealth      education household size
         0.965          1.000          0.695          0.155
```

`step` drops terms while AIC falls; `k = log(500)` turns the penalty into BIC's. The last line is the share of bootstrap samples whose final model kept each variable: an honest report of how much the "selected model" depends on the sample.

MUST-KNOW NOTES (MD-LEVEL)

- Stepwise selection is a search; the criterion (p-value, AIC, BIC) decides its answer.
- Known problems: unstable choices, inflated apparent significance, and p-values and CIs that ignore the selection (Heinze et al., 2018).
- Do not use it to choose confounders: confounder status is a causal question (recall §11.6.1), which no p-value or fit statistic can settle.
- If selection is unavoidable, report inclusion frequencies across bootstrap samples, as above (Heinze et al., 2018).

15.1.4 Likelihood Ratio Test

DEFINITION

The **likelihood ratio test (LRT)** compares two **nested** models: the reduced model is the full model with some parameters fixed at 0. It tests H_0: those extra parameters are all 0. The statistic is

$$LR = -2 \times (\ln L_{reduced} - \ln L_{full}) = deviance_{reduced} - deviance_{full}$$

Under H_0, LR follows approximately a χ^2 distribution (recall §5.7.2.4) with **df = the difference in the number of estimated parameters**. The approximation is a large-sample result (Wilks' theorem; proof beyond this book).

SIMPLIFIED EXPLANATION

AIC and BIC score models; the LRT is a hypothesis test (recall Chapter 8). The Derivation in §15.1.1 guarantees that LR ≥ 0. The question is whether the gain in fit is bigger than chance alone would produce with that many extra parameters. Nesting matters: every parameter of the reduced model must be in the full model, and both must use the same observations. The df counts parameters, not variables: adding maternal education (three levels) adds 2 parameters and 2 df.

ILLUSTRATION / EXAMPLE

§13.3 fitted the CBHI membership model with education. Does education, as a whole, add to distance and wealth?

WORKED CALCULATION

1. ln L_reduced (model A) = −276.126; ln L_full (model B) = −273.581.
2. LR = −2 × (−276.126 + 273.581) = −2 × (−2.545) = **5.09**.
3. df = 5 − 3 = **2** (two education coefficients).
4. p = P(χ^2_2 > 5.09) = **0.078**. For 2 df the χ^2 upper tail is exactly e^(−x/2), so p = e^(−2.545) = 0.078.
5. Conclusion: weak evidence that maternal education adds to distance and wealth. Note the contrast with §13.3, where the single coefficient for secondary education had p = 0.027. The LRT asks a different question (both education coefficients together, 2 df) and spends its evidence on both.

MUST-KNOW NOTES (MD-LEVEL)

- LR = −2(ln L_reduced − ln L_full); χ^2 with df = difference in the number of parameters; approximate, large-sample.
- Only for nested models fitted to the same observations. For non-nested models, use AIC or BIC.
- For a multi-level categorical variable, the LRT (joint test) is the right test of "does this variable matter?", not the p-value of any one level.

CHECKPOINT 15.1

1. A colleague says, "Model B has lower AIC than model A, so model B is the true model." What is the overstatement?
2. Why does adding a purely random predictor never lower a model's maximised likelihood? Give the argument, not just the conclusion.
3. A logistic model on 400 children has ln L = −210.4 with 4 parameters. Adding a four-category feeding-practice variable gives ln L = −205.9. Compute the LR statistic, its df, the change in AIC and the change in BIC. Which criteria favour the larger model?

(Answers in the Appendix at the end of the book.)

15.1.5 Modelling Strategy – Explanatory, Predictive and Descriptive Aims

DEFINITION

A **modelling strategy** is the plan, fixed before the analysis, for which variables enter a model and in what form. It follows from the model's **aim**:

- **Explanatory (causal):** estimate the effect of one exposure on an outcome, free of confounding.
- **Predictive:** predict the outcome for new individuals as accurately as possible.
- **Descriptive:** summarise how the outcome is distributed across groups, without a causal or predictive claim.

Shmueli (2010) argues that "conflation between explanation and prediction is common, yet the distinction must be understood", because the two aims lead to different choices at every step of modelling.

SIMPLIFIED EXPLANATION

Explanatory models are built from a causal diagram (recall §11.6.1). Draw the DAG, find a sufficient adjustment set by the backdoor criterion, and put those variables in the model whether or not they are "significant". Do not adjust for mediators or colliders. The coefficient of interest is one number, the exposure's; the other coefficients are there to remove confounding and should not be interpreted as effects in their own right. Check that the data can support the model: the events-per-variable convention (recall §13.2.4) limits how many parameters you can afford.

Predictive models choose variables for accuracy in new data. Causal structure does not matter; a variable that is a consequence of the outcome's causes can be a good predictor. What matters is not overfitting: prefer a pre-specified set of predictors chosen from the literature, penalised methods that shrink coefficients towards 0 (shrinkage: pulling estimates towards 0 to offset overfitting), and honest validation in data not used for fitting. §15.5.2 (validating clinical prediction models by discrimination and calibration) develops this.

Descriptive models report, for example, prevalence of anaemia by district and wealth quintile. They should be pre-specified and interpreted as associations only.

Two habits to drop.

1. **Univariable p-value screening** (keeping only variables with $p < 0.05$ or 0.20 in a crude analysis). It "wrongly rejects potentially important variables when the relationship between an outcome and a risk factor is confounded" (Sun, Shook & Kay, 1996): a crude association can be hidden by confounding and appear only after adjustment.
2. **Stepwise selection for explanatory models** (§15.1.3). It chooses variables by their association with the outcome, not by their causal role.

Form matters as well as choice. Keep continuous variables continuous, and model curvature with categories, a trend test or splines rather than cut-points (recall §12.4.5).

ILLUSTRATION / EXAMPLE

Three questions from the same CBHI survey:

Aim	Question	Strategy
Explanatory	Does distance to the facility reduce CBHI membership?	Assumed DAG: wealth and maternal education affect both where households live relative to the facility and whether they join; adjust for both, keep them whatever their p-values (model B, §15.1.2). Report the distance OR, 0.91 per km (recall §13.3).
Predictive	Which households should the scheme's field workers visit first?	Pre-specify candidate predictors; fit with shrinkage; judge by discrimination and calibration in new villages (§15.5.2), not by p-values.
Descriptive	What share of households in each education group are members?	Tabulate 31/168 (18.5%), 50/177 (28.2%), 69/155 (44.5%); no adjustment and no causal claim.

In the explanatory model, a univariable screen at $p < 0.05$ would have kept distance (crude $p < 0.001$), wealth ($p < 0.001$) and education ($p < 0.001$), and dropped household size ($p = 0.56$). Here the screen happens to agree with the DAG, but that is a property of this dataset, not of the method: in another dataset a confounder with a weak crude association would be screened out.

MUST-KNOW NOTES (MD-LEVEL)

- State the aim first. Explanatory → DAG-based adjustment set, one coefficient interpreted. Predictive → accuracy in new data, shrinkage, validation. Descriptive → pre-specified tabulation, association only.
- Do not screen variables by univariable p-values (Sun et al., 1996), and do not use stepwise selection to choose confounders (Heinze et al., 2018).
- In an explanatory model, do not read the confounders' coefficients as their effects. They may themselves be confounded, or be direct rather than total effects, even when the exposure's estimate is not: the "Table 2 fallacy" (Westreich & Greenland, 2013). Interpret the exposure coefficient only.
- Keep continuous variables continuous; use splines for curvature (recall §12.4.5).

CHECKPOINT 15.2

1. A thesis protocol says, "Variables with $p < 0.2$ in univariable analysis will be entered into a multivariable model, and backward elimination will give the final model of risk factors for low birthweight." Give two problems, one about each step.
2. For a model predicting which pregnant women will miss their third antenatal visit, is it acceptable to include a variable that is not a cause of missing the visit? Would the same variable be acceptable in a model estimating the effect of ASHA home visits on attendance?
3. In the CBHI bootstrap of §15.1.3, household size entered 15.5% of the final AIC-selected models. Out of 200 bootstrap samples, how many is that, and what does it tell you about a single stepwise run?

(Answers in the Appendix at the end of the book.)

15.2 Clustered and Hierarchical Data

DEFINITION

Clustered (hierarchical) data have observations grouped inside larger units: households in villages, children in anganwadi centres, patients in PHCs, or repeated measurements inside one person. Observations in the same cluster tend to be more alike than observations from different clusters, so they are not independent.

SIMPLIFIED EXPLANATION

Most methods in this book assume independent observations (recall §5.4). Households in one village share a water source, a sub-centre, an ASHA and a local economy, so their outcomes move together. Dependence arises in three ways, which often occur together:

- **Design clustering:** the survey sampled villages first, then households (recall cluster and multistage sampling, §6.2.4–6.2.5).
- **Natural hierarchy:** children are nested in schools whether or not the sample was drawn that way.
- **Repeated measurements:** blood pressure at 0, 1 and 3 months in the same patient.

All three make an analysis that treats every observation as independent too confident. The ICC (§15.2.1) measures the dependence; the design effect (§15.2.2) prices it; random-effects models (§15.2.3–15.2.4), GEE (§15.2.5) and design-based survey analysis (§15.2.6) analyse it.

ILLUSTRATION / EXAMPLE

The CBHI survey sampled 20 villages and 25 households in each. Figure 15.2 shows membership village by village: from 8% in V04 and V07 to 56% in V02, around an overall 30%. If households were independent, village percentages based on 25 households would vary with a variance of about 0.30 × 0.70 ÷ 25 = 0.0084. The observed variance of the 20 village proportions is 0.0203, about 2.4 times as large: households in the same village resemble each other.

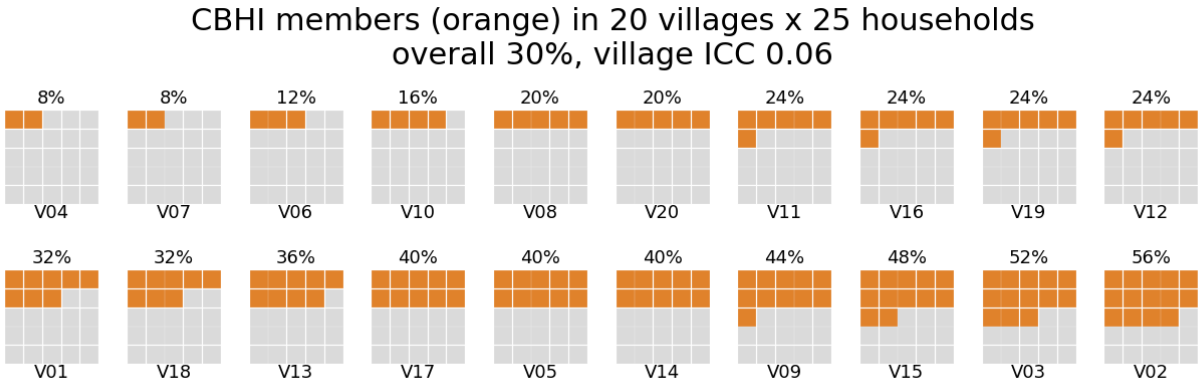


Figure 15.2. Each block is one village of 25 households (data/cbhi_survey.csv), sorted by membership; an orange square is a member household. Village membership ranges from 8% to 56% around the overall 30%, more spread than chance alone would give with 25 households per village.

MUST-KNOW NOTES (MD-LEVEL)

- Design clustering, natural hierarchy and repeated measures are different sources of the same problem: non-independent observations.
- Ignoring positive clustering usually gives standard errors that are too small, CIs that are too narrow and p-values that are too small.

15.2.1 Intraclass Correlation Coefficient (ICC)

DEFINITION

The **intraclass (intracluster) correlation coefficient**, ρ, is the share of the total variance of an outcome that lies between clusters:

$\rho = \sigma^2_b \div (\sigma^2_b + \sigma^2_w)$

where σ^2_b is the variance of the true cluster means (between clusters) and σ^2_w is the variance of individuals around their own cluster mean (within clusters). Equivalently, ρ is the correlation between the outcomes of two individuals drawn from the same cluster.

SIMPLIFIED EXPLANATION

If ρ = 0, knowing a household's village tells you nothing about its outcome. If ρ = 1, every household in a village has the same outcome, and the village is the only thing that varies. The ICC is estimated from one-way ANOVA (recall §9.5.1), which already splits variation into between-group and within-group parts. Chapter 16 (§16.3.2, test–retest reliability) uses the same ratio, with subjects in place of clusters and repeated ratings in place of households, to measure reliability; there the ICC comes in several forms, and only the absolute-agreement forms are lowered by a systematic shift between occasions or raters.

DERIVATION

1. Estimating ρ from ANOVA. With k clusters of m individuals each, one-way ANOVA gives the between-cluster mean square MSB and the within-cluster mean square MSW. Their expected values are

$E(MSW) = \sigma^2_w$

$E(MSB) = \sigma^2_w + m\,\sigma^2_b$

(each cluster mean carries within-cluster noise σ^2_w/m plus the real between-cluster variation σ^2_b; multiplying by m gives MSB). Solving:

$\hat{\sigma}^2_b = (MSB - MSW) \div m$, and $\hat{\rho} = \hat{\sigma}^2_b \div (\hat{\sigma}^2_b + MSW) = (MSB - MSW) \div [MSB + (m - 1)\,MSW]$.

If MSB happens to be smaller than MSW, $\hat{\rho}$ is **negative**. The true ρ in this model cannot be below 0, but the estimate can, by chance, when real clustering is absent or very weak.

2. Why DEFF ≈ 1 + (m − 1)ρ. This is the result §6.2.6 previewed. Take one cluster of m individuals, each with variance $\sigma^2 = \sigma^2_b + \sigma^2_w$, and every pair correlated ρ. The variance of the cluster total is the sum of m variances and m(m − 1) covariances, each $\rho\sigma^2$:

$\text{Var}(y_1 + \ldots + y_m) = m\sigma^2 + m(m-1)\rho\sigma^2 = m\sigma^2[1 + (m-1)\rho]$

Divide by m^2 for the cluster mean: $\text{Var}(\bar{y}_\text{cluster}) = (\sigma^2/m)[1 + (m-1)\rho]$. The overall mean of k independent clusters has variance $(\sigma^2/km)[1 + (m-1)\rho]$. A simple random sample (SRS) of the same n = km individuals has variance $\sigma^2/(km)$. The ratio is

DEFF = Var(cluster design) ÷ Var(SRS) = $1 + (m-1)\rho$.

The formula is exact for equal cluster sizes and this correlation model; with unequal clusters it is an approximation using the average cluster size.

ILLUSTRATION / EXAMPLE

The CBHI survey: membership (0/1) by village, k = 20, m = 25.

WORKED CALCULATION

1. One-way ANOVA of membership on village gives MSB = 0.5074 (19 df) and MSW = 0.1987 (480 df).
2. $\sigma^2_b = (0.5074 - 0.1987) \div 25 = 0.01235$.
3. $\hat{\rho} = 0.01235 \div (0.01235 + 0.1987) =$ **0.0585**, which rounds to **0.06**. About 6% of the variation in membership lies between villages and 94% within them.
4. Check with the second form: $(0.5074 - 0.1987) \div (0.5074 + 24 \times 0.1987) = 0.3087 \div 5.276 = 0.0585$.
5. The same calculation for "used a health facility in the past month" gives $\hat{\rho} =$ **−0.006**: MSB is slightly below MSW, so there is no detectable village clustering of facility use, and the negative estimate should be read as "about 0".

MUST-KNOW NOTES (MD-LEVEL)

- $\rho = \sigma^2_b/(\sigma^2_b + \sigma^2_w)$: the between-cluster share of variance, or the correlation of two members of one cluster.
- ANOVA estimate: $(\text{MSB} - \text{MSW})/[\text{MSB} + (m-1)\text{MSW}]$. It can be negative by chance; report such a value and interpret it as no detectable clustering.
- An ICC of a few hundredths is not negligible: its cost grows with cluster size (§15.2.2).
- The ICC depends on the outcome: membership clusters by village here (0.06); facility use does not (−0.006).

15.2.2 Design Effect

DEFINITION

Recall from §6.2.6 that the **design effect** is DEFF = Var(estimate under the actual design) ÷ Var(estimate under SRS of the same size), approximately $1 + (m-1)\rho$ for a cluster sample (derived in §15.2.1). The **effective sample size** is n ÷ DEFF. **DEFT** = √DEFF is the factor by which the design inflates the **standard error**.

SIMPLIFIED EXPLANATION

DEFF multiplies the variance, so it multiplies the standard error by √DEFF, not by DEFF. With DEFF = 2.44 the SE grows by a factor of 1.56, not 2.44. DEFF is a planning and precision concept: it tells you how many households you need (§7.3 uses it) and how far a naive SE understates the truth. It is not by itself an analysis. To estimate an adjusted association from clustered data you still need a method that handles the clustering: a mixed model (§15.2.4), GEE (§15.2.5) or a design-based survey analysis (§15.2.6).

ILLUSTRATION / EXAMPLE

CBHI membership: m = 25, ρ = 0.06, prevalence 30% (150/500).

WORKED CALCULATION

1. DEFF = 1 + (25 − 1) × 0.06 = 1 + 1.44 = **2.44**.
2. Effective sample size = 500 ÷ 2.44 ≈ **205** households.
3. SRS standard error = √(0.30 × 0.70 ÷ 500) = **0.0205**.
4. DEFT = √2.44 = **1.56**. Cluster-corrected SE ≈ 0.0205 × 1.56 = **0.032**.
5. 95% CI for membership: treating households as independent, 0.30 ± 1.96 × 0.0205 = 26.0% to 34.0%; allowing for villages, 0.30 ± 1.96 × 0.032 = 23.7% to 36.3%.
6. The common error: multiplying the SE by DEFF gives 0.0205 × 2.44 = 0.050, an interval more than half as wide again as it should be.
7. Direct check: the SE of a mean of 20 equal-sized village proportions is SD(village proportions)/√20 = √0.0203/√20 = 0.0319. $(0.0319 \div 0.0205)^2 \approx 2.4$; with the unrounded SEs (0.03185 and 0.02051, the latter with n − 1 in place of n) it is **2.41**, the observed DEFF, close to the formula's 2.44.

IN R (OPTIONAL)

```
library(survey)
cb ← read.csv("data/cbhi_survey.csv")
a ← anova(lm(cbhi_member ~ village_id, data = cb)); a
MSB ← a[1, 3]; MSW ← a[2, 3]; m ← 25
c(sigma2_b = (MSB - MSW) / m,
  ICC = (MSB - MSW) / (MSB + (m - 1) * MSW))
# households treated as SRS
srs ← svydesign(ids = ~1, data = cb)
# villages as clusters
clu ← svydesign(ids = ~village_id, data = cb)
svymean(~cbhi_member, srs); svymean(~cbhi_member, clu)
# observed DEFF
(SE(svymean(~cbhi_member, clu)) / SE(svymean(~cbhi_member, srs)))^2
```

```
Response: cbhi_member
            Df Sum Sq Mean Sq F value    Pr(>F)
village_id  19   9.64 0.50737  2.5539 0.0003532 ***
Residuals  480  95.36 0.19867
  sigma2_b        ICC
0.01234807 0.05851757
            mean     SE
cbhi_member  0.3 0.0205
            mean     SE
cbhi_member  0.3 0.0319
            cbhi_member
cbhi_member    2.411208
```

The mean squares give the ICC of the Worked Calculation. `svydesign(ids = ~1)` treats the 500 households as a simple random sample (SE 0.0205); `ids = ~village_id` declares the villages as clusters, and the SE rises to 0.0319. The squared ratio is the observed DEFF, 2.41. (R also prints a warning that no weights were given; here every household had the same chance of selection, so none are needed.)

MUST-KNOW NOTES (MD-LEVEL)

- DEFF inflates the **variance**; SE_cluster = SE_SRS × √DEFF (DEFT).
- Effective n = n ÷ DEFF. Plan cluster surveys with DEFF (recall §6.2.6, §7.3); power for comparative cluster designs is in §15.6 (sample size and power for comparing groups).
- DEFF is not an analysis method. For adjusted associations use §15.2.4–15.2.6.

CHECKPOINT 15.3

1. A survey samples 800 households from 40 villages (20 households per village) with an ICC of 0.10. Compute the DEFF, the effective sample size and the factor by which standard errors grow.
2. One-way ANOVA of a 0/1 outcome across 30 anganwadi centres, 10 children each, gives MSB = 0.40 and MSW = 0.20. Estimate the ICC.
3. A colleague reports only the inflated variance from a DEFF and stops there. What is missing from the analysis of the exposure–outcome association?

(Answers in the Appendix at the end of the book.)

15.2.3 Random Effects – The General Idea

DEFINITION

A **random effect** lets a parameter, usually an intercept and sometimes a slope, differ from cluster to cluster, with the cluster-specific values treated as draws from a common distribution, typically Normal with mean 0 and variance σ^2_u. A **fixed effect** is estimated as a separate parameter with no such shared distribution. A model with both is a **mixed-effects model**.

SIMPLIFIED EXPLANATION

Every regression so far had one intercept for everybody. A **random intercept** gives each village its own baseline, u_j, spread around the overall intercept with variance σ^2_u. Three choices are possible: one intercept for all villages (ignores clustering), 20 separate fixed intercepts (uses up 19 parameters and says nothing about villages outside the sample), or 20 intercepts drawn from one distribution (the random-effects compromise). The compromise **borrows strength**: a village with few households, or an extreme one, has its estimate pulled towards the overall average, because its own data are weak evidence. This pulling is called **shrinkage**.

"Fixed" and "random" are technical labels. They describe how the cluster values are modelled, not whether the clusters were chosen at random.

ILLUSTRATION / EXAMPLE

Figure 15.3 shows a random-intercept model: every village has the same slope for wealth but its own starting level.

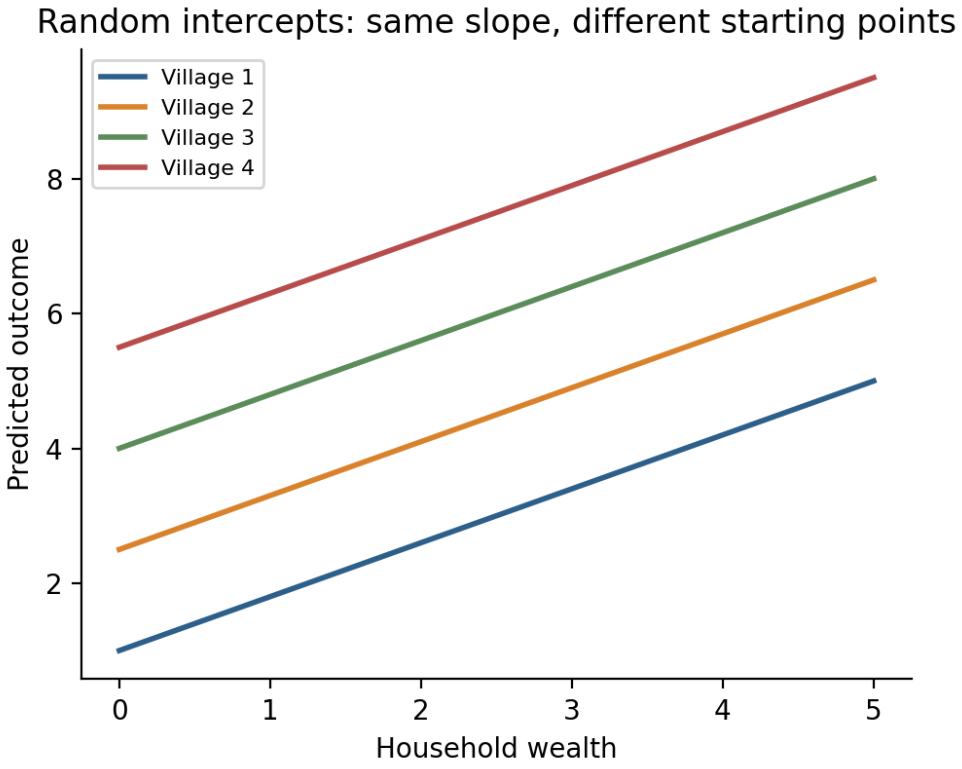


Figure 15.3. Schematic random-intercept model: parallel lines, one per village, differing only in intercept.

MUST-KNOW NOTES (MD-LEVEL)

- Random effect = cluster-specific values drawn from a shared distribution; fixed effect = separate unconnected parameters. Mixed model = both.
- Shrinkage (borrowing strength) stabilises estimates for small or extreme clusters.
- Meta-analysis uses "random effects" in the same mathematical sense, for true effects that vary between studies (§15.7.3: fixed- versus random-effects pooling).

15.2.4 Multilevel Mixed-Effects Models

DEFINITION

A **multilevel (mixed-effects) model** is a regression with fixed effects for the predictors of interest and random effects for the clustering. For a continuous outcome y of individual i in cluster j, the random-intercept model is

$y_{ij} = b_0 + b_1 x_{ij} + u_j + e_{ij}$, with $u_j \sim N(0, \sigma^2_u)$ and $e_{ij} \sim N(0, \sigma^2_e)$.

Adding a **random slope** lets the effect of x vary by cluster: $y_{ij} = (b_0 + u_{0j}) + (b_1 + u_{1j})x_{ij} + e_{ij}$.

SIMPLIFIED EXPLANATION

The model splits variation into between-cluster (σ^2_u) and within-cluster (σ^2_e) parts, exactly as the ICC does ($\sigma^2_u/(\sigma^2_u + \sigma^2_e)$ is the ICC of §15.2.1, after allowing for x). The coefficient b_1 is a **cluster-specific (conditional)** effect: the effect of x within a cluster, holding that cluster's own u_j fixed. The same idea works for binary and count outcomes (§15.2.4.1).

Repeated measures are clusters too. A person measured several times is a cluster of measurements. Recall §9.5.5: RM-ANOVA needs every subject at the same fixed times and drops anyone with a missed visit, and the summary-measure method (a slope per patient, then a one-sample t-test) is a simple valid alternative. A mixed model with a random intercept and random slope per person fits every patient's line in one model, uses every available measurement, and allows unequal or missed visits.

ILLUSTRATION / EXAMPLE

Recall the six patients of §9.5.5, systolic BP at 0, 1 and 3 months. Fit sbp = $b_0 + b_1 \times$ month with a random intercept and a random slope for each patient (Figure 15.4).

WORKED CALCULATION

1. Average slope $\hat{b}_1$ = **−2.62 mmHg per month**, SE 0.545. With complete balanced data it equals the mean of the six patients' own least-squares slopes (−3.86, −3.07, −1.00, −2.79, −4.00, −1.00; mean −2.62).
2. With only 6 patients, use t with 6 − 1 = 5 df rather than 1.96: 95% CI = −2.62 ± 2.571 × 0.545 = **−4.02 to −1.22 mmHg per month**, close to the summary-measure interval of §9.5.5 (−4.10 to −1.23).
3. Shrinkage: patient 3's own slope is −1.00; the model's estimate for patient 3 is −1.25, pulled towards the average. Patient 5's −4.00 becomes −3.90.
4. Missed visit: delete patient 5's 3-month reading. RM-ANOVA would drop patient 5 entirely. The mixed model still uses patient 5's two remaining readings and gives an average slope of −2.60.

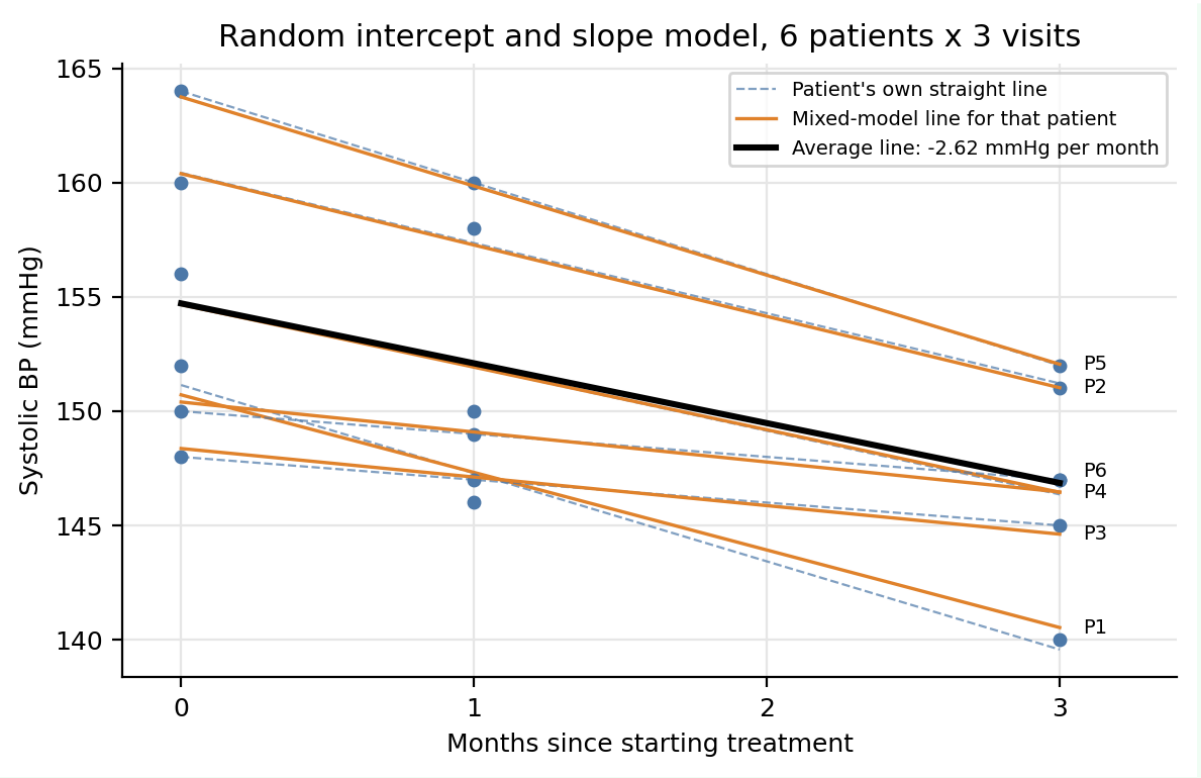


Figure 15.4. Systolic BP of the six patients of §9.5.5. Dashed: each patient's own least-squares line. Orange: that patient's line from the random intercept and slope model, pulled slightly towards the average (black, −2.62 mmHg per month).

IN R (OPTIONAL)

```
library(lme4)
bp ← data.frame(patient = rep(1:6, each = 3),
                month = rep(c(0, 1, 3), 6),
                sbp = c(152,146,140, 160,158,151, 148,147,145,
                        156,150,147, 164,160,152, 150,149,147))
# random intercept and slope
fit ← lmer(sbp ~ month + (month | patient), data = bp)
summary(fit)
# each patient's own line (shrunk)
round(coef(fit)$patient, 2)
# patient 5 misses the 3-month visit
bp2 ← bp[!(bp$patient == 5 & bp$month == 3), ]
round(fixef(update(fit, data = bp2)), 2)
```

```
...
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 patient  (Intercept) 38.717   6.222
          month        1.462   1.209    -0.73
 Residual              1.500   1.225
Number of obs: 18, groups:  patient, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept) 154.7143     2.5752  60.079
month        -2.6190     0.5451  -4.804
...
  (Intercept) month
1      150.71 -3.39
2      160.39 -3.12
3      148.37 -1.25
4      154.66 -2.74
5      163.75 -3.90
6      150.40 -1.31
(Intercept)       month
     154.71       -2.60
```

`(month | patient)` requests a random intercept and a random slope for each patient. The fixed effects are the average line: 154.7 mmHg at baseline, falling 2.62 mmHg per month. The random-effects block says baseline BP varies a lot between patients (SD 6.2 mmHg) and slopes vary less (SD 1.2 mmHg per month); the residual SD, 1.2 mmHg, is the scatter of readings around each patient's own line. `lmer` prints no p-values; with 6 patients the t-based interval of the Worked Calculation is the honest one. The last line refits without patient 5's final visit: 17 readings, slope −2.60.

MUST-KNOW NOTES (MD-LEVEL)

- Mixed model = fixed effects for predictors + random effects for clusters. Coefficients are cluster-specific (conditional).
- Repeated measures: random intercept (and slope) per person. Uses every observation, allows missed and unequal visits; with complete balanced data it agrees closely with the summary-measure method of §9.5.5.
- Random slopes let the effect of x vary between clusters.
- With few clusters (6 patients here, 20 villages below), variance components are imprecise and z-based intervals are too narrow; use t-based or small-sample methods.

15.2.4.1 Multilevel Logistic Regression

DEFINITION

Multilevel logistic regression adds random effects to logistic regression (recall §13.2.3): logit(p_ij) = b_0 + b_1x_ij + ... + u_j, with u_j ~ N(0, σ^2_u).

SIMPLIFIED EXPLANATION

Exponentiated coefficients are odds ratios, but **cluster-specific** ones: the OR comparing two households in the same village. On the logit scale the individual-level variance is fixed at $\pi^2/3 \approx 3.29$ (the variance of the standard logistic distribution), so a **latent-scale ICC** is $\sigma^2_u \div (\sigma^2_u + 3.29)$.

ILLUSTRATION / EXAMPLE

The §13.3 CBHI membership model, refitted with a random intercept for village.

WORKED CALCULATION

1. Ordinary logistic model (§13.3): distance OR 0.91 per km (0.85 to 0.97), SE of log OR 0.034.
2. Random-intercept model: distance OR **0.91** per km, SE 0.035; village variance σ^2_u = **0.020**.
3. Latent-scale ICC = 0.020 ÷ (0.020 + 3.29) = **0.006**. To compare like with like, refit with no covariates: σ^2_u = 0.285, latent ICC = 0.285 ÷ (0.285 + 3.29) = **0.080**. (The 0.06 of §15.2.1 is on the 0/1 scale, not the latent scale, so it is not directly comparable.) Distance, wealth and education, which themselves differ between villages, explain most of the village clustering; little is left for the random intercept.

IN R (OPTIONAL)

```
library(lme4)
# cb as in §15.1.2
ml ← glmer(cbhi_member ~ distance_km + wealth_index +
             maternal_education + (1 | village_id),
           data = cb, family = binomial)
round(exp(fixef(ml)), 2)
VarCorr(ml)
# latent-scale ICC
v ← as.numeric(VarCorr(ml)$village_id); v / (v + pi^2 / 3)
m0 ← update(ml, . ~ 1 + (1 | village_id))
v0 ← as.numeric(VarCorr(m0)$village_id)
# no covariates: same scale
c(v0, v0 / (v0 + pi^2 / 3))
```

```
                 (Intercept)                  distance_km
                        0.62                         0.91
               wealth_index    maternal_educationprimary
                        1.35                         1.28
maternal_educationsecondary+
                        1.86
 Groups     Name        Std.Dev.
 village_id (Intercept) 0.14195
[1] 0.006087197
[1] 0.28497373 0.07971646
```

`(1 | village_id)` adds a random intercept for village. The ORs barely move from §13.3 (wealth 1.35 against 1.34, secondary education 1.86 against 1.87). The village SD on the logit scale is 0.142 (variance 0.020), and the latent ICC 0.006. The last line is the model without covariates: village variance 0.285, latent ICC 0.080.

MUST-KNOW NOTES (MD-LEVEL)

- ORs from a multilevel logistic model are cluster-specific (conditional).
- Latent-scale ICC = $\sigma^2_u/(\sigma^2_u + \pi^2/3)$.
- Covariates that vary between clusters can absorb much of the clustering; compare the latent ICC before and after adjustment (same scale).

15.2.5 Generalised Estimating Equations

DEFINITION

Generalised estimating equations (GEE) fit a regression model for the mean of the outcome and handle within-cluster correlation through a **working correlation structure** and a **cluster-robust (sandwich) standard error**. The coefficients are **population-averaged (marginal)** effects.

SIMPLIFIED EXPLANATION

Recall the robust (sandwich) SE of §13.7: it replaces the variance the model assumes with the variance the residuals actually show. GEE uses the cluster version: residuals are added up within each village before being squared, so correlation inside a village is counted. The working correlation is a guess at how members of a cluster are correlated:

- **exchangeable:** every pair in a cluster equally correlated (households in a village);
- **independence:** no correlation assumed (the robust SE still corrects for it);
- **autoregressive (AR-1):** correlation fading with time between measurements (repeated visits).

If the working correlation is wrong, the coefficients remain consistent and the sandwich SE still valid, provided there are **enough clusters**. That proviso is the catch. The sandwich SE is estimated from the clusters, so with few clusters it is itself imprecise and tends to be too small: "hypothesis tests using the robust covariance estimator can have inflated size when the number of independent clusters is small" (Mancl & DeRouen, 2001). Their simulations covered 10–40 clusters: the robust and jackknife estimators gave tests whose size could be 2–3 times the nominal level, while their bias-corrected estimator stayed close to it. A second limit: when some measurements are missing, standard GEE is valid only if they are missing completely at random (defined in §15.4.1: missingness unrelated to any data); weighted GEE or multiple imputation (§15.4.3) is needed when they are not (Birhanu et al., 2011).

Mixed model or GEE? For a continuous outcome with an identity link, the two give very similar coefficients. For binary or count outcomes they differ: averaging a curved (logistic) relationship over villages flattens it, so a population-averaged OR is closer to 1 than a cluster-specific OR when village variation is large. Choose by the question. "What is the average effect across the district's households?" is a population-averaged question (GEE). "How do villages differ, and what is the effect within a village?" is a cluster-specific question (mixed model).

ILLUSTRATION / EXAMPLE

The CBHI membership model, fitted by GEE with exchangeable working correlation, 20 villages as clusters.

WORKED CALCULATION

1. GEE distance OR = exp(−0.0934) = **0.91** per km, robust SE 0.024. 95% CI = exp(−0.0934 ± 1.96 × 0.024) = **0.87 to 0.95**.
2. The estimated exchangeable correlation is 0.005 (on the residual scale, not the latent scale of §15.2.4.1, but just as small): after adjustment, little correlation is left.
3. The robust SE (0.024) is *smaller* than the model-based SE of the ordinary and random-intercept models (0.034 and 0.035). With only 20 villages, that is a warning, not a gain in precision (Mancl & DeRouen, 2001): report the more conservative interval, 0.85 to 0.97, or use a small-sample correction.

IN R (OPTIONAL)

```
library(geepack)
# geeglm needs each cluster's rows together
cb ← cb[order(cb$village_id), ]
g ← geeglm(cbhi_member ~ distance_km + wealth_index +
             maternal_education,
           id = factor(village_id), data = cb, family = binomial,
           corstr = "exchangeable")
summary(g)
b ← coef(g)["distance_km"]
se ← summary(g)$coefficients["distance_km", "Std.err"]
round(exp(b + c(0, -1.96, 1.96) * se), 2)
```

```
...
 Coefficients:
                            Estimate  Std.err   Wald Pr(>|W|)
(Intercept)                -0.47683  0.30703  2.412   0.1204
distance_km                -0.09344  0.02396 15.206 9.64e-05 ***
wealth_index                0.29691  0.05928 25.083 5.49e-07 ***
maternal_educationprimary   0.24508  0.29098  0.709   0.3997
maternal_educationsecondary+  0.61430  0.27053  5.156   0.0232 *
...
Estimated Correlation Parameters:
      Estimate Std.err
alpha 0.004553 0.01553
Number of clusters:   20  Maximum cluster size: 25
[1] 0.91 0.87 0.95
```

`id` names the cluster and `corstr` the working correlation. `Std.err` is the cluster-robust SE; `Wald` is (estimate/SE)^2, compared with χ^2 on 1 df. `alpha` is the estimated within-village correlation (0.005). The last line gives the OR per km and its robust 95% CI.

MUST-KNOW NOTES (MD-LEVEL)

- GEE: population-averaged coefficients, working correlation, cluster-robust SE (the clustered version of §13.7's sandwich).
- Robust to a wrong working correlation **only with enough clusters**. With few clusters the sandwich SE is too small (Mancl & DeRouen, 2001): use a small-sample correction, and distrust a robust SE smaller than the model-based one.

- Standard GEE needs missing observations to be missing completely at random (Birhanu et al., 2011).
- Five different things: cluster sampling (a design), ICC (a descriptive measure), DEFF (a precision and planning measure), mixed model (cluster-specific analysis), GEE (population-averaged analysis).

CHECKPOINT 15.4

1. A programme manager wants to know which villages have unusually low baseline CBHI membership so that she can target them. Would a mixed-effects model or GEE better suit this goal, and why?
2. Explain why "fixed" and "random" effects are not statements about whether something was chosen at random.
3. A cluster trial randomises 8 PHCs. GEE with exchangeable correlation gives a log OR of −0.40 for the intervention, with a robust SE of 0.10; the model-based SE is 0.16. Compute the 95% CI for the OR with each SE. Which SE should worry you, and what would you do?

(Answers in the Appendix at the end of the book.)

15.2.6 Analysing Complex Survey Data – Weights, Strata and PSUs

DEFINITION

A **complex survey** combines stratification, clustering and unequal selection probabilities, as NFHS does. Three design variables must enter the analysis:

- the **sampling weight** of each respondent, $w = 1/\pi$, where π is the respondent's probability of selection;
- the **stratum**, the separately sampled population subgroup (recall §6.2.3);
- the **primary sampling unit (PSU)**, the cluster selected at the first stage (recall §6.2.5).

Design-based analysis uses all three to compute estimates that represent the population and standard errors that reflect how the sample was drawn.

SIMPLIFIED EXPLANATION

Weights fix representativeness. Surveys often over-sample small groups (urban areas, a small district) so that each has enough respondents. The DHS guide, the manual for NFHS-type data, warns: "If weights are not used, all calculations will be biased toward the levels and relationships in the over-sampled strata", and explains that "sample weights are inversely proportional to the probability of selection" (DHS Program, n.d.).

Strata and PSUs fix the standard error. Stratification usually lowers it and clustering raises it (recall §15.2.2). The same guide notes that analysis requires "three pieces of information – the primary sampling unit or cluster variable, the stratification variable, and the weight variable".

Subpopulations: do not subset the data first. To estimate anaemia among urban women, keep the whole design and restrict the calculation, rather than deleting rural women and re-declaring the design. Survey

software does this; R's survey package, for example, will "restrict a survey design to a subpopulation, keeping the original design information about number of clusters, strata" (Lumley, 2004, and package documentation). Deleting rows first can drop whole PSUs or strata, which changes the degrees of freedom and, in some designs, the SE.

DERIVATION

Why $w = 1/\pi$. A woman selected with probability $\pi = 1/375$ is one of 375 women in the population with the same chance. Her answer stands for all 375. So:

1. Estimated population size = $\Sigma\, w_i$ (sum of the weights over the sample).
2. Estimated number anaemic = $\Sigma\, w_i y_i$ ($y_i = 1$ if anaemic).
3. Weighted prevalence = $\Sigma\, w_i y_i \div \Sigma\, w_i$.

The unweighted mean $\Sigma y_i / n$ gives every sampled woman the same influence, so over-sampled strata count too much.

ILLUSTRATION / EXAMPLE

data/nfhs_extract.csv is a synthetic NFHS-style extract: 640 women aged 15–49 in 2 districts, each split into urban and rural strata (4 strata), with 8 PSUs of 20 women per stratum (32 PSUs). Every stratum has 160 women in the sample, although the urban strata are much smaller in the population, so urban women are over-sampled.

Stratum	Women in population	Sampled	Weight w	Anaemic (sample)
District A urban	60,000	160	375	67 (41.9%)
District A rural	140,000	160	875	95 (59.4%)
District B urban	30,000	160	187.5	76 (47.5%)
District B rural	170,000	160	1,062.5	86 (53.8%)
Total	400,000	640		324 (50.6%)

WORKED CALCULATION

1. Unweighted prevalence = 324 ÷ 640 = **50.6%**.
2. $\Sigma\, w_i = 160 \times (375 + 875 + 187.5 + 1{,}062.5) = 400{,}000$, the population total.
3. $\Sigma\, w_i y_i = 375 \times 67 + 875 \times 95 + 187.5 \times 76 + 1{,}062.5 \times 86 = 25{,}125 + 83{,}125 + 14{,}250 + 91{,}375 = 213{,}875$.
4. Weighted prevalence = 213,875 ÷ 400,000 = **53.5%**. The unweighted figure is too low because urban women, who are less often anaemic here, are 50% of the sample but only 22.5% of the population.
5. Naive SE = $\sqrt{(0.506 \times 0.494 \div 640)} = 1.98$ percentage points. Design-based SE = **2.92** points; 95% CI **47.8% to 59.2%**; DEFF **2.19**.
6. Subpopulations, from the full design: urban **43.8%**, rural **56.3%**.
7. Degrees of freedom for design-based t-tests = PSUs − strata = 32 − 4 = 28, not 640 − 1: the design has 32 independent units, not 640.

IN R (OPTIONAL)

```r
library(survey)
nf ← read.csv("data/nfhs_extract.csv")
nf$education ← factor(nf$education,
                      levels = c("none", "primary", "secondary+"))
# unweighted: wrong
mean(nf$anaemic)
des ← svydesign(ids = ~psu, strata = ~stratum, weights = ~weight,
                data = nf, nest = TRUE)
p ← svymean(~anaemic, des, deff = TRUE); p; confint(p)
# subpopulations, not subsets
svyby(~anaemic, ~residence, des, svymean)
fit ← svyglm(anaemic ~ education + residence, design = des,
             family = quasibinomial())
round(exp(cbind(OR = coef(fit), confint(fit))), 2)
```

```
[1] 0.50625
            mean       SE   DEff
anaemic 0.534687 0.029166 2.1884
             2.5 %    97.5 %
anaemic 0.4775222 0.5918528
      residence   anaemic         se
rural     rural 0.5629032 0.03293289
urban     urban 0.4375000 0.06273764
                      OR 2.5 % 97.5 %
(Intercept)         1.63  1.02   2.59
educationprimary    0.86  0.51   1.47
educationsecondary+ 0.55  0.30   1.03
residenceurban      0.69  0.37   1.28
```

`svydesign` declares the PSUs (`ids`), strata and weights; `nest = TRUE` tells R that PSU labels restart within strata. `svymean` gives the weighted prevalence (53.5%), its design-based SE (2.92 points) and the DEFF (2.19). `svyby` estimates each subpopulation from the full design. `svyglm` fits a survey-weighted logistic regression (the `quasibinomial` family avoids a warning about non-integer weighted counts; the estimates are the same as `binomial`): women with secondary education have 0.55 times the odds of anaemia of women with none, adjusted for residence, but the 95% CI (0.30 to 1.03) includes 1. An unweighted `glm` on the same data gives a similar OR, 0.56, with a CI of 0.37 to 0.86 that excludes 1, because it ignores both the weights and the clustering.

MUST-KNOW NOTES (MD-LEVEL)

- Weight = 1/π; weighted estimate = Σwy/Σw. Without weights, estimates are biased towards over-sampled strata (DHS Program, n.d.).
- Declare PSU, stratum and weight together. Design df = PSUs − strata.
- Estimate subpopulations with the software's subpopulation (domain) option on the full design, not by deleting rows first (Lumley, survey package documentation).
- Design-based analysis estimates population-averaged quantities, like GEE; it answers "what is true of this population?", not "what happens within one cluster?".

CHECKPOINT 15.5

1. A block survey samples 100 women from each of two strata: a town of 5,000 women and rural areas with 45,000. Compute each stratum's weight. If 30 town women and 50 rural women are anaemic, compute the unweighted and weighted prevalence.
2. An NFHS analysis of 1,200 children in 60 PSUs across 6 strata reports a naive SE of 1.2 percentage points and a design-based SE of 1.8. What is the DEFF, and how many degrees of freedom does the design have?
3. A student wants the prevalence of anaemia among adolescent girls and deletes all other women from the dataset before declaring the survey design. What should she do instead, and why?

(Answers in the Appendix at the end of the book.)

15.3 Confounding Handling Techniques

DEFINITION

Confounding handling techniques prevent or remove the distortion of an exposure–outcome association by a confounder (recall §11.6). At the design stage: **restriction**, **matching** and randomisation (§15.9). At the analysis stage: **stratification**, **multivariable adjustment** and **propensity-score methods**.

SIMPLIFIED EXPLANATION

Confounding is a feature of the world, not an error of conduct, so most of it can be handled after data collection, provided the confounder was measured. Which variables count as confounders is decided from a causal diagram (recall §11.6.1), not from the data. The techniques differ in how many confounders they can handle and what they cost.

ILLUSTRATION / EXAMPLE

Does CBHI membership increase use of a health facility? Household wealth affects both joining and facility use. A researcher could restrict the study to one wealth band; match each member to a non-member of similar wealth; stratify by wealth and pool the strata; put wealth in a logistic model beside membership (recall §12.4.1, §13.2.5); or build a propensity score from wealth, education and other confounders (§15.3.1–15.3.2).

MUST-KNOW NOTES (MD-LEVEL)

- **Restriction** to a single level of a variable removes confounding by that variable, at the cost of sample size and generalisability. Restriction to a band (such as one wealth band) can leave residual confounding within the band.
- **Matching** in a case-control study makes a matched analysis necessary: the matched-pair OR b/c (recall §10.8) or conditional logistic regression (recall §13.2.9). In the matched EBF case-control study of §10.8, the matched OR was 2.4 and the unmatched OR 2.08. A matched variable's own effect cannot then be studied.

- **Stratification** with Mantel–Haenszel pooling (recall §11.7) handles one or two confounders; the EBF–diarrhoea RR fell from a crude 3.54 to RR_MH 2.93 after stratifying by PHC. With several confounders the strata multiply and empty out.
- **Multivariable adjustment** handles many confounders at once, if the model's form is right (recall §12.4.5).
- **Propensity scores** condense many confounders into one number (§15.3.1).
- None of these handles an unmeasured confounder. Only randomisation balances unmeasured variables as well as measured ones, on average across possible allocations (§15.9: randomised trials).

15.3.1 Propensity Scores – The Idea

DEFINITION

The **propensity score** is a subject's probability of receiving the exposure (being in the exposed group), given measured baseline covariates X:

$e(X) = P(E = 1 \mid X)$

It is the same quantity for exposed and unexposed subjects: for a non-member household, it is the probability that a household like it would have joined, not the probability of the (non-)membership it actually has. In Austin's (2009) words, "the propensity score is a subject's probability of treatment, conditional on observed baseline covariates." It is not the probability of the outcome, and it is not an effect estimate.

SIMPLIFIED EXPLANATION

Multivariable adjustment models the *outcome*. A propensity score models the *exposure* instead. Fit a logistic regression with exposure (CBHI membership) as the dependent variable and the confounders as predictors; each household's fitted probability is its score. The property that makes this useful: "conditional on the true propensity score, treated and untreated subjects have similar distributions of observed baseline covariates" (Austin, 2009). Two households with the same score may differ in wealth and education individually, but among all households with that score, members and non-members have the same *distribution* of wealth and education. Matching or weighting on one number then balances all the covariates in it.

Which covariates go in: decide from the DAG, before fitting. Recall §11.6.1: adjust for a set that blocks every backdoor path, and never for a descendant of the exposure.

- **Include** confounders (causes of both exposure and outcome).
- **Include** variables that affect the outcome but not the exposure. In simulations, such variables "should always be included in a PS model. The inclusion of these variables will decrease the variance of an estimated exposure effect without increasing bias" (Brookhart et al., 2006).
- **Avoid** variables that affect the exposure but not the outcome (instrument-like variables). "Including variables that are related to the exposure but not to the outcome will increase the variance of the estimated exposure effect without decreasing bias" (Brookhart et al., 2006). An example: which ASHA enrolled the village, if the ASHA affects joining but not facility use by any other route.

- **Exclude** mediators and colliders. A mediator lies on the causal path *after* the exposure, so what matters is when a variable was measured: a variable measured after enrolment, such as a household's satisfaction with CBHI services, may be a consequence of membership and must be left out. A collider need not come after the exposure: a pre-exposure variable that is a common effect of two unmeasured causes, one of the exposure and one of the outcome, satisfies the classical three-condition test of §11.6, yet adjusting for it opens a non-causal path (this is why §11.6.1 prefers the backdoor criterion).

A logistic regression cannot discover which variables are confounders. It only estimates a probability from the variables a human chose.

Positivity (overlap). For every combination of covariates, both exposure groups must be possible. If every poor household were a non-member, the data could never show what membership does for a poor household, and no statistics could supply it. Check overlap by plotting the scores of both groups before using them. Poor overlap shows up as unmatched subjects, extreme weights, or an answer that describes only a narrow slice of the sample.

ILLUSTRATION / EXAMPLE

In the CBHI survey, the DAG drawn by the study team treats distance to the facility, the wealth index, maternal education and household size, all recorded as pre-existing household characteristics, as the adjustment set for the question "does membership increase facility use in the past month?". The propensity model is the logistic regression of membership on these four.

Figure 15.5 (left) shows the scores. Both groups span most of the range: members from 0.09 to 0.77, non-members from 0.04 to 0.79. But the tails are thin. Below about 0.15 there are many non-members and few members; above 0.60 there are 26 members but only 10 non-members. The overlap is adequate in the middle and sparse at the top, which §15.3.2 will pay for. The right panel is a simulated contrast in which one covariate almost fixes exposure: most exposed subjects have scores near 1 and most unexposed near 0, and there is almost no common ground to compare.

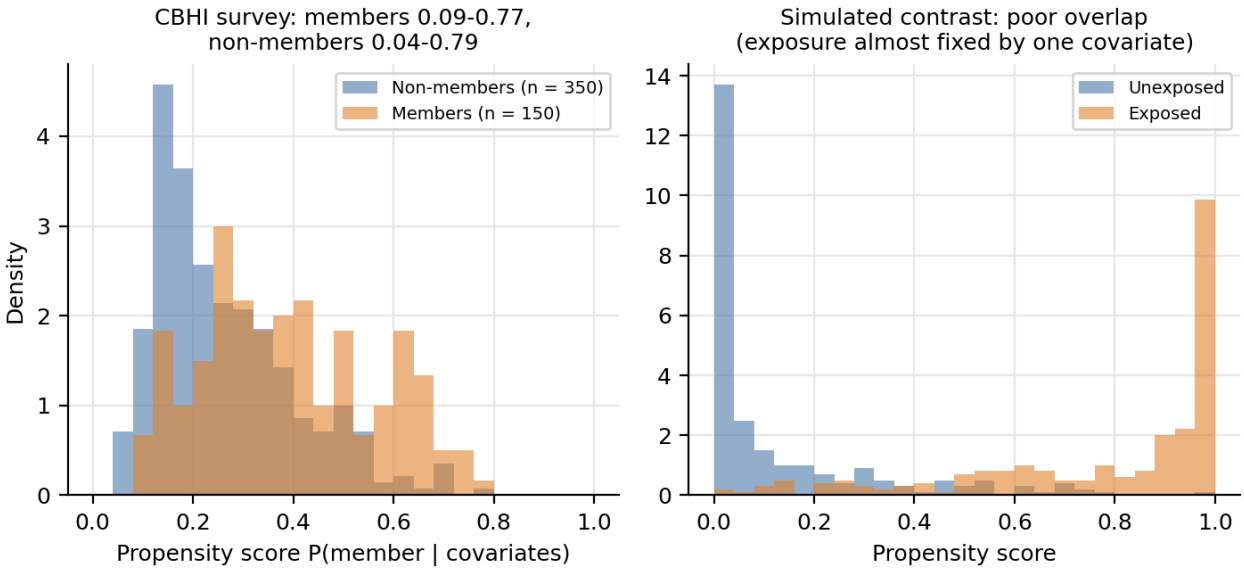


Figure 15.5. Left: propensity scores for CBHI membership (logistic model on distance, wealth, education and household size; 150 members, 350 non-members). Overlap is adequate between about 0.15 and 0.6, but thin above 0.6, where members outnumber non-members 26 to 10. Right: simulated data (illustrative only) with poor overlap, exposure being nearly determined by one covariate.

MUST-KNOW NOTES (MD-LEVEL)

- Propensity score = P(exposed | measured covariates), for everyone. Not the outcome probability, not an effect, and no help against unmeasured confounding.
- Choose covariates from the DAG: include confounders and outcome-only predictors; avoid exposure-only predictors, mediators (anything measured after exposure that it may affect) and colliders (Brookhart et al., 2006).
- Check positivity (overlap) before using the scores.

15.3.2 Using Propensity Scores – Matching, Weighting, and Balance

DEFINITION

A propensity score is used in one of three main ways:

- **Matching:** pair each exposed subject with an unexposed subject of similar score, commonly within a **caliper** (a maximum allowed difference); unmatched subjects are dropped.
- **Inverse probability of treatment weighting (IPW):** keep everyone and give each subject a weight equal to "the inverse of the probability of receiving the treatment that the subject actually received" (Austin, 2011a): 1/e(X) for the exposed, 1/[1 − e(X)] for the unexposed.
- **Stratification:** group subjects by quintiles of the score and compare within each group.

Balance is then checked with the **standardised mean difference (SMD)** of each covariate:

$$SMD = (\bar{x}_E - \bar{x}_U) \div \sqrt{[(s^2_E + s^2_U)/2]}$$

and for a binary covariate with proportions p_E and p_U, $SMD = (p_E - p_U) \div \sqrt{\{[p_E(1 - p_E) + p_U(1 - p_U)]/2\}}$ (Austin, 2009).

SIMPLIFIED EXPLANATION

Matching. Nearest-neighbour matching within a caliper of 0.2 standard deviations of the logit of the score is a common choice; it "minimized the mean squared error of the estimated treatment effect in several scenarios" (Austin, 2011a).

Weighting. The Derivation shows why the weights balance covariates. With these weights the comparison describes the whole sample, the **average treatment effect (ATE)**.

Stratification. Stratifying on quintiles of the score "eliminates approximately 90% of the bias due to measured confounders when estimating a linear treatment effect" (Austin, 2011a, summarising work by Rosenbaum and Rubin that extends a result of Cochran for a single confounder).

Balance, not p-values. The SMD expresses a difference in units of pooled SD, so it can be compared across variables and "is not influenced by sample size" (Austin, 2009). A t-test after matching is influenced by it: matching shrinks the sample, so a covariate that is still imbalanced can give p = 0.6. A common convention treats |SMD| < 0.1 as negligible imbalance; Austin (2009) notes there is "no clear consensus" and that "some researchers have proposed that a standardized difference of 0.1 (10 per cent) denotes meaningful imbalance".

The estimand. Say which effect, in which population. Matching exposed subjects to controls estimates the **average treatment effect in the treated (ATT)**: "the average effect of treatment on those subjects who ultimately received the treatment" (Austin, 2011a). If some exposed subjects cannot be matched, it is the effect among the *matched* exposed only. The IPW weights above estimate the ATE.

The outcome analysis is a separate decision. After 1:1 matching, should the analysis respect the pairs? Austin's simulations favour paired methods: they gave "empirical type I error rates that were closer to the advertised rate" and "narrower 95 per cent confidence intervals" (Austin, 2011b). Stuart (2010) records that "there is debate about whether the analysis needs to account for the matched pair nature of the data" and gives reasons why it may not be necessary. State and justify your choice; a reasonable default for a binary outcome is a paired analysis (McNemar test and matched-pair OR b/c, recall §10.8), with an unpaired analysis as a sensitivity check. After weighting, use a weighted analysis with a robust SE (recall §13.7); after stratification, pool the strata as in §11.7.

DERIVATION

Why IPW balances covariates. Take the N_x households in the sample that share covariate values x, with score e = e(x).

1. About N_x × e of them are members; each gets weight 1/e. Weighted number of members = N_x × e × (1/e) = N_x.
2. About N_x × (1 − e) are non-members; each gets weight 1/(1 − e). Weighted number = N_x.
3. So in the weighted data, members and non-members each have N_x households with covariates x, for every x: both groups have the covariate distribution of the whole sample.
4. The weighted outcome comparison therefore estimates the effect in the whole population (the ATE). When e is close to 0 for a member (or close to 1 for a non-member), the weight explodes: this is the positivity problem in numbers.

ILLUSTRATION / EXAMPLE

Research question: is CBHI membership associated with facility use in the past month, after balancing distance, wealth, maternal education and household size? Crude: 50 of 150 members (33.3%) and 56 of 350 non-members (16.0%) used a facility, a crude risk difference of 17.3 percentage points.

WORKED CALCULATION

1. **SMD before matching, distance:** members 6.41 km (SD 3.42), non-members 8.12 km (SD 3.03). SMD = (6.41 − 8.12) ÷ √[(3.42² + 3.03²)/2] = −1.71 ÷ √10.46 = −1.71 ÷ 3.23 = **−0.53**. Wealth: **0.70**. Secondary education: 46.0% against 24.6%, SMD = 0.214 ÷ √[(0.46 × 0.54 + 0.246 × 0.754)/2] = 0.214 ÷ 0.466 = **0.46**. Large imbalances.
2. **Matching:** 1:1 nearest neighbour on the logit of the score, caliper 0.2 SD, without replacement. **133** of 150 members find a partner. The 17 unmatched members all have scores between 0.59 and 0.70: the thin upper tail of Figure 15.5.
3. **Balance after matching:** SMDs for distance −0.06, wealth −0.03, secondary education 0.06, no education 0.02, household size 0.04. All below 0.1 in absolute value.
4. **Paired outcome table** (133 pairs): both used a facility 8; member only 35; non-member only 16; neither 74. Facility use: members 43/133 = 32.3%, matched non-members 24/133 = 18.0%.

5. **Paired risk difference** = (35 − 16)/133 = **0.143**. SE = √[35 + 16 − (35 − 16)²/133] ÷ 133 = √48.29 ÷ 133 = 0.0522; 95% CI 0.143 ± 1.96 × 0.0522 = **0.040 to 0.245**. Matched-pair OR = 35/16 = **2.19**; McNemar $\chi^2 = (35 - 16)^2 \div 51 = 7.08$, p = **0.008**.
6. **Estimand:** ATT among the 133 matched members. It does not describe the 17 unmatched members, all from the upper tail of the score (0.59 to 0.70).
7. **IPW (all 500 households):** weights range from 1.05 to 11.4; after weighting all SMDs are below 0.03. Weighted facility use: members 27.6%, non-members 15.1%, a difference of **12.5 percentage points** (ATE).
8. **Limitation:** trust in the formal health system was not measured. If it drives both joining and facility use, none of these estimates removes it.

IN R (OPTIONAL)

```
cb ← read.csv("data/cbhi_survey.csv")
cb$sec  ← as.integer(cb$maternal_education == "secondary+")
cb$none ← as.integer(cb$maternal_education == "none")
ps_fit ← glm(cbhi_member ~ distance_km + wealth_index +
               maternal_education + household_size,
             family = binomial, data = cb)
# P(member | covariates), for every household
cb$ps ← fitted(ps_fit)
tapply(cb$ps, cb$cbhi_member, range)              # overlap check
smd ← function(x, t) (mean(x[t == 1]) - mean(x[t == 0])) /
  sqrt((var(x[t == 1]) + var(x[t == 0])) / 2)
vars ← c("distance_km", "wealth_index", "sec", "none",
         "household_size")
# 1:1 nearest neighbour on logit(ps), caliper 0.2 SD
lps ← qlogis(cb$ps); cal ← 0.2 * sd(lps)
tr ← which(cb$cbhi_member == 1); tr ← tr[order(-cb$ps[tr])]
pool ← which(cb$cbhi_member == 0); pairs ← NULL
for (i in tr) {
  j ← pool[which.min(abs(lps[pool] - lps[i]))]
  if (abs(lps[j] - lps[i]) ⩽ cal) {
    pairs ← rbind(pairs, c(i, j)); pool ← setdiff(pool, j)
  }
}
nrow(pairs)                                       # members matched
md ← cb[c(pairs), ]
round(rbind(
  before = sapply(vars, function(v) smd(cb[[v]], cb$cbhi_member)),
  after  = sapply(vars, function(v) smd(md[[v]], md$cbhi_member))), 2)
y ← cb$facility_use_past_month
table(member = y[pairs[, 1]], matched_nonmember = y[pairs[, 2]])
tab ← table(y[pairs[, 1]], y[pairs[, 2]])
b ← tab["1", "0"]; c ← tab["0", "1"]; n ← nrow(pairs)
se_rd ← sqrt(b + c - (b - c)^2 / n) / n
rd ← (b - c) / n
round(c(RD = rd, lower = rd - 1.96 * se_rd, upper = rd + 1.96 * se_rd,
        OR_pairs = b / c), 3)
mcnemar.test(tab, correct = FALSE)$p.value
# inverse-probability weighting
library(survey)
cb$w ← ifelse(cb$cbhi_member == 1, 1 / cb$ps, 1 / (1 - cb$ps))
summary(cb$w)
dw ← svydesign(ids = ~village_id, weights = ~w, data = cb)
svyby(~facility_use_past_month, ~cbhi_member, dw, svymean)
```

```
$`0`
[1] 0.04267006 0.79124671
$`1`
[1] 0.08761804 0.77169950
[1] 133
       distance_km wealth_index  sec  none household_size
before       -0.53         0.70 0.46 -0.41           0.06
after        -0.06        -0.03 0.06  0.02           0.04
      matched_nonmember
member  0  1
     0 74 16
     1 35  8
      RD    lower    upper OR_pairs
   0.143    0.040    0.245    2.188
[1] 0.007801728
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.045   1.218   1.460   1.994   2.083  11.413
  cbhi_member facility_use_past_month         se
0           0               0.1514145 0.01845493
1           1               0.2762926 0.04267730
```

The first block is the overlap check (ranges by group). The loop is greedy nearest-neighbour matching, members with the highest scores first; 133 are matched. The SMD table shows balance before and after matching. The 2 × 2 table of pairs gives the paired risk difference, its CI, the matched-pair OR and the McNemar p-value of the Worked Calculation. The last block is IPW: `svydesign` with the weights and villages as clusters gives weighted facility use by membership with robust SEs. (Dedicated packages exist for matching and balance tables; the base-R version shows every step.)

MUST-KNOW NOTES (MD-LEVEL)

- Matching, weighting and stratification are three uses of the score; “propensity score” does not mean “matching”.
- SMD = difference in means ÷ pooled SD. Convention: |SMD| < 0.1 = negligible imbalance (Austin, 2009; no clear consensus). Check every covariate; do not rely on post-matching p-values.
- IPW weight = 1/e for the exposed, 1/(1 − e) for the unexposed → ATE. Extreme weights signal poor overlap.
- Matching the exposed → ATT (among the matched). State the estimand.
- Paired or unpaired analysis after matching is debated (Austin, 2011b; Stuart, 2010): justify your choice and check the other.
- Say “propensity-score methods improve comparability on measured covariates”, never “propensity scores remove confounding”. Doubly robust estimation and sensitivity analysis for unmeasured confounding (§17.3.1: quantitative bias analysis and the E-value) are beyond this section.

CHECKPOINT 15.6

1. After propensity-score matching, a covariate’s p-value comparing the two groups is 0.62. A colleague concludes that balance was achieved. Why is this premature, and what should be reported instead?

2. Before matching, mean age of mothers is 27.0 years (SD 4.0) among members and 25.0 years (SD 5.0) among non-members. Compute the SMD. Is the imbalance negligible by the usual convention?
3. A team matches 1:1 on the propensity score and reports an odds ratio from ordinary (unpaired) logistic regression on the matched sample. What must the team justify, and what are the two positions in the literature?

(Answers in the Appendix at the end of the book.)

15.4 Missing Data

DEFINITION

Missing data are values that were intended to be recorded but were not. Values that do not apply (a question skipped by design) are **structurally missing** and are not missing data in this sense.

SIMPLIFIED EXPLANATION

Every method so far assumed a complete dataset. Real data lose values: skipped questions, missed visits, lost samples. Whether the losses bias the answer depends on *why* values are missing (§15.4.1), and that decides the handling (§15.4.2–15.4.3).

ILLUSTRATION / EXAMPLE

In data/cbhi_survey.csv, `sought_formal_care` is blank for the 351 households without a sick child, `reason_not_joining` is blank for the 150 members, and `satisfaction_1to5` is blank for the 350 non-members. All three are structurally missing: the question did not apply. A household that refused to state its income would be genuinely missing. In the §12.6 journey, 3 of 240 children (1.25%) lacked a wealth index and were dropped.

MUST-KNOW NOTES (MD-LEVEL)

- First separate “not applicable” from “not obtained”.
- How to handle genuinely missing values depends on the missingness mechanism (§15.4.1).
- If people with missing data differ systematically from the rest, dropping them distorts the sample in the same way as selection bias (§17.2.1: distortion that arises when remaining in the analysis depends on both exposure and outcome).

15.4.1 Types of Missingness

DEFINITION

Missingness mechanisms are classified as (Sterne et al., 2009):

- **Missing completely at random (MCAR):** the probability of being missing is unrelated to any values, observed or unobserved.

- **Missing at random (MAR):** the probability of being missing depends only on *observed* variables; within groups defined by those variables, it is unrelated to the missing values themselves.
- **Missing not at random (MNAR):** the probability of being missing depends on the unobserved value itself, even after allowing for the observed variables.

SIMPLIFIED EXPLANATION

"Random" is technical here. MAR does not mean haphazard; it means "explainable by what we did observe". The mechanism cannot be proved from the data: "it is not possible to distinguish between missing at random and missing not at random using observed data" (Sterne et al., 2009). You argue for a mechanism from knowledge of how the data were collected.

ILLUSTRATION / EXAMPLE

- **MCAR:** a haemoglobin sample is lost because a centrifuge failed, regardless of the woman.
- **MAR:** women with no schooling are more likely to leave the wealth items blank; within each education group, whether the items are blank is unrelated to the household's actual wealth.
- **MNAR:** the richest households refuse the income question because their income is high.

MUST-KNOW NOTES (MD-LEVEL)

- MCAR: unrelated to anything. MAR: explained by observed variables. MNAR: depends on the missing value itself.
- MAR versus MNAR cannot be tested from the observed data (Sterne et al., 2009); argue from how the data were collected.
- Standard GEE needs MCAR (recall §15.2.5).

15.4.2 Missing Data Handling Techniques

DEFINITION

The main approaches are **complete-case analysis** (listwise deletion: analyse only records with no missing values), **single imputation** (fill each gap with one value, such as the mean), and **multiple imputation** (§15.4.3).

SIMPLIFIED EXPLANATION

Complete-case analysis always loses precision. Whether it also loses validity depends on the mechanism and on the analysis. It is unbiased under MCAR. "When it is plausible that data are missing at random, but not completely at random, analyses based on complete cases may be biased" (Sterne et al., 2009). In a regression, however, "missing data in predictor variables also do not cause bias in analyses of complete cases if the reasons for the missing data are unrelated to the outcome" (Sterne et al., 2009). So complete cases are sometimes fine; decide case by case.

Mean imputation keeps the sample size but pretends to know what it does not. Every filled-in value sits at the centre, so the variable's variance shrinks (recall §3.6), its correlations with other variables are di-

luted, and the SE is too small because the imputed values are treated as data. If missingness is related to the value, the mean is also biased.

ILLUSTRATION / EXAMPLE

Take the CBHI wealth index (true mean −0.196) and delete 72 of the 500 values (14.4%) by a MAR mechanism: 25% of households with no maternal education lose their value, against 10.7% with primary and 7.1% with secondary education, and members are less often missing than non-members. (The deletion is simulated by computer, so the truth is known.)

WORKED CALCULATION

1. Complete-case mean of the 428 remaining values = **−0.080** (SE 0.086). It is too high by 0.116, because the poorer, less-educated households are the ones missing.
2. Mean imputation fills the 72 gaps with −0.080. The mean stays at −0.080 (still biased); the SD falls from 1.77 in the complete cases to 1.64; the naive SE becomes $1.64/\sqrt{500} = 0.073$, smaller than the complete-case SE, although no information was added.

MUST-KNOW NOTES (MD-LEVEL)

- Complete cases: unbiased under MCAR; can be biased under MAR; in regression, unbiased if missingness in predictors is unrelated to the outcome (Sterne et al., 2009). Always loses precision.
- Mean imputation shrinks variance, distorts associations and gives falsely small SEs. Avoid it for inference.
- Under MNAR no method using only the observed data is guaranteed to work; use sensitivity analysis (§15.4.3).

15.4.3 Multiple Imputation

DEFINITION

Multiple imputation (MI) fills each missing value m times with values drawn from their predictive distribution given the observed data, analyses each of the m completed datasets with the planned analysis, and combines the m results by **Rubin's rules**:

- pooled estimate $\bar{Q}$ = the mean of the m estimates;
- within-imputation variance $\bar{U}$ = the mean of the m squared SEs;
- between-imputation variance B = the variance of the m estimates;
- total variance $T = \bar{U} + (1 + 1/m) \times B$, and $SE = \sqrt{T}$ (van Buuren, *Flexible Imputation of Missing Data*, §2.3).

SIMPLIFIED EXPLANATION

A missing value is not known, but it is not unknown either: education, distance and membership narrow down what a household's wealth could be. MI draws a plausible value from that narrowed range, with deliberate randomness, so the same gap is filled differently in each completed dataset. The analysis is re-

peated in each; if the m answers agree, the missing values matter little, and if they disagree, B is large. Rubin's rules add that disagreement to the ordinary sampling variance, so "standard errors ... take account of the variability in results between the imputed datasets, reflecting the uncertainty associated with the missing values" (Sterne et al., 2009). The (1 + 1/m) term allows for using a finite number of imputations.

MI is not "guessing the right value". It works by representing the spread of plausible values honestly.

The imputation model is not the analysis model. The imputation model should include the analysis variables (including the outcome) and **auxiliary** variables that predict the missing values or their missingness, even if they play no part in the analysis. Standard MI is justified under MAR, and only if the variables that explain the missingness are in the imputation model. It cannot repair MNAR: "if individuals are more likely to miss appointments because they are depressed on the day of the appointment, then it may be impossible to make the missing at random assumption plausible, even if a large number of variables is included in the imputation model" (Sterne et al., 2009).

Sensitivity analysis. When MNAR is plausible, repeat the analysis under less convenient assumptions (for example, imputed incomes shifted upwards by a stated amount) and report whether the conclusion survives.

ILLUSTRATION / EXAMPLE

Continue the §15.4.2 example: 72 wealth values deleted under MAR. Impute m = 20 times, using maternal education, distance, membership and household size, and estimate the mean wealth index in each completed dataset.

WORKED CALCULATION

1. The 20 estimates range from −0.235 to −0.126 (Figure 15.6). $\bar{Q}$ = their mean = **−0.176**.
2. $\bar{U}$ = 0.00625 (mean of the 20 squared SEs); B = 0.000567 (variance of the 20 estimates).
3. T = 0.00625 + (1 + 1/20) × 0.000567 = 0.00625 + 0.000596 = 0.00685. SE = √0.00685 = **0.083**.
4. 95% CI: **−0.339 to −0.013**. The true mean, −0.196, lies inside; the complete-case mean, −0.080, is well away from it.
5. The fraction of missing information is about 0.09: about 9% of the information about the mean was lost to the missing values, less than the 14.4% of values missing, because the observed variables in the imputation model recover part of it.

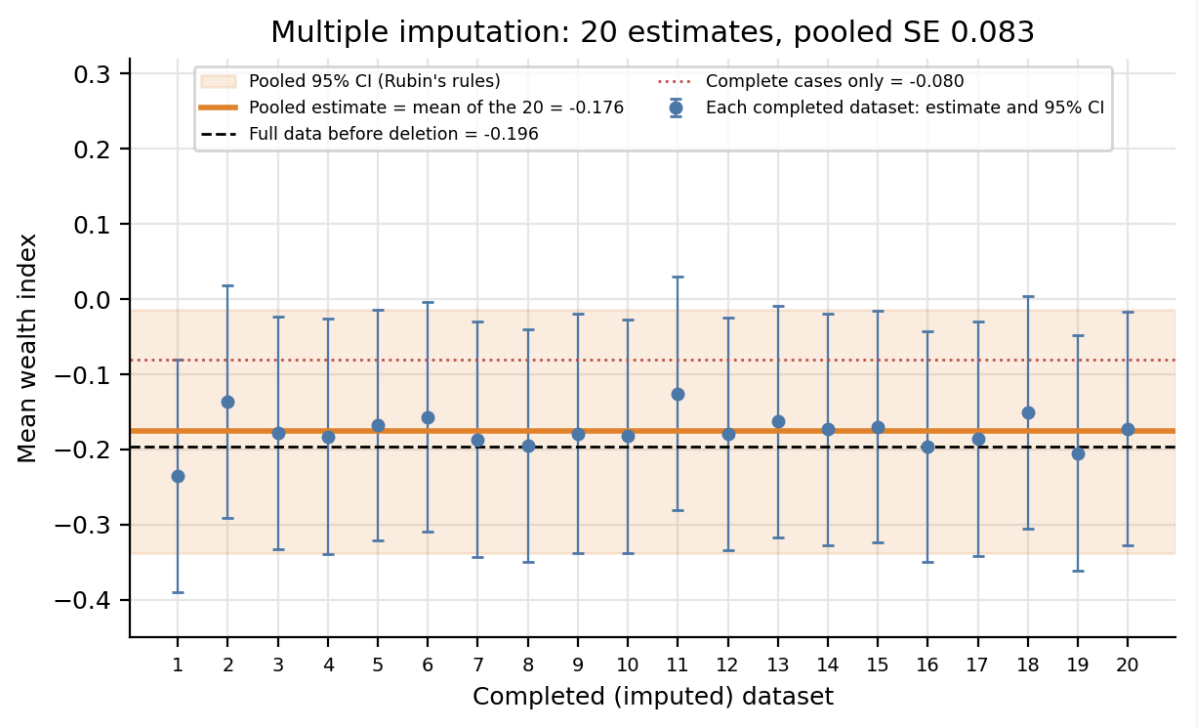


Figure 15.6. Mean wealth index estimated in each of 20 imputed datasets (points, 95% CIs). The orange line is the pooled estimate, exactly the mean of the 20 points (−0.176); the shaded band is its pooled 95% CI from Rubin's rules. Dashed: true mean before deletion (−0.196). Dotted: complete-case mean (−0.080).

IN R (OPTIONAL)

```
library(mice)
cb ← read.csv("data/cbhi_survey.csv")
cb$maternal_education ← factor(cb$maternal_education,
  levels = c("none", "primary", "secondary+"))
# the truth, before we delete anything
mean(cb$wealth_index)
# delete wealth: MAR, more often if no education
set.seed(15)
p_miss ← plogis(-2.2 + 1.5 * (cb$maternal_education == "none") -
                0.8 * cb$cbhi_member)
cb$wealth_obs ← ifelse(runif(500) < p_miss, NA, cb$wealth_index)
# complete-case mean
sum(is.na(cb$wealth_obs)); mean(cb$wealth_obs, na.rm = TRUE)
imp ← mice(cb[, c("wealth_obs", "maternal_education", "distance_km",
                  "cbhi_member", "household_size")],
           m = 20, method = "pmm", seed = 15, printFlag = FALSE)
# analysis model: the mean
fits ← with(imp, lm(wealth_obs ~ 1))
round(summary(pool(fits), conf.int = TRUE)[,
  c("estimate", "std.error", "2.5 %", "97.5 %")], 3)
round(pool(fits)$pooled[, c("ubar", "b", "t", "fmi")], 5)
```

```
[1] -0.19578
[1] 72
[1] -0.08025701
  estimate std.error  2.5 % 97.5 %
1   -0.176     0.083 -0.339 -0.013
     ubar       b       t     fmi
1 0.00625 0.00057 0.00685 0.09168
```

The first three lines are the truth (−0.196), the number deleted (72) and the biased complete-case mean (−0.080). `mice` creates 20 completed datasets by predictive mean matching (`pmm`: each gap is filled with an observed wealth value from a household whose predicted wealth is close). `with` runs the analysis in each and `pool` applies Rubin's rules: estimate −0.176, SE 0.083. The last line shows $\bar{U}$, B, T and the fraction of missing information, as in the Worked Calculation.

MUST-KNOW NOTES (MD-LEVEL)

- MI: m completed datasets → m analyses → Rubin's rules. $\bar{Q}$ = mean; $T = \bar{U} + (1 + 1/m)B$.
- Include the outcome and auxiliary variables in the imputation model; it need not match the analysis model.
- Valid under MAR with an adequate imputation model; not a remedy for MNAR (Sterne et al., 2009). Pair it with a sensitivity analysis.
- Report the amount of missing data, the mechanism assumed, the imputation model, m, and the complete-case result alongside the MI result.

CHECKPOINT 15.7

1. 15% of income values are missing, mostly among the wealthiest respondents, who are reluctant to disclose income. Will standard MAR-based multiple imputation remove the bias? What should the analyst do?
2. Five imputed datasets give estimates 2.0, 2.4, 2.2, 1.8 and 2.1 for a regression coefficient, each with SE 0.30. Apply Rubin's rules to find the pooled estimate and its SE.
3. Why might the imputation model for missing income include "survey enumerator ID", a variable that will never appear in the analysis model?

(Answers in the Appendix at the end of the book.)

15.5 Diagnostic Test Statistics

Sensitivity, specificity, predictive values and likelihood ratios (recall §5.5.1–5.5.3) judge a test at one chosen cut-off. Many tests give a continuous reading (MUAC, haemoglobin, a risk score), and the cut-off is a choice. §15.5.1 shows how a test performs over every possible cut-off at once and how to pick one. §15.5.2 applies the same tools to the predicted probabilities of a regression model and adds the checks a prediction model needs.

15.5.1 ROC Curve and AUC

DEFINITION

The **receiver operating characteristic (ROC) curve** plots a test's sensitivity against (1 − specificity) at every possible cut-off. The **area under the curve (AUC)** is the probability that a randomly chosen person with the disease has a more abnormal test value than a randomly chosen person without it. It runs from 0 to 1: 0.5 is chance, 1 is perfect separation, and a value below 0.5 means the test works in the opposite direction to the one specified.

SIMPLIFIED EXPLANATION

Each cut-off gives one pair of sensitivity and specificity (recall §5.5.1), so each cut-off is one point on the curve. A very lenient cut-off calls almost everyone positive: sensitivity near 1, specificity near 0, the top-right corner. A very strict cut-off does the reverse, the bottom-left corner. A test that separates the groups perfectly climbs straight up the left edge to sensitivity 1 before moving right. A useless test runs along the diagonal, where sensitivity equals the false-positive proportion at every cut-off.

The AUC compresses the curve into one number. It does not depend on any cut-off, so it is the natural way to compare candidate tests before a cut-off is chosen.

DERIVATION

Why is the area a probability? Take n_1 diseased and n_0 non-diseased people. Step the cut-off down one value at a time. Each time the cut-off passes a diseased person, the curve rises by $1/n_1$. Each time it passes a non-diseased person, it moves right by $1/n_0$. The area added by one step to the right equals its width, $1/n_0$, times the current height, which is the proportion of diseased people already passed, that is, with a more abnormal value than this non-diseased person. Adding over all non-diseased people:

AUC = (number of diseased–non-diseased pairs in which the diseased person is more abnormal) ÷ ($n_1 \times n_0$),

with tied pairs counted as one-half. This count is the Mann–Whitney U of §10.2 (recall: U counts the pairs in which one group's member has the larger value), so $\mathbf{AUC = U/(n_1 n_0)}$ (Hanley & McNeil, 1982). If the direction is specified the wrong way round, every "win" becomes a "loss" and the AUC becomes 1 − AUC, hence values below 0.5.

ILLUSTRATION / EXAMPLE

Mid-upper-arm circumference (MUAC) is quick to measure in an anganwadi. Suppose (hypothetical data) 20 children aged 6–59 months also have their weight-for-height measured, and 10 of them have severe acute malnutrition (SAM) by that reference standard. Low MUAC means "positive".

	MUAC (cm)
SAM (n = 10)	10.8, 11.2, 11.4, 11.6, 11.9, 12.0, 12.3, 12.6, 12.9, 13.4
Not SAM (n = 10)	11.8, 12.4, 12.7, 13.0, 13.2, 13.5, 13.7, 14.0, 14.3, 14.8

WORKED CALCULATION

1. Pairs: 10 × 10 = 100. Count the pairs in which the SAM child has the *lower* MUAC. For each non-SAM child, count the SAM children below it: 11.8 cm has 4 below it (10.8, 11.2, 11.4, 11.6); 12.4 has 7; 12.7 has 8; 13.0 and 13.2 have 9 each; the five from 13.5 upwards have 10 each. Total 4 + 7 + 8 + 9 + 9 + 50 = **87**, no ties.
2. AUC = 87/100 = **0.87**.
3. One cut-off: "positive if MUAC ≤ 12.6 cm". Sensitivity = 8/10 = 0.8 (SAM children at or below 12.6). Specificity = 8/10 = 0.8 (non-SAM children above 12.6: all but 11.8 and 12.4). The point is (0.2, 0.8).
4. Youden's J = Se + Sp − 1 (recall §5.5.3) at each cut-off. It peaks at **0.6** at three cut-offs: ≤ 12.3 cm (Se 0.7, Sp 0.9), ≤ 12.6 cm (0.8, 0.8) and ≤ 12.9 cm (0.9, 0.7).
5. J cannot choose among them, because J weights a missed case and a false alarm equally. A missed SAM child can die; a false positive costs a referral visit. A screening programme would take **≤ 12.9 cm** (Se 0.9). A programme with scarce referral places might take ≤ 12.3 cm (Sp 0.9, LR+ = 0.7/0.1 = 7.0).

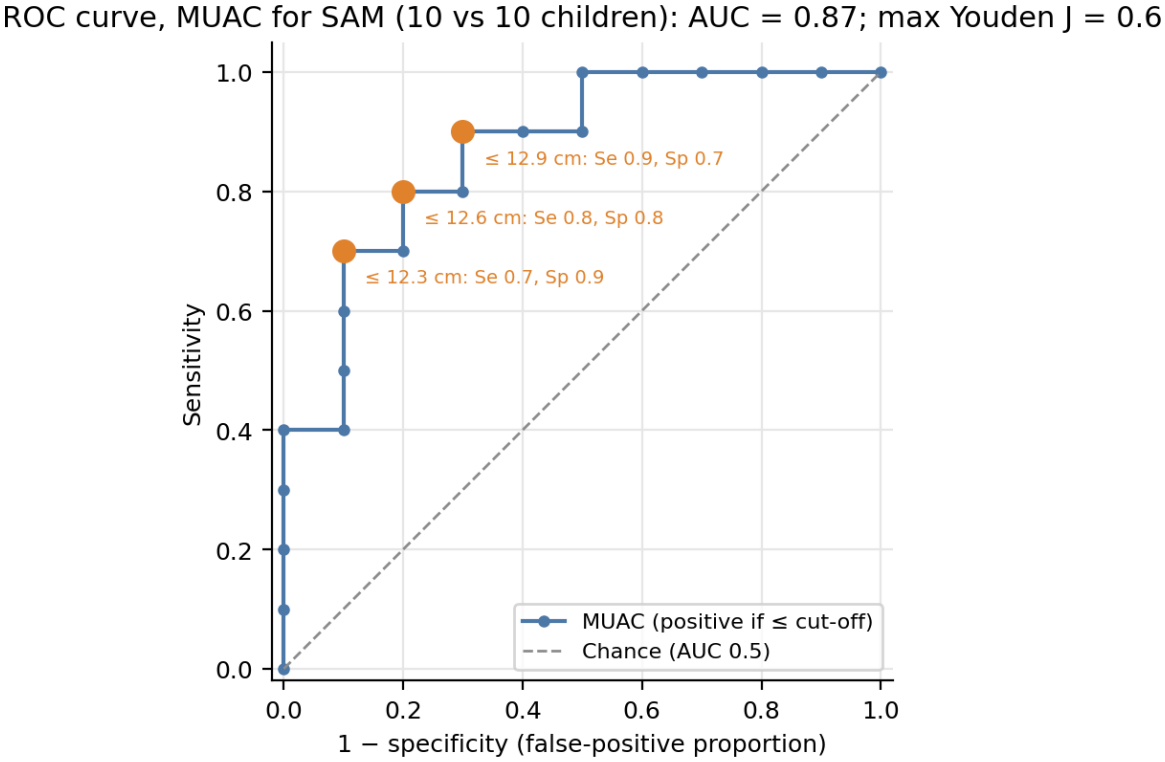


Figure 15.7. ROC curve for MUAC as a screen for SAM (hypothetical 10 vs 10 children). Each step is one cut-off. The three orange points share the maximum Youden J of 0.6. AUC = 0.87.

IN R (OPTIONAL)

```
library(pROC)
# first 10 SAM, last 10 not SAM
muac ← c(10.8, 11.2, 11.4, 11.6, 11.9, 12.0, 12.3, 12.6, 12.9, 13.4,
         11.8, 12.4, 12.7, 13.0, 13.2, 13.5, 13.7, 14.0, 14.3, 14.8)
sam  ← rep(1:0, each = 10)
# low MUAC = positive
r ← roc(sam, muac, direction = ">", levels = c(0, 1))
auc(r); ci.auc(r)
coords(r, "best", best.method = "youden")
# direction reversed
auc(roc(sam, muac, direction = "<", levels = c(0, 1)))
```

```
Area under the curve: 0.87
95% CI: 0.7124-1 (DeLong)
  threshold specificity sensitivity
1     12.95         0.7         0.9
2     12.65         0.8         0.8
3     12.35         0.9         0.7
Area under the curve: 0.13
```

`direction = ">"` tells `roc` that the non-SAM group has the higher values. The AUC of 0.87 is step 2. Its DeLong 95% CI runs from 0.71 to 1 (the upper limit is cut at 1): with only 20 children, the AUC is imprecise. `coords` returns the three tied Youden cut-offs of step 4; `pROC` reports each threshold as the midpoint between two observed values (12.95 lies between 12.9 and 13.0). The last line specifies the direction wrongly and gets 1 − 0.87 = 0.13.

MUST-KNOW NOTES (MD-LEVEL)

- **AUC runs from 0 to 1.** 0.5 is chance. Below 0.5 the test is informative but the "positive" direction was specified the wrong way round; reversing it gives 1 − AUC.
- **Rough labels are a convention.** An often-quoted scale, from Hosmer and Lemeshow's logistic regression textbook, calls 0.7–0.8 "acceptable", 0.8–0.9 "excellent" and above 0.9 "outstanding" (Mandrekar, 2010). Whether an AUC is good enough depends on the use: a triage score at 0.75 may help; a confirmatory test at 0.85 may not.
- **The curve does not choose the cut-off.** Youden's J is one rule, and it assumes equal costs of the two errors. The clinical cost of a false negative against a false positive, and the prevalence where the test will be used (recall §5.5.2), should decide.
- **Once a cut-off is chosen,** sensitivity, specificity, predictive values and likelihood ratios at that cut-off are computed as in §5.5.1–5.5.3. Report their CIs (recall §7.2.4) and the AUC's CI.
- **Two tests with the same AUC can differ where it matters.** One curve may be higher at the high-sensitivity end, which is the end a screening programme uses.
- The reference standard must be independent of the test. If the test result helps define "diseased", sensitivity and the AUC are inflated.

CHECKPOINT 15.8

1. An ROC curve for a new biomarker lies almost on the diagonal. A second analyst reports an AUC of 0.21 for another marker. What does each result mean, and what should the second analyst check first?
2. A study has 8 diseased and 5 non-diseased people. Of the 40 diseased–non-diseased pairs, the diseased person has the higher biomarker value in 30, the values are tied in 4, and the non-diseased person is higher in 6. Compute the AUC.
3. In the MUAC example, compute Youden's J and LR+ at the cut-offs ≤ 12.3 cm and ≤ 12.9 cm. Which would you recommend for community screening, and why does J not settle the question?

(Answers in the Appendix at the end of the book.)

15.5.2 Clinical Prediction Models – Discrimination, Calibration, Validation

DEFINITION

A **clinical prediction model** combines several predictors to estimate each person's probability of an outcome, now (diagnostic) or in future (prognostic). Its performance has two parts. **Discrimination** is how well it ranks people with the outcome above people without it, measured by the **c-statistic**, the AUC of §15.5.1 applied to the model's predicted probabilities. **Calibration** is whether the predicted probabilities are numerically right: among households given 30%, about 30% should have the outcome. **Validation** estimates how both will hold up in people not used to build the model.

SIMPLIFIED EXPLANATION

A regression model built to *explain* (is distance associated with membership, adjusting for confounders?) is judged by its coefficients. A model built to *predict* (which households are likely to join?) is judged by its predictions (recall §15.1.5, where the two aims lead to different modelling strategies). The same CBHI logistic model of §13.3 can be read both ways. Here it is read as a predictor.

Every performance measure computed on the data used to fit the model is too optimistic. The model has partly fitted the noise of that sample, so it looks better there than it will in new households. **Internal validation** estimates this **optimism** by refitting the model in bootstrap resamples (recall §7.2.4). **External validation** tests the model in new data: another district, or a later year.

DERIVATION

Calibration-in-the-large on the development data is perfect by construction. The log-likelihood of a logistic model (recall §13.2.3) is $\Sigma[y_i \ln p_i + (1 - y_i) \ln(1 - p_i)]$. Its derivative with respect to the intercept (its slope as the intercept changes; the calculus is beyond this book) is $\Sigma(y_i - p_i)$. At the maximum this is zero, so $\boldsymbol{\Sigma p_i = \Sigma y_i}$: the mean predicted probability equals the observed proportion. The same argument makes the **calibration slope** exactly 1 on the development data. Both checks become informative only in new data, or after bootstrap correction.

Calibration slope and intercept. Refit the outcome on the model's linear predictor (its log-odds): $\text{logit}(P(y = 1)) = a + b \times LP$. The slope b shows the spread of the predictions. $b < 1$ means the predic-

tions are too extreme, too high for high-risk and too low for low-risk people, the signature of overfitting. The intercept a, with b fixed at 1, is **calibration-in-the-large**: a < 0 means the model overestimates risk on average (Van Calster et al., 2019).

Bootstrap optimism. (1) Fit the model to a bootstrap resample and measure its performance there (training). (2) Measure the same fitted model on the original data (test). (3) Optimism = training − test. (4) Average over 200 or more resamples and subtract the average from the apparent performance (Steyerberg et al., 2001).

Hosmer–Lemeshow (HL) test. Sort people by predicted probability, split them into g groups (usually tenths), and compare observed events O with expected events E (the sum of predicted probabilities) in each group of size n:

$HL = \Sigma\,(O - E)^2 \div [E(1 - E/n)]$, referred to χ^2 with $g - 2$ degrees of freedom.

ILLUSTRATION / EXAMPLE

The CBHI model of §13.3 predicts membership from distance, wealth and maternal education in 500 households (150 members). Its c-statistic is 0.72: a randomly chosen member has a higher predicted probability than a randomly chosen non-member in 72% of the 150 × 350 = 52,500 member–non-member pairs. The mean predicted probability is 0.39 among members and 0.26 among non-members (recall §13.3). Over all 500 households it is 0.30, exactly the observed 30%, as the Derivation requires.

WORKED CALCULATION

1. Sort the 500 households by predicted probability and cut them into ten groups of 50.

Tenth	Mean predicted	Expected members E	Observed members O
1	0.10	4.8	5
2	0.14	6.8	9
3	0.17	8.4	6
4	0.20	10.1	6
5	0.24	12.2	16
6	0.29	14.3	16
7	0.34	17.1	18
8	0.40	20.0	21
9	0.49	24.5	20
10	0.64	31.8	33

2. Largest term, tenth 4: $(6 - 10.15)^2 \div [10.15 \times (1 - 10.15/50)] = 17.2 \div 8.09 = 2.13$. The ten terms sum to **HL = 7.57** on 10 − 2 = 8 df, **p = 0.48**.
3. The expected column sums to 150, the observed number of members (Derivation).
4. Bootstrap validation (200 resamples, a computer task; the optional In R box shows it): the apparent c-statistic 0.716 has an optimism of 0.010, so the corrected c-statistic is **0.706**. The corrected calibration slope is **0.96**: new households would get predictions about 4% too extreme on the

log-odds scale. With 150 events and 4 coefficients the optimism is small, as expected (recall §13.2.4 on events per variable).

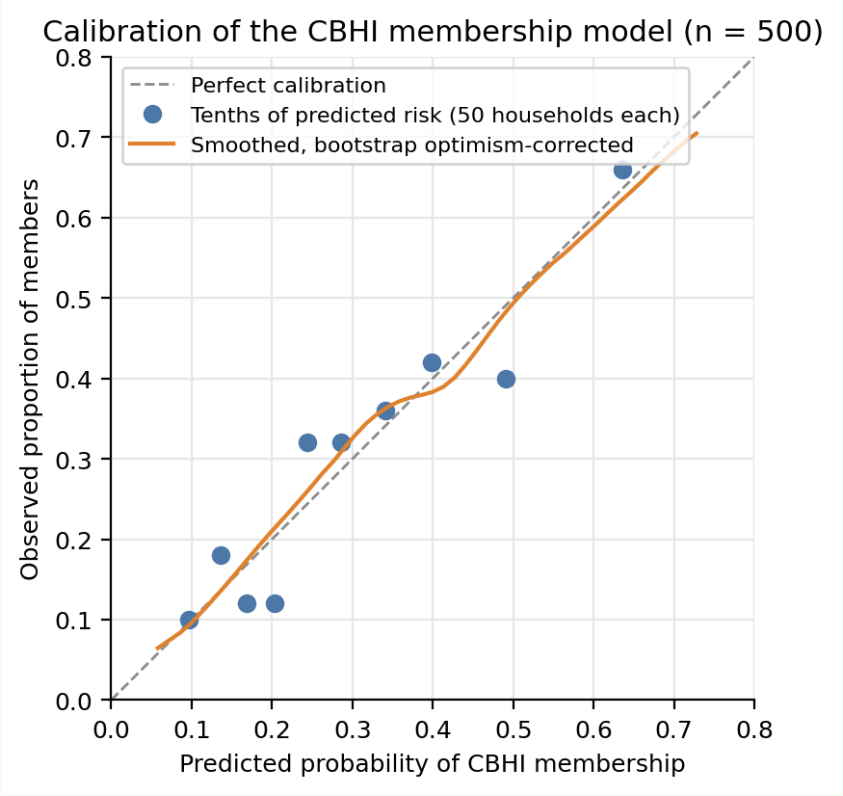


Figure 15.8. Calibration of the CBHI membership model. Points: observed proportion of members against mean predicted probability in each tenth of predicted risk (step 1). Orange line: smoothed calibration curve corrected for optimism by 200 bootstrap resamples. The dashed diagonal is perfect calibration.

IN R (OPTIONAL)

```
library(rms)
cb ← read.csv("data/cbhi_survey.csv")
cb$maternal_education ← factor(cb$maternal_education,
  levels = c("none", "primary", "secondary+"))
f ← lrm(cbhi_member ~ distance_km + wealth_index +
          maternal_education,
        data = cb, x = TRUE, y = TRUE)
round(f$stats[c("C", "Dxy", "Brier")], 3)
set.seed(2026)
round(validate(f, B = 200)[c("Dxy", "Intercept", "Slope"), 1:5], 3)
plot(calibrate(f, B = 200))        # the calibration plot
```

```
    C   Dxy Brier
0.716 0.432 0.184
          index.orig training   test optimism index.corrected
Dxy            0.432    0.444  0.424    0.020           0.412
Intercept      0.000    0.000 -0.013    0.013          -0.013
Slope          1.000    1.000  0.960    0.040           0.960
```

`C` is the c-statistic, 0.716. `rms` reports discrimination as Somers' **Dxy** = 2(C − 0.5), so C = 0.5 + Dxy/2. The `validate` rows follow the four bootstrap steps: `training` is the average performance in the resamples, `test` the average on the original data, and `optimism` their difference. Corrected Dxy

0.412 gives C = 0.5 + 0.412/2 = 0.706 (step 4). The apparent intercept 0 and slope 1 are the "by construction" values of the Derivation; the corrected slope is 0.96. `Brier` is the mean squared difference between predicted probability and outcome (0 is perfect).

MUST-KNOW NOTES (MD-LEVEL)

- **Discrimination and calibration are separate.** Multiplying every predicted odds by 2 leaves the ranking, and so the c-statistic, unchanged but ruins calibration. Report both.
- **The HL test is weak evidence either way.** It depends on an arbitrary grouping, has low power in small samples, and says nothing about how the model is miscalibrated. Van Calster et al. (2019) "recommend against" it; prefer a calibration plot with slope and intercept. In very large samples a trivial miscalibration becomes significant: in simulations with 50,000 patients and a 0.4% deviation, the HL test was significant in all models (Kramer & Zimmerman, 2007). Here p = 0.48 means "no evidence of miscalibration in 500 households", not "well calibrated".
- **Apparent performance is optimistic.** Bootstrap internal validation gives "stable estimates with low bias"; splitting the sample into development and test halves is inefficient (Steyerberg et al., 2001).
- **External validation is the real test.** A model built in one district may be miscalibrated in another where membership is commoner (calibration-in-the-large) or the predictors act differently (slope). Recalibrating the intercept and slope is often enough to update it.
- **Reporting.** TRIPOD+AI (Collins et al., 2024) is the 27-item checklist for studies that develop or validate prediction models, whether by regression or machine learning. It supersedes TRIPOD 2015 (22 items; Collins et al., 2015).
- **Sample size for prediction models** is set by the number of events, the number of predictor parameters and the outcome prevalence, not by a fixed events-per-variable rule (recall §13.2.4).

CHECKPOINT 15.9

1. Two logistic prediction models for 1-year TB treatment default both have a c-statistic of 0.78 in a validation sample. Model A's calibration slope is 1.02 and intercept 0.01; model B's slope is 1.0 but it predicts every patient's risk about twice too high. What does the shared c-statistic tell you, and what does it fail to tell you? Which model would you give to a counsellor who quotes risks to patients?
2. A model's apparent Somers' Dxy is 0.60. Bootstrap validation gives an average optimism of 0.06 in Dxy and a corrected calibration slope of 0.85. Compute the apparent and corrected c-statistics, and explain what a slope of 0.85 says about the predictions.
3. A team reports "Hosmer–Lemeshow p = 0.48, so the model is well calibrated" for 500 households. Why is this conclusion too strong? What would you expect the same test to show if the same model, with the same small miscalibration, were checked in 50,000 households?

(Answers in the Appendix at the end of the book.)

15.6 Sample Size and Power Calculation

DEFINITION

A **power-based sample size** is the number of participants needed so that a planned comparison has a stated probability (the power, $1 - \beta$) of giving a significant result at level α if the true effect equals a stated, clinically important size. It differs from the **precision-based** sample size of §7.3 (recall: n chosen so that one estimate has a stated margin of error). Estimating one prevalence needs §7.3; comparing groups, testing an association, or running a trial needs this section.

SIMPLIFIED EXPLANATION

Sample size is a design decision, not a formula lookup. It starts with the **primary research question**, which fixes the **primary outcome** (what is measured) and the **primary comparison** (between which groups). Only then does "effect size" mean anything: the smallest difference that would matter clinically or for the programme, not the smallest that happens to be detectable. The remaining inputs (the expected variability or event rate, α and power; recall §8.4) are plugged into a formula that matches the planned analysis. A t-test, a comparison of proportions, a correlation and a log-rank test each have their own formula, which is why the analysis (recall §8.8) is chosen before the sample size.

All the formulas below come from one idea, derived in §8.4: the expected difference must lie ($z_{\alpha/2} + z_{\beta}$) standard errors from zero. With $\alpha = 0.05$ two-sided and 80% power, $z_{\alpha/2} = 1.960$, $z_{\beta} = 0.842$, and $(1.960 + 0.842)^2 =$ **7.849**. For 90% power, $z_{\beta} = 1.282$ and the constant is 10.51.

DERIVATION

Two means (recall §8.4). SE of the difference with n per group = $\sigma\sqrt{(2/n)}$. Setting $\Delta/SE = z_{\alpha/2} + z_{\beta}$ gives

$$n \text{ per group} = 2(z_{\alpha/2} + z_{\beta})^2\sigma^2 \div \Delta^2.$$

Two proportions p_1, p_2. Under H_0 the difference has $SE_0 = \sqrt{(2\bar{p}\bar{q}/n)}$, where $\bar{p} = (p_1 + p_2)/2$ and $\bar{q} = 1 - \bar{p}$; under H_1 it has $SE_1 = \sqrt{[(p_1q_1 + p_2q_2)/n]}$. The test rejects when the observed difference exceeds $z_{\alpha/2} \times SE_0$. Power is $P(\text{difference} > z_{\alpha/2}\, SE_0) = \Phi[(\Delta - z_{\alpha/2}\, SE_0)/SE_1]$. Setting this to $1 - \beta$ gives $\Delta = z_{\alpha/2}\, SE_0 + z_{\beta}\, SE_1$. Multiply by $\sqrt{n}$ and square:

$$n \text{ per group} = [z_{\alpha/2}\sqrt{(2\bar{p}\bar{q})} + z_{\beta}\sqrt{(p_1q_1 + p_2q_2)}]^2 \div (p_1 - p_2)^2.$$

Cohort study, via the risk ratio. Put p_0 = risk in the unexposed and $p_1 = RR \times p_0$, then use the two-proportion formula.

Case-control study, via the odds ratio. The proportions compared are the proportions *exposed* among controls (p_0) and cases (p_1). The odds of exposure in cases are $OR \times p_0/(1 - p_0)$ (recall §10.6.4: the exposure OR equals the disease OR), so $p_1 = OR \times p_0 \div [1 + p_0(OR - 1)]$. Then use the two-proportion formula.

Correlation. Fisher's $z = \frac{1}{2} \ln[(1 + r)/(1 - r)]$ has SE $1/\sqrt{(n - 3)}$ (recall §11.2). Setting $z(r)/SE = z_{\alpha/2} + z_{\beta}$ gives $n = [(z_{\alpha/2} + z_{\beta})/z(r)]^2 + 3$.

Unequal allocation, k controls per case (or k:1 arms). The variance of a difference is $\sigma^2(1/n_1 + 1/n_2)$. With a fixed total N and $n_2 = kn_1$, $n_1 = N/(1 + k)$, and the variance is $\sigma^2(1 + k)^2/(kN)$. At 1:1 it is $4\sigma^2/N$.

So k:1 needs $\mathbf{(1 + k)^2/(4k)}$ times the total of a 1:1 design: 1.125 for 2:1, 1.33 for 3:1, 1.8 for 5:1. The inverse, $4k/(1 + k)^2$, is the relative efficiency.

Cluster-randomised trial. Randomising clusters (villages, anganwadi centres) of m people each inflates the variance by the design effect 1 + (m − 1)ICC (recall §6.2.6 and §15.2.2), so multiply the individually randomised n by it and divide by m to get clusters per arm (Hayes & Bennett, 1999).

Survival (events-driven). For a log-rank test or Cox model with a proportion a allocated to one arm, the number of **events** needed is $D = (z_{\alpha/2} + z_{\beta})^2 \div [a(1 - a)(\ln HR)^2]$, which is $4(z_{\alpha/2} + z_{\beta})^2/(\ln HR)^2$ for 1:1 allocation (Schoenfeld, 1983; derivation beyond this book; not to be confused with the Schoenfeld residuals of §14.6.2). Power depends on events, not people. The number to enrol is D divided by the expected proportion who will have the event during follow-up.

ILLUSTRATION / EXAMPLE

Five planning questions from one district, each at $\alpha = 0.05$ (two-sided) and 80% power.

WORKED CALCULATION

1. *Two means.* Iron–folic acid with counselling is expected to raise mean haemoglobin in anaemic adolescent girls by 0.5 g/dL; SD 1.2 g/dL. $n = 2 \times 7.849 \times 1.2^2 \div 0.5^2 = 2 \times 7.849 \times 1.44 \div 0.25 = 90.4$ → **91 per group** (92 by the t-distribution, which software computes).
2. *Two proportions.* Peer counselling is expected to raise exclusive breastfeeding at 6 months from 40% to 55%. $\bar{p} = 0.475$. $\sqrt{(2 \times 0.475 \times 0.525)} = 0.7062$; $\sqrt{(0.40 \times 0.60 + 0.55 \times 0.45)} = 0.6982$. $n = (1.960 \times 0.7062 + 0.8416 \times 0.6982)^2 \div 0.15^2 = (1.3842 + 0.5876)^2 \div 0.0225 = 3.888 \div 0.0225 = 172.8$ → **173 per group**.
3. *Case-control.* Not-EBF is expected in 30% of controls; the smallest OR worth detecting is 2. $p_1 = 2 \times 0.30 \div (1 + 0.30) = 0.4615$. The two-proportion formula gives 140.7 → **141 cases and 141 controls**.
4. *Cohort.* Diarrhoea risk in exclusively breastfed infants 20%; RR to detect 1.5, so $p_1 = 0.30$. $n = 293.2$ → **294 per group**.
5. *Correlation.* To detect r = 0.30 between years of service and anganwadi workers' knowledge scores: $z(0.30) = \tfrac{1}{2} \ln(1.3/0.7) = 0.3095$. $n = (2.802/0.3095)^2 + 3 = 81.9 + 3 = 84.9$ → **85 workers**.
6. *Cluster RCT.* Question 2 run by randomising anganwadi centres, 20 infants per centre, ICC 0.05 from earlier surveys. $DEFF = 1 + 19 \times 0.05 = 1.95$. $n = 172.8 \times 1.95 = 337.0$ per arm → 337/20 = 16.9 → **17 centres per arm** (340 infants).
7. *Events-driven.* A CBHI retention intervention should cut the dropout hazard by 30% (HR 0.70). $D = 4 \times 7.849 \div (\ln 0.70)^2 = 31.40 \div 0.1272 = 246.8$ → **247 dropouts**. If 30% of households are expected to drop out during follow-up, enrol 246.8/0.30 = 822.6 → **823 households**.
8. *Unequal allocation.* Question 1 with 2:1 allocation and the same 182 girls: the variance rises by 9/8, so Δ/SE falls from 2.802 to $2.802/\sqrt{1.125} = 2.641$, and power $= \Phi(2.641 - 1.960) = \Phi(0.681) =$ **0.75**, not 0.80. Keeping 80% needs the unrounded 1:1 total, $2 \times 90.4 = 180.8$, times 1.125 = 203.4: 68 + 136 = **204 girls**.
9. *Losses.* Inflate the final n for expected loss to follow-up exactly as for non-response (recall §7.3): 91 per group with 15% expected loss becomes 91/0.85 = 107.1 → 108 per group.

IN R (OPTIONAL)

```
library(pwr)
power.t.test(delta = 0.5, sd = 1.2, power = 0.80)$n        # two means
# two proportions
power.prop.test(p1 = 0.40, p2 = 0.55, power = 0.80)$n
# case-control, OR 2
power.prop.test(p1 = 0.30, p2 = 0.4615, power = 0.80)$n
# correlation
pwr.r.test(r = 0.3, power = 0.80)$n
z ← qnorm(0.975) + qnorm(0.80)
# events for HR 0.7
4 * z^2 / log(0.7)^2
# power, 2:1 at the 1:1 total
pnorm(z / sqrt(9/8) - qnorm(0.975))
```

```
[1] 91.38944
[1] 172.7999
[1] 140.7206
[1] 84.07364
[1] 246.7871
[1] 0.7521895
```

Line by line: 91.4 (step 1, by the t-distribution, so 92 per group); 172.8 (step 2, the same formula as the Derivation); 140.7 (step 3); 84.1 (step 5; `pwr.r.test` adds a small-sample correction, and both round up to 85); 246.8 events (step 7); and power 0.752 for 2:1 allocation (step 8). Every n is rounded *up*.

MUST-KNOW NOTES (MD-LEVEL)

- **Inputs, not outputs.** n depends on α, power, the effect size and the variability or baseline risk (recall §8.4). The effect size is the smallest *important* effect, not the effect you hope for or found in a small pilot.
- **Halving Δ quadruples n.** Every formula has Δ^2 (or $(\ln \text{HR})^2$, or $z(r)^2$) in the denominator.
- **Clustering must be planned for.** Omitting the design effect when clusters are randomised or sampled gives a trial whose effective sample size is DEFF times smaller than planned, so it is underpowered. With few clusters per arm, the analysis needs small-sample care (recall §15.2.5).
- **Unequal allocation costs little until it is extreme.** 2:1 needs 12.5% more participants; 3:1, 33% more.
- **Survival studies are sized in events.** A long follow-up or a higher-risk population reaches the events sooner.
- **Report the calculation** with every input and its source, so a reader can reproduce n (CONSORT, the trial reporting guideline of §15.9, asks for this).
- **After the study, report the estimate and its CI, not "observed power"** (recall §8.4). An underpowered study cannot be rescued by reanalysis.

CHECKPOINT 15.10

1. A trial was sized at 91 girls per arm for 80% power with 1:1 allocation. For recruitment reasons the investigators switch to 2:1 (intervention:control) but keep the same total of 182. Compute the relative efficiency and the new power. How many girls would 2:1 need to keep 80% power?
2. A calculation gives 200 per group, but 20% loss to follow-up is expected. What should the enrolment target be, and why is 200 × 1.2 slightly wrong?
3. A cohort study will compare 1-year diarrhoea risk in infants of mothers with and without a latrine at home. Risk without a latrine is 15%; the smallest important difference is a fall to 10%. Compute n per group for 80% power at α = 0.05, and then allow for 10% loss to follow-up.

(Answers in the Appendix at the end of the book.)

15.7 Meta-Analysis Statistics

DEFINITION

Meta-analysis is the statistical combination of the results of several independent studies of the same question into one pooled estimate. It is usually the analysis stage of a **systematic review**, a review that finds, selects and appraises all eligible studies by a pre-specified, reproducible method.

SIMPLIFIED EXPLANATION

Each study is reduced to two ingredients used throughout this book: an **effect estimate** (a mean difference, a log risk ratio, a log odds ratio) and its **standard error** (recall §6.7). A simple average of the estimates would treat a small, imprecise trial like a large, precise one and throw away what the SE says. Meta-analysis instead weights each study by the inverse of its variance, $w = 1/SE^2$, so precise studies count more. Ratio measures are pooled on the log scale, where their sampling distribution is closer to Normal (recall §10.6.4), and the result is exponentiated at the end.

DERIVATION

Why inverse-variance weights? A weighted average $\bar{y} = \Sigma w_i y_i / \Sigma w_i$ of independent estimates has variance $\Sigma w_i^2 SE_i^2 / (\Sigma w_i)^2$. With $w_i = 1/SE_i^2$ this becomes $\Sigma w_i / (\Sigma w_i)^2 = 1/\Sigma w_i$, and no other choice of weights gives a smaller variance. So

pooled estimate $= \Sigma w_i y_i \div \Sigma w_i$, $SE(\text{pooled}) = 1 \div \sqrt{\Sigma w_i}$.

Because Σw_i is larger than any single w_i, the pooled SE is smaller than every study's own SE.

ILLUSTRATION / EXAMPLE

Ten trials each test whether an antenatal supplement reduces low birthweight. Each is small, with a wide CI. Pooled, they give one estimate whose CI is narrower than any single trial's.

MUST-KNOW NOTES (MD-LEVEL)

- A meta-analysis is only as good as its studies. Pooling biased studies gives a more precise wrong answer: precision does not remove bias (recall §6.4). Bias in the design sense (systematic error in how studies select, measure or analyse people) is catalogued in §17.2; the risk of bias of each included study should be assessed before pooling.
- The studies must answer a comparable question: similar populations, interventions, comparators and outcome definitions. Heterogeneity (§15.7.2) measures how far the results disagree beyond chance.
- §15.7.1 draws the forest plot, §15.7.2 measures heterogeneity, §15.7.3 pools by fixed- and random-effects models, and §15.7.4 checks for missing studies and grades the evidence.

15.7.1 Forest Plot

DEFINITION

A **forest plot** shows each study's effect estimate as a square, with a horizontal line for its 95% CI, and the pooled estimate as a diamond at the bottom whose width is the pooled CI. The square's area is proportional to the study's weight.

SIMPLIFIED EXPLANATION

The plot shows, on one page, both the individual pieces of evidence and how they combine. A vertical line marks "no effect" (RR = 1, or a mean difference of 0). A CI that crosses it is compatible with no effect at the 5% level (recall §8.5). Ratio measures are drawn on a **log axis**: on it, an RR of 0.5 and an RR of 2 lie the same distance from 1, and each CI, computed as ln RR ± 1.96 SE, is symmetric about its square.

ILLUSTRATION / EXAMPLE

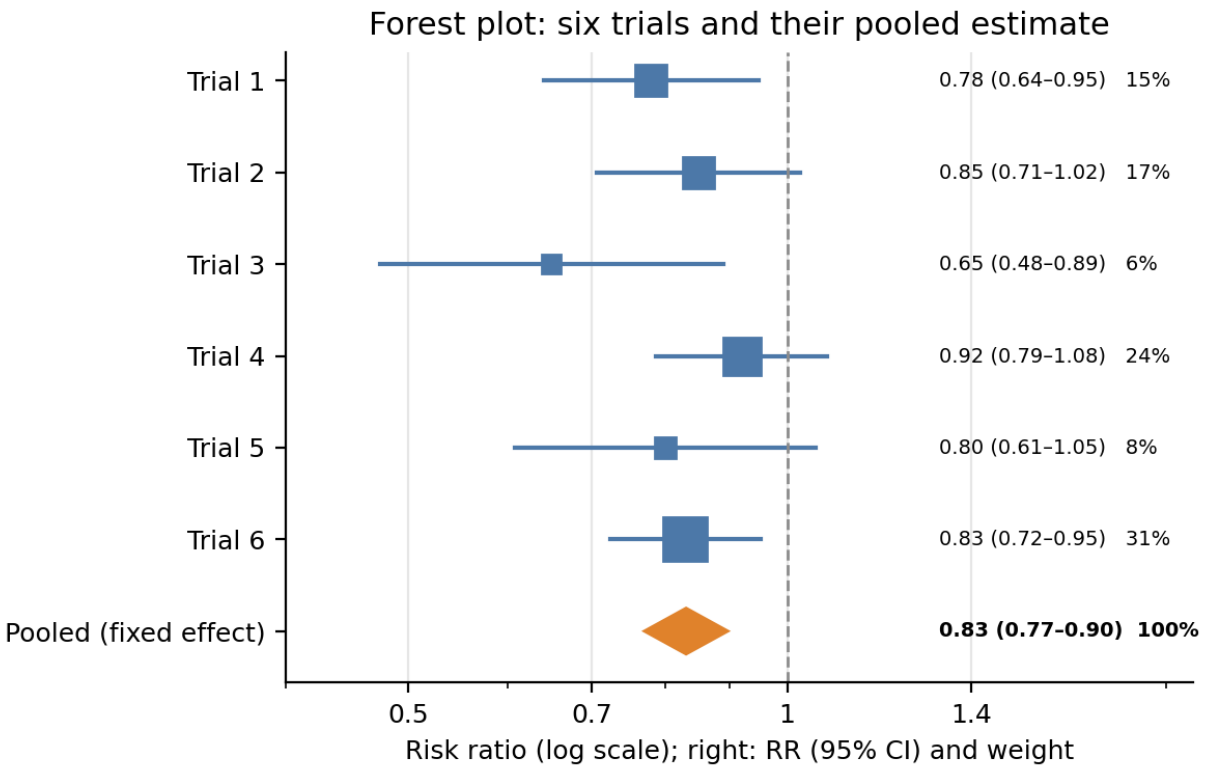


Figure 15.9. Forest plot of six trials of a hypothetical intervention (risk ratios, log axis). Square area is proportional to the fixed-effect weight (right-hand column). The diamond is the fixed-effect pooled RR 0.83 (95% CI 0.77 to 0.90).

Trial 6 carries the most weight (31%) because it has the narrowest CI; trial 3 carries 6%. Three of the six CIs (trials 1, 3 and 6) exclude 1, three cross it, and the pooled CI, 0.77 to 0.90, is narrower than any trial's.

MUST-KNOW NOTES (MD-LEVEL)

- Weight follows precision, not sample size as such: a large trial with few events can still have a wide CI.
- The diamond is usually narrower than every study's CI (Derivation, §15.7). Under a random-effects model (one that lets the true effect vary between studies, §15.7.3) with substantial heterogeneity it is wider than the fixed-effect diamond, and may be little narrower than the most precise study (§15.7.3).
- Read the spread of the squares before the diamond: widely scattered, non-overlapping CIs warn that one pooled number may not describe the studies (§15.7.2).

15.7.2 Heterogeneity – I^2 and Q Statistic

DEFINITION

Heterogeneity is variation in the *true* effects of the studies beyond what sampling error alone would produce. **Cochran's Q** tests H_0: all studies share one true effect. **I^2** is the percentage of the total variation across studies that is due to heterogeneity rather than chance.

SIMPLIFIED EXPLANATION

Even if every study estimated exactly the same true effect, their estimates would differ by sampling error (recall §6.4). Q asks whether they differ *more* than their SEs allow, for example because the studies had different populations, doses or outcome definitions. I^2 turns Q into a descriptive percentage that does not depend on the number of studies in the same way.

DERIVATION

With k studies, fixed-effect weights w_i and fixed-effect pooled estimate $\bar{y}$:

$Q = \Sigma\, w_i(y_i - \bar{y})^2$.

Each term is a squared z-score, $[(y_i - \bar{y})/SE_i]^2$. Under H_0, Q follows approximately χ^2 with **df = k − 1** (one df is used by estimating $\bar{y}$; recall §5.7.2.4). A χ^2 variable has mean equal to its df, so under H_0 Q is expected to be about k − 1. The excess, Q − df, measures heterogeneity, and as a share of Q it gives

$I^2 = (Q - df)/Q \times 100\%$, set to 0 when Q < df (Higgins et al., 2003).

ILLUSTRATION / EXAMPLE

Three trials of a community intervention against childhood diarrhoea (data/meta_three_trials.csv):

Trial	Intervention: events/n	Control: events/n	RR	ln RR	SE(ln RR)
A	38/400	56/400	0.68	−0.388	0.198
B	60/450	64/450	0.94	−0.065	0.167
C	15/150	27/150	0.56	−0.588	0.301

$SE(\ln RR) = \sqrt{(1/a - 1/n_1 + 1/c - 1/n_0)}$ (recall §10.6.4). Trial A: $\sqrt{(1/38 - 1/400 + 1/56 - 1/400)} = \sqrt{0.0392} = 0.198$.

WORKED CALCULATION

1. Fixed-effect weights $1/SE^2$: A 25.5, B 35.9, C 11.1; Σw = 72.5 (§15.7.3 step 1). Pooled $\bar{y} = -0.258$.
2. Q terms: A $25.5 \times (-0.388 + 0.258)^2 = 0.43$; B $35.9 \times (-0.065 + 0.258)^2 = 1.35$; C $11.1 \times (-0.588 + 0.258)^2 = 1.20$. **Q = 2.98** on df = 2, **p = 0.23**.
3. **$I^2 = (2.98 - 2)/2.98 = 32.8\%$.**

MUST-KNOW NOTES (MD-LEVEL)

- **Q has low power when studies are few or small.** A non-significant Q is not evidence of no heterogeneity. For this reason p < 0.10, rather than 0.05, is sometimes used (Higgins et al., 2003; Cochrane Handbook, ch. 10). With k = 3, p = 0.23 says little.
- **I^2 labels are tentative conventions.** Higgins et al. (2003) "tentatively" called 25%, 50% and 75% low, moderate and high. The Cochrane Handbook gives overlapping bands: 0–40% might not be important, 30–60% moderate, 50–90% substantial, 75–100% considerable, and adds that importance depends on the size and direction of the effects and the strength of evidence for

heterogeneity. Here $I^2 = 32.8\%$ lies in the overlap of the first two bands; its CI would be very wide with three trials.

- **I^2 is a proportion, not an amount.** It grows as the studies get larger (smaller SEs), even if the spread of true effects is unchanged. τ (§15.7.3) measures the spread itself.
- **Judgement comes first.** Whether the studies are similar enough in population, intervention and outcome to share one true effect is decided on clinical and methodological grounds before looking at I^2. I^2 can confirm or challenge that judgement; it should not make it by a cut-off ("I^2 over 50%, so random effects").
- **Explore, do not just report.** When heterogeneity is present, prespecified subgroup analyses (by baseline risk, dose, setting) can suggest why, and the pooled estimate is interpreted with caution.

15.7.3 Fixed vs Random Effects Models

DEFINITION

A **fixed-effect** model assumes that every study estimates one common true effect and that the studies differ only by sampling error. A **random-effects** model assumes that the true effects themselves vary across studies around a mean, with between-study variance $\boldsymbol{\tau^2}$, and estimates that mean. A **prediction interval** gives the range in which the true effect of a new, similar study is expected to lie.

SIMPLIFIED EXPLANATION

The phrase "random effects" appeared in §15.2.3 for patients nested in clusters (recall: cluster-specific values treated as draws from a shared distribution). Here the "clusters" are studies. In a random-effects model each study's estimate has two sources of variance: its own sampling variance SE^2 and the between-study variance τ^2. The weights become $1/(SE^2 + \tau^2)$. Because τ^2 is added to every study, the weights become more equal: small studies gain relative weight, large studies lose it, and the pooled CI widens.

DERIVATION

DerSimonian–Laird estimate of τ^2. Under heterogeneity, Q exceeds its null expectation $df = k - 1$. Setting the expected value of Q equal to its observed value and solving (DerSimonian & Laird, 1986) gives

$$\tau^2 = \max\{0, (Q - df) \div [\Sigma w_i - \Sigma w_i^2/\Sigma w_i]\}.$$

Random-effects pooling. $w_i^* = 1/(SE_i^2 + \tau^2)$; pooled $\hat{\mu} = \Sigma w_i y_i / \Sigma w_i$; $SE(\hat{\mu}) = 1/\sqrt{\Sigma w_i^*}$.

Prediction interval. A new study's true effect varies around μ with variance τ^2, and μ itself is estimated with variance $SE(\hat{\mu})^2$. Adding them: $\hat{\mu} \pm t \times \sqrt{(\tau^2 + SE(\hat{\mu})^2)}$, where Higgins et al. (2009) use t with $k - 2$ df. A z-value (1.96) is also used (the metafor default).

ILLUSTRATION / EXAMPLE

Pooling the three diarrhoea trials of §15.7.2 both ways.

WORKED CALCULATION

1. *Fixed-effect weights.* Variances: A $0.198^2 = 0.0392$, B 0.0278, C 0.0904. Weights: 25.5, 35.9, 11.1; $\Sigma w = 72.5$ (percentages 35.2, 49.5, 15.3).
2. *Fixed-effect pooled.* $\Sigma wy = 25.5 \times (-0.388) + 35.9 \times (-0.065) + 11.1 \times (-0.588) = -18.72$; $\bar{y} = -18.72/72.5 = -0.258$. SE $= 1/\sqrt{72.5} = 0.117$. 95% CI: $-0.258 \pm 1.96 \times 0.117 = -0.488$ to -0.028. Back-transformed: **RR 0.77 (0.61 to 0.97)**.
3. τ^2. $\Sigma w^2 = 25.5^2 + 35.9^2 + 11.1^2 = 2{,}064$; $\Sigma w - \Sigma w^2/\Sigma w = 72.5 - 28.5 = 44.0$. $\tau^2 = (2.98 - 2)/44.0 =$ **0.0222** ($\tau = 0.149$ on the ln RR scale).
4. *Random-effects weights.* $1/(0.0392 + 0.0222) = 16.3$; $1/(0.0278 + 0.0222) = 20.0$; $1/(0.0904 + 0.0222) = 8.9$; $\Sigma = 45.2$ (percentages 36.1, 44.2, 19.7). Trial C's share rises from 15% to 20%.
5. *Random-effects pooled.* $\hat{\mu} = -0.284$; SE $= 1/\sqrt{45.2} = 0.149$. 95% CI −0.576 to 0.008. **RR 0.75 (0.56 to 1.01).** The random-effects CI now crosses 1.
6. *Prediction interval.* $\sqrt{(0.0222 + 0.149^2)} = 0.211$. With z: $-0.284 \pm 1.96 \times 0.211 \rightarrow$ **RR 0.50 to 1.14**. With t on $k - 2 = 1$ df ($t = 12.71$): **0.05 to 10.9**, useless, which is the honest message: three trials cannot say how much the effect varies between settings.

IN R (OPTIONAL)

```
library(metafor)
d  ← read.csv("data/meta_three_trials.csv")
es ← escalc("RR", ai = events_int, n1i = n_int,
                  ci = events_ctl, n2i = n_ctl,
            data = d, slab = study)
fe ← rma(yi, vi, data = es, method = "FE")
re ← rma(yi, vi, data = es, method = "DL")
re
predict(re, transf = exp)
forest(re, atransf = exp)          # forest plot on a log axis
```

```
Random-Effects Model (k = 3; tau^2 estimator: DL)
tau^2 (estimated amount of total heterogeneity): 0.0222 (SE = 0.0681)
tau (square root of estimated tau^2 value):      0.1490
I^2 (total heterogeneity / total variability):   32.83%
...
Test for Heterogeneity:
Q(df = 2) = 2.9774, p-val = 0.2257
Model Results:
estimate      se     zval    pval    ci.lb   ci.ub
 -0.2841  0.1488  -1.9092  0.0562  -0.5757  0.0076  .
...
   pred  ci.lb  ci.ub  pi.lb  pi.ub
 0.7527 0.5623 1.0076 0.4982 1.1372
```

`escalc("RR")` computes each trial's ln RR (`yi`) and its variance (`vi`, the SE^2 of the table). $\tau^2 = 0.0222$, $I^2 = 32.83\%$ and $Q = 2.98$ ($p = 0.23$) are steps 3 and §15.7.2. The model estimate −0.284 (SE 0.149) is step 5 on the log scale; `predict(..., transf = exp)` back-transforms it to RR 0.75 (0.56 to 1.01) and gives the z-based prediction interval 0.50 to 1.14 (step 6). `rma(method = "FE")`, printed the same way, gives step 2.

MUST-KNOW NOTES (MD-LEVEL)

- **Fixed-effect weights follow precision; random-effects weights are more equal.** Large studies dominate a fixed-effect analysis. When small studies differ systematically from large ones (§15.7.4), a random-effects analysis gives them more influence, which can be a disadvantage.
- **Only the random-effects model widens the CI.** A fixed-effect model estimates one common effect and does not widen its CI for heterogeneity. A random-effects model estimates the mean of a distribution of true effects, and its CI (and more so its prediction interval) reflects the variation between studies.
- **The two models answer different questions.** Fixed effect: "what is the common effect in these studies?" Random effects: "what is the average effect across settings like these?" The prediction interval answers "what might happen in the next setting?" Riley et al. (2011) recommend reporting it; it "can be calculated when the meta-analysis contains at least three studies, although the interval may be very wide with so few studies", as step 6 shows.
- **Choose on substantive grounds, before seeing the result.** Picking whichever model gives the more favourable answer is results-driven analysis, like choosing a one-tailed test after the data (recall §8.6). When the choice is uncertain, report both as a sensitivity analysis and say whether the conclusion changes: here fixed effect 0.77 (0.61 to 0.97) and random effects 0.75 (0.56 to 1.01) differ in whether the CI excludes 1, so the conclusion is fragile.
- τ^2 is itself poorly estimated with few studies (here SE 0.068 against an estimate of 0.022).

CHECKPOINT 15.11

1. A meta-analysis of ten trials shows $I^2 = 68\%$. A colleague says: "I^2 is over 50%, so we must use a random-effects model." What reasoning does this shortcut, and what should come first?
2. A forest plot's pooled diamond is narrower than every study's CI. Explain why, using inverse-variance weights. When might a random-effects diamond be barely narrower than the most precise study?
3. Five studies give $Q = 12.0$. Compute I^2. Another five-study meta-analysis gives $Q = 3.0$: what is I^2, and why is it reported as that value?

(Answers in the Appendix at the end of the book.)

15.7.4 Publication Bias, Funnel Plots and PRISMA/GRADE

DEFINITION

Publication bias (more generally, non-reporting bias) arises when whether a study is published, or which results it reports, depends on its findings. **Small-study effects** are the tendency of smaller studies to show larger effects, for whatever reason. A **funnel plot** plots each study's effect estimate against a measure of its precision (usually its SE, with the axis reversed). **Egger's test** is a regression test for funnel-plot asymmetry. **Trim-and-fill** estimates how many studies may be missing and imputes them to give an adjusted pooled estimate.

SIMPLIFIED EXPLANATION

Small trials have wide sampling scatter. If all were published, their results would spread evenly on both sides of the pooled effect, wider at the bottom of the plot (large SE) and narrower at the top: an inverted funnel. If small trials that happened to find no benefit stay in desk drawers, one lower corner of the funnel is empty, and the pooled estimate is pulled towards a larger benefit.

Asymmetry does not prove publication bias. Smaller studies may also be of poorer quality (which tends to exaggerate effects), may have been done in higher-risk populations where the true effect is larger (real heterogeneity), or the asymmetry may be chance.

DERIVATION

Egger's test. If there is no small-study effect, every study estimates the same θ, and its z-score $y_i/SE_i = \theta \times (1/SE_i) + \text{noise}$: a straight line through the origin when plotted against precision $1/SE_i$. Egger et al. (1997) regress the "standard normal deviate" y_i/SE_i on $1/SE_i$. An intercept away from 0 means that the least precise studies sit systematically on one side: asymmetry. (An equivalent form, used by R's metafor package, regresses y_i on SE_i with weights $1/SE_i^2$ and tests the slope on SE.)

ILLUSTRATION / EXAMPLE

A simulation makes the mechanism visible. Forty trials of an intervention whose true RR is 0.85 are generated, with 25 to 800 children per arm. Every trial with a significant benefit is "published", as is every trial with more than 400 per arm; each of the rest is published with probability 0.2. Seventeen trials are published.

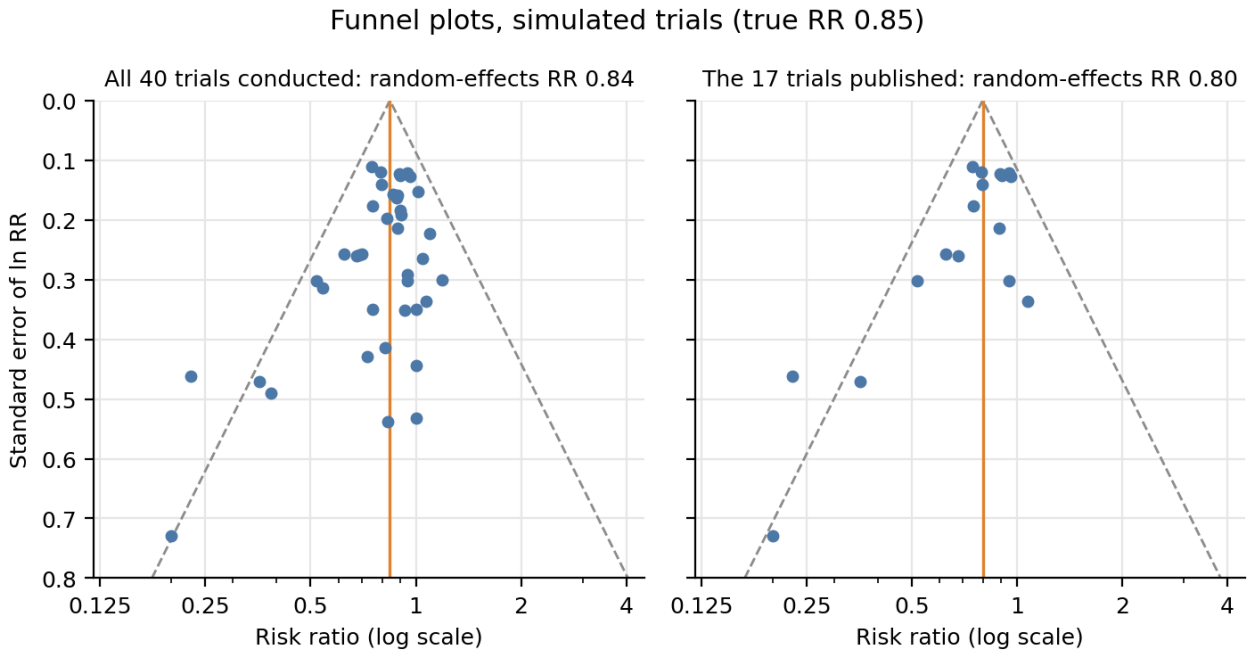


Figure 15.10. Funnel plots of 40 simulated trials (true RR 0.85). Left: all trials conducted, symmetric about the pooled RR 0.84. Right: the 17 "published" trials. Small trials to the right of the pooled line (no benefit) are missing, and the random-effects pooled RR moves to 0.80. Dashed lines: pooled ln RR ± 1.96 SE.

WORKED CALCULATION

1. All 40 trials: random-effects RR **0.84**, close to the true 0.85.
2. The 17 published trials: RR **0.80** (95% CI 0.72 to 0.89). The benefit is overstated: a 20% risk reduction instead of 16%.
3. Egger's test on the published trials: **p = 0.004**, strong evidence of asymmetry.
4. Trim-and-fill imputes **3** missing trials on the right, and the adjusted RR is **0.82** (0.72 to 0.94), partly corrected.

IN R (OPTIONAL)

```
library(metafor)
# simulated: 40 trials, true RR 0.85
set.seed(33)
# children per arm
n   ← round(exp(runif(40, log(25), log(800))))
es  ← escalc("RR", ai = rbinom(40, n, 0.17), n1i = n,
                   ci = rbinom(40, n, 0.20), n2i = n)
es$pub ← es$yi/sqrt(es$vi) < -1.96 | n > 400 | runif(40) < 0.2
# the 17 "published" trials
pub ← rma(yi, vi, data = es[es$pub, ], method = "DL")
all40 ← rma(yi, vi, data = es, method = "DL")   # all 40 trials
round(exp(c(all = coef(all40), published = coef(pub))), 2)
regtest(pub, model = "lm")                    # Egger's test
trimfill(pub)
```

```
      all.intrcpt published.intrcpt
             0.84              0.80
Regression Test for Funnel Plot Asymmetry
Model:     weighted regression with multiplicative dispersion
Predictor: standard error
Test for Funnel Plot Asymmetry: t = -3.3788, df = 15, p = 0.0041
...
Estimated number of missing studies on the right side: 3 (SE = 2.7999)
Random-Effects Model (k = 20; tau^2 estimator: DL)
...
Model Results:
estimate      se     zval    pval    ci.lb    ci.ub
 -0.1934  0.0681  -2.8385  0.0045  -0.3270  -0.0599  **
```

The first line compares steps 1 and 2. `regtest(model = "lm")` is Egger's test in its weighted-regression form (step 3): the SE of each trial predicts its effect, t = −3.38 on 15 df. `trimfill` adds 3 imputed trials (k = 20) and re-pools: exp(−0.193) = 0.82, CI exp(−0.327) to exp(−0.060) = 0.72 to 0.94 (step 4). The "SE = 2.80" beside the 3 missing trials shows how uncertain that number is. `funnel(pub)` draws the plot.

MUST-KNOW NOTES (MD-LEVEL)

- **At least 10 studies.** Tests for funnel-plot asymmetry "should be used only when there are at least 10 studies", because their power is low with fewer (Cochrane Handbook, ch. 13). The three-trial example of §15.7.3 cannot be assessed this way. Egger et al. (1997) themselves warned that results from "a limited number of small trials" need "considerable caution".

- **Asymmetry is not diagnosis.** The Cochrane Handbook states that funnel-plot asymmetry "should not be considered to be diagnostic for the presence of non-reporting bias"; heterogeneity, poor quality of small studies and chance produce it too.
- **Trim-and-fill is a sensitivity analysis, not a correction.** Its adjusted estimate assumes the asymmetry is due to publication bias (Duval & Tweedie, 2000). With large heterogeneity and no publication bias it can distort the estimate, and its performance varies widely, so Peters et al. (2007) recommend it "as a form of sensitivity analysis as intended by the authors of the method". Report it beside, not instead of, the main estimate.
- **Prevention beats detection.** Searching trial registries (CTRI in India, ClinicalTrials.gov), theses and grey literature, and contacting authors, finds unpublished studies directly.
- **PRISMA 2020** is the reporting guideline for systematic reviews: a 27-item checklist and a flow diagram showing records identified, screened, excluded (with reasons) and included (Page et al., 2021).
- **GRADE** rates the certainty of a body of evidence for each outcome as high, moderate, low or very low. Randomised trials start as high and observational studies as low; certainty is lowered for risk of bias (study limitations), inconsistency (heterogeneity, §15.7.2), indirectness, imprecision (a wide CI, as in §15.7.3) and publication bias (Guyatt et al., 2008).

CHECKPOINT 15.12

1. A meta-analysis of six trials reports "Egger's test p = 0.31, so there is no publication bias." Give two reasons this conclusion is unsafe.
2. In a meta-analysis of 12 trials, the four smallest (SE of ln RR 0.50 each) report RRs of 0.40 to 0.50, and the eight largest (SE 0.10 each) cluster around 0.90. Compute the share of the total fixed-effect weight carried by the four small trials. Would a random-effects analysis increase or decrease their influence, and why does that matter here?
3. Trim-and-fill imputes four "missing" trials and moves the pooled RR from 0.70 to 0.82. A colleague writes "the true effect is RR 0.82". How would you rewrite the sentence?

(Answers in the Appendix at the end of the book.)

15.8 Introduction to Bayesian Statistics

DEFINITION

Bayesian statistics treats an unknown parameter as uncertain and describes that uncertainty with a probability distribution. A **prior distribution** states what is believed about the parameter before the new data. Bayes' theorem (recall §5.5) combines it with the **likelihood** of the data (recall §13.1) to give the **posterior distribution**: posterior ∝ likelihood × prior. A 95% **credible interval** is a range that contains the parameter with 95% posterior probability, *given the prior and the model*.

SIMPLIFIED EXPLANATION

The methods of this book so far are **frequentist**: probability is a long-run frequency over repeated samples (recall §6.5 and the careful reading of a confidence interval in §7.2). Bayesian inference starts instead from probability as a degree of belief (recall subjective probability, §5.1.3) and updates it with data.

You have already done a Bayesian update. In §5.5.3, a referral-clinic patient with pre-test probability 30% had a positive test with LR+ 9.5, and the post-test probability became 80.3%: pre-test odds 0.4286 × 9.5 = 4.07, probability 4.07/5.07 = 0.803. The pre-test probability is the prior, the likelihood ratio carries the data, and the post-test probability is the posterior. The same logic applies to a whole range of parameter values.

DERIVATION

Beta–binomial updating. Let p be a prevalence. A convenient prior is the **Beta(a, b) distribution**, a distribution on 0 to 1 with density proportional to p^(a−1)(1 − p)^(b−1) and mean a/(a + b). It behaves like a earlier "successes" and b earlier "failures". The binomial likelihood of x cases in n (recall §5.7.1.2) is proportional to p^x(1 − p)^(n−x). Multiplying:

posterior ∝ p^(a+x−1)(1 − p)^(b+n−x−1), which is **Beta(a + x, b + n − x)**.

The posterior mean (a + x)/(a + b + n) is a weighted average of the prior mean and the sample proportion x/n, with weights (a + b) and n.

ILLUSTRATION / EXAMPLE

Last year's survey in a block suggested that about 40% of pregnant women attending a PHC were anaemic, with about the information of 50 women: prior Beta(20, 30). This year 30 of 60 women tested are anaemic (x/n = 0.50).

WORKED CALCULATION

1. Prior mean: 20/50 = 0.40 (95% prior interval 0.27 to 0.54).
2. Posterior: Beta(20 + 30, 30 + 30) = Beta(50, 60). Mean 50/110 = **0.455**, between the prior 0.40 and the data 0.50; the prior carries weight 50/110 = 45%.
3. 95% credible interval: the 2.5th and 97.5th centiles of Beta(50, 60), from software, are **0.363 to 0.548**. By hand, a Normal approximation works well here: the posterior SD is √[50 × 60 ÷ (110^2 × 111)] = 0.047, and 0.455 ± 1.96 × 0.047 = 0.362 to 0.547.
4. P(prevalence > 50% | data, prior) = **0.17**.
5. With a flat prior, Beta(1, 1), the posterior is Beta(31, 31): mean 0.50 and 95% credible interval 0.377 to 0.623, almost the same as the frequentist Wilson interval for 30/60 (0.377 to 0.623; recall §7.2.4). With little prior information, the two approaches give similar numbers but different interpretations.

IN R (OPTIONAL)

```
a ← 20; b ← 30                    # prior: Beta(20, 30), mean 0.40
x ← 30; n ← 60                    # new data: 30 of 60 anaemic
c(prior = a/(a + b), posterior = (a + x)/(a + b + n))
qbeta(c(0.025, 0.975), a + x, b + n - x)   # 95% credible interval
# same data, flat Beta(1, 1) prior
qbeta(c(0.025, 0.975), 1 + x, 1 + n - x)
# P(prevalence > 50% | data, prior)
1 - pbeta(0.5, a + x, b + n - x)
```

```
    prior posterior
0.4000000 0.4545455
[1] 0.3628812 0.5477760
[1] 0.376968 0.623032
[1] 0.1690927
```

The lines are steps 1–2, 3, 5 and 4. `qbeta` gives quantiles of the Beta posterior and `pbeta` its cumulative probability; nothing is simulated, because the conjugate posterior has a known form.

MUST-KNOW NOTES (MD-LEVEL)

- **The prior matters, so it must be stated and defended.** It may come from earlier studies or expert opinion, or be deliberately vague ("uninformative"). Report a sensitivity analysis with other priors, as in step 5.
- **A credible interval is a probability statement about the parameter, conditional on the prior and the model.** "Given this prior and this binomial model, there is a 95% probability that the prevalence lies between 36% and 55%." A frequentist 95% CI does not allow that sentence (recall §7.2). If the prior or model is wrong, the credible interval is wrong too.
- **Evidence accumulates naturally.** Today's posterior can be tomorrow's prior. Frequentist methods also accumulate evidence, through cumulative meta-analysis (§15.7) and group-sequential trial designs, but Bayesian updating expresses it more directly.
- **Uses in health research** include hierarchical models (for example, small-area estimates of district prevalence that borrow strength from neighbouring districts, as random effects do in §15.2.3), adaptive trial designs, and the interpretation of diagnostic tests. Most real models need simulation (Markov chain Monte Carlo) rather than the closed form above; a full treatment is beyond this book.

CHECKPOINT 15.13

1. With the prior Beta(20, 30), a survey in a second PHC finds 12 of 40 pregnant women anaemic. Give the posterior distribution and its mean. Is the posterior mean closer to the prior mean or to the sample proportion, and why?
2. A report says: "The 95% credible interval for the risk difference was 1% to 9%, so there is a 95% probability that the true risk difference lies in this range." What condition must be added for the sentence to be correct? Could the same sentence be written about a 95% confidence interval?

3. A clinician's pre-test probability of tuberculous lymphadenitis is 10%. A test with LR+ 6 is positive. Compute the post-test probability, and name the prior, the likelihood and the posterior in this calculation.

(Answers in the Appendix at the end of the book.)

15.9 Randomised Trials – Randomisation, ITT, Baseline Adjustment, Cluster and Non-Inferiority Designs

DEFINITION

A **randomised controlled trial (RCT)** assigns participants (or clusters) to interventions by a chance mechanism and compares their outcomes. Its design features protect that comparison:

- **Randomisation**: the sequence of allocations is generated by chance. **Simple** randomisation assigns each participant independently (like a coin toss); **block** (restricted) randomisation guarantees equal numbers after every block of, say, 4; **stratified** randomisation uses separate blocked sequences within strata of a strong prognostic factor. **Minimisation** allocates each new participant to the arm that minimises the imbalance, summed over chosen prognostic factors, usually with a random element.
- **Allocation concealment**: whoever enrols a participant cannot know or influence the next assignment until it is made.
- **Blinding (masking)**: participants, care providers or outcome assessors do not know the assigned arm.
- **Intention-to-treat (ITT)**: everyone randomised is analysed in the arm to which they were randomised, whatever they actually received. **Per-protocol**: only those who followed the protocol are analysed.

SIMPLIFIED EXPLANATION

Randomisation is the reason RCTs can support causal conclusions (recall §11.5 and §11.6): chance, not the clinician or the patient, decides who gets what, so the arms are alike *on average* in both measured and unmeasured characteristics. Each other feature protects what randomisation achieves. Concealment stops a recruiter who knows the next allocation from steering sicker patients away from it; without it a trial can be subverted, and inadequate concealment "leads to exaggerated estimates of treatment effect, on average" (Schulz & Grimes, 2002). Blinding stops knowledge of the arm from changing care, reporting or outcome assessment. ITT keeps the randomised groups intact: dropping people who stopped treatment breaks the comparison that randomisation made fair, because those who stop usually differ from those who continue.

Randomisation balances groups on average, not in every trial. Simple randomisation of small trials can give unequal arms; block randomisation fixes the numbers; stratification and minimisation also balance chosen prognostic factors. Minimisation "provides better balanced treatment groups" than restricted or unrestricted randomisation and handles more factors than stratification (Scott et al., 2002; Pocock & Simon, 1975).

DERIVATION

Why simple randomisation can be unbalanced. With 20 participants and a fair coin, the number in arm A is binomial ($n = 20$, $p = 0.5$; recall §5.7.1.2). P(a split of 7:13 or worse) = $2 \times P(X \le 7)$ = **0.26**. One small trial in four is this unbalanced.

Blocks of 4. A block contains two A's and two B's in random order: $C(4, 2) = 6$ possible blocks (AABB, ABAB, ABBA, BAAB, BABA, BBAA; recall combinations, §5.7.1.2). After each block the arms are equal. A fixed block size can let a recruiter guess the last allocation of a block, so block sizes are often varied at random.

Adjusting for the baseline value. Let the outcome Y (say follow-up haemoglobin) and its baseline X have the same SD σ and correlation ρ. Three analyses compare arms on:

- the final value Y: variance σ^2;
- the change Y − X: variance $\sigma^2 + \sigma^2 - 2\rho\sigma^2 = \mathbf{2\sigma^2(1 - \rho)}$;
- Y − βX with β chosen to minimise the variance, which is analysis of covariance (**ANCOVA**, a regression of Y on arm and X; recall §12.4): $\mathrm{Var}(Y - \beta X) = \sigma^2(1 + \beta^2 - 2\beta\rho)$, minimised at $\beta = \rho$, giving $\mathbf{\sigma^2(1 - \rho^2)}$.

Since $1 - \rho^2 \le 1$ and $1 - \rho^2 \le 2(1 - \rho)$ for every ρ, ANCOVA has the smallest variance. The change score beats the final value only when $\rho > 0.5$. At $\rho = 0.5$ the variances are 1, 1 and 0.75 ($\times \sigma^2$); at $\rho = 0.7$ they are 1, 0.6 and 0.51. The sample size scales with the variance, so ANCOVA at $\rho = 0.5$ needs 25% fewer participants than either alternative.

Non-inferiority. Let d = new − standard (for cure proportions, higher is better) and let −M be the **non-inferiority margin**, the largest loss of efficacy that would still be acceptable. H_0: $d \le -M$; H_1: $d > -M$ (recall §8.2). Reject H_0 at one-sided $\alpha = 0.025$ if the lower limit of the one-sided 97.5% CI exceeds −M. That lower limit is the lower limit of the usual two-sided 95% CI, so: **non-inferiority is shown if the whole 95% CI lies above −M**. **Equivalence** needs the whole CI inside (−M, +M) (ICH, 1998).

ILLUSTRATION / EXAMPLE

Baseline adjustment. A simulated trial randomises 120 anaemic adolescent girls, in blocks of 4, to iron–folic acid (IFA) with counselling or to control, with a true effect of 0.4 g/dL. The simulation sets the baseline–follow-up correlation at about 0.6; this sample happens to give 0.70.

Non-inferiority. A new, shorter treatment regimen is compared with the standard, whose cure proportion is about 90%. Clinicians agree that a loss of up to 5 percentage points would be acceptable in exchange for the shorter course: M = 5 points.

WORKED CALCULATION

1. *Trial A*, 600 per arm: cure 89.5% (new) vs 90.0% (standard). d = −0.5 points. SE = $\sqrt{(0.895 \times 0.105/600 + 0.90 \times 0.10/600)} = \sqrt{(0.0001566 + 0.0001500)} = 0.0175$, that is 1.75 points. 95% CI: $-0.5 \pm 1.96 \times 1.75$ = **−3.9 to +2.9 points**. The lower limit −3.9 lies above −5: **non-inferior**. The CI includes 0, so superiority is not shown.

2. *Trial B*, 400 per arm: 88.0% vs 90.0%. d = −2.0; SE = 2.21 points; 95% CI **−6.3 to +2.3**. The lower limit is below −5: non-inferiority is **not shown**, although the CI includes 0. "No significant difference" is not non-inferiority.
3. *Scenario E* of the figure: a CI of −4.6 to −0.4 is non-inferior (above −5) and yet significantly worse than the standard (below 0). Both statements are true; which matters depends on whether the margin was well chosen.

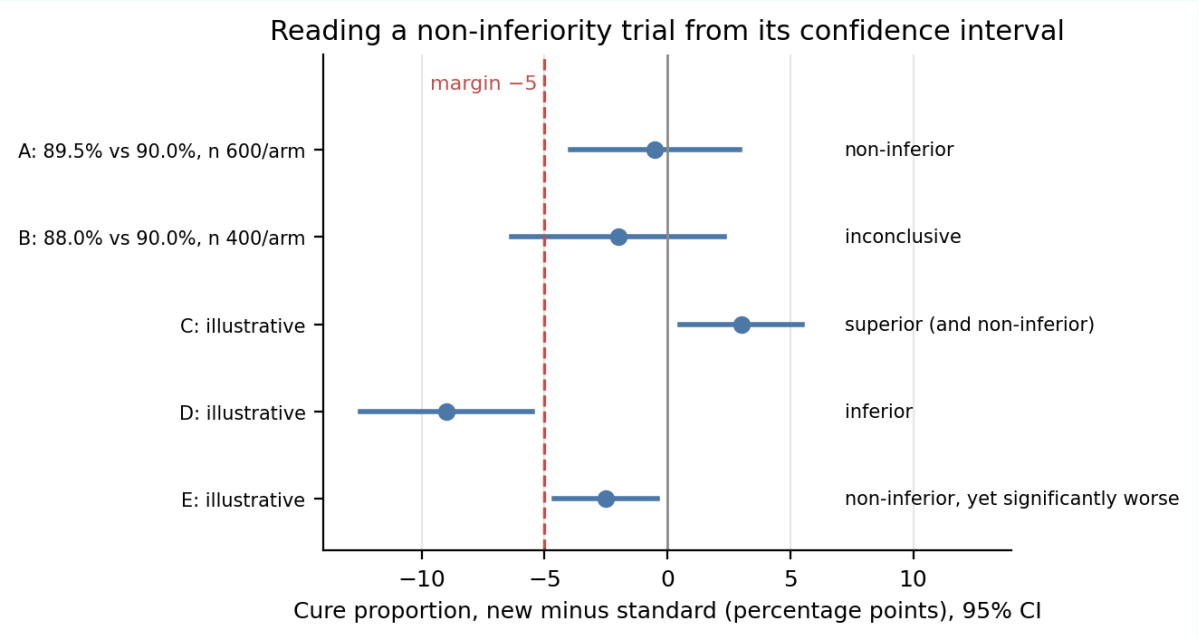


Figure 15.11. Reading a non-inferiority trial from its 95% CI for the difference in cure proportion (new − standard), margin −5 points. A and B are computed in the Worked Calculation; C–E are illustrative.

IN R (OPTIONAL)

```
# simulated trial: 120 anaemic adolescent girls
set.seed(9)
# permuted blocks of 4
blocks ← replicate(30, sample(c("A", "A", "B", "B")))
arm ← factor(as.vector(blocks), labels = c("control", "IFA"))
table(arm)
# baseline Hb, g/dL
hb0 ← rnorm(120, 10.5, 1.0)
hb1 ← 10.5 + 0.6 * (hb0 - 10.5) + 0.4 * (arm == "IFA") +
  rnorm(120, 0, 0.8)
cor(hb0, hb1)
fits ← list(final  = lm(hb1 ~ arm),
            change = lm(I(hb1 - hb0) ~ arm),
            ancova = lm(hb1 ~ arm + hb0))
t(sapply(fits, function(f) round(coef(summary(f))["armIFA", 1:2], 3)))
```

```
arm
control     IFA
     60      60
[1] 0.6999205
       Estimate Std. Error
final     0.351      0.188
change    0.204      0.144
ancova    0.246      0.135
```

`replicate(30, sample(...))` makes 30 shuffled blocks of AABB, so the arms end exactly 60:60. The sample correlation is 0.70. The three rows estimate the same true effect (0.4 g/dL) from the same girls. Their SEs, 0.188, 0.144 and 0.135, stand in the ratio 1 : 0.77 : 0.72, close to the $\sqrt{1} : \sqrt{0.6} : \sqrt{0.51}$ of the Derivation at $\rho = 0.7$. The estimates differ because, by chance, the IFA arm started 0.15 g/dL higher: the final-value analysis counts that head start as effect (0.351 = 0.204 + 0.146), the change score subtracts all of it, and ANCOVA adjusts by the fitted amount.

MUST-KNOW NOTES (MD-LEVEL)

- **Do not test baseline differences.** In a properly randomised trial every baseline difference is due to chance by definition, so a p-value for it tests a null that is known to be true. Such tests are "superfluous and could be harmful" (Schulz & Grimes, 2002) and "philosophically unsound" (Senn, 1994). Report baseline characteristics descriptively in Table 1, and adjust for prognostic variables named in the protocol whatever their baseline balance.
- **Prefer ANCOVA** for a continuous outcome measured at baseline and follow-up. Vickers and Altman (2001): analysis of covariance is "the preferred general approach" and "generally has greater statistical power"; its gain over a change score is small when the correlation is high (say $r > 0.8$).
- **ITT is the primary analysis of a superiority trial**, and it needs outcome data on everyone randomised; handle missing outcomes as in §15.4 rather than silently dropping people. A survey of trials in four major journals found ITT "often inadequately described and inadequately applied" (Hollis & Campbell, 1999). A per-protocol analysis estimates something different and loses the protection of randomisation.
- **In non-inferiority trials, ITT is not conservative.** Dropouts in both arms tend to make the arms look alike, so the full analysis set "may be biased toward demonstrating equivalence" (ICH, 1998). Both ITT and per-protocol results are therefore usually reported, and non-inferiority is more convincing when both support it; CONSORT has an extension for non-inferiority and equivalence trials (Piaggio et al., 2012). The margin is fixed in advance on clinical grounds, and it must be smaller than the benefit the standard treatment is known to have, or "non-inferior" could mean no better than no treatment.
- **Cluster-randomised trials** (villages, anganwadi centres, PHCs) are used when the intervention acts on a group or would contaminate neighbours. Plan the sample size with the design effect (§15.6 step 6) and analyse with methods that respect clustering: cluster-level summaries, mixed models or GEE (recall §15.2.4–15.2.5). Analysing individuals as if independent gives CIs that are too narrow. CONSORT has a cluster extension (Campbell et al., 2012).
- **Cross-over trials** give each participant both treatments in random order, separated by a **washout** period, and compare within person, like the paired designs of §9.2.3. They suit stable chronic conditions and short-acting treatments. Watch for a **period effect** (outcomes differ by

period whatever the treatment) and **carry-over** (the first treatment still acting in the second period).

- **Reporting.** CONSORT 2025 (Hopewell et al., 2025), a 30-item checklist with a participant flow diagram, updates CONSORT 2010 (Schulz et al., 2010). Report the randomisation method, concealment, blinding, the sample size calculation (§15.6) and the analysis population.

CHECKPOINT 15.14

1. A trial's Table 1 reports baseline weight with "p = 0.04" and the authors write that "randomisation failed". Explain why the p-value is the wrong tool and what the authors should do about a prognostically important imbalance.
2. In a planned trial of a salt-reduction intervention, the correlation between baseline and 6-month systolic BP is expected to be 0.6. A final-value analysis would need 200 participants per arm. Compute the variance factors for a change-score analysis and for ANCOVA, and the corresponding sample sizes per arm.
3. A non-inferiority trial of a community-based versus facility-based treatment for a childhood infection uses a margin of −10 percentage points. The ITT difference in cure is −3 points (95% CI −8.5 to +2.5); the per-protocol difference is −6 points (95% CI −11.8 to −0.2). What do you conclude, and why are both analyses reported?

(Answers in the Appendix at the end of the book.)

Measurement Theory – Constructs, Validity, and Reliability

Every chapter so far has assumed the data were trustworthy: a recorded weight, a recorded blood pressure, a recorded yes/no. Much Community Medicine research instead measures things no scale can read directly: satisfaction, quality of life, knowledge, stigma. This chapter asks two questions of any instrument: does it measure the right thing (validity, §16.2), and does it measure it consistently (reliability, §16.3, with agreement between two methods in §16.3.5)? It then links the two (§16.4), shows how items are compressed and combined into scores (§16.5–16.6), and ends with the full validation workflow (§16.7). No later analysis can repair a poor instrument: garbage in, garbage out.

16.1 What Is a Construct

DEFINITION

A **construct** is an abstract characteristic that cannot be observed directly. It is inferred from measurable **indicators** (questionnaire items, test scores, observed behaviours) that theory says should reflect it.

SIMPLIFIED EXPLANATION

Weight is not a construct: you put a person on a scale and read a number. “Quality of life”, “anxiety”, “trust in the health system” or “knowledge of CBHI benefits” cannot be read off any instrument. Researchers capture them indirectly. They ask a chosen set of questions, each an indicator, and combine the answers into a score that stands in for the construct.

ILLUSTRATION / EXAMPLE

“Patient satisfaction” at a PHC is a construct. A questionnaire might ask ten questions (was the waiting time acceptable? were staff respectful? were explanations clear? would you recommend this facility?) and combine the answers into one satisfaction score.

MUST-KNOW NOTES (MD-LEVEL)

- Constructs differ from directly observed variables (age, weight, blood pressure; recall §1.1). A construct needs a whole instrument, and a body of evidence, before you can trust that its score reflects the intended idea.

- Constructs that sound alike can be distinct: "patient satisfaction", "quality of care" and "trust in provider". Showing that an instrument measures the *specific* construct intended, and not an overlapping one, is the job of construct validity (§16.2.4).

16.2 Validity of a Measurement Instrument

DEFINITION

Validity is the degree to which an instrument measures the construct it claims to measure, rather than something else.

SIMPLIFIED EXPLANATION

Imagine a "satisfaction questionnaire" that mostly asks about waiting time. It may give a perfectly consistent score, but that score measures "perceived waiting time", not the broader construct it claims. Validity asks whether the instrument hits the target construct or something next to it.

ILLUSTRATION / EXAMPLE

A weighing scale that consistently reads 2 kg too high gives the same wrong answer every time. It is reliable but not valid: it does not measure true weight. §16.4 develops this contrast.

MUST-KNOW NOTES (MD-LEVEL)

- Validity asks "are we measuring the right thing?" Reliability (§16.3) asks "are we measuring it consistently?" An instrument can satisfy one without the other.
- Four kinds of validity evidence follow: content (§16.2.1), face (§16.2.2), criterion (§16.2.3) and construct (§16.2.4). No single kind is enough. A well-validated instrument gathers evidence of several kinds, in the order set out in §16.7.

16.2.1 Content Validity

DEFINITION

Content validity is the degree to which an instrument's items cover every relevant domain of the construct, without leaving out important dimensions or including irrelevant ones.

SIMPLIFIED EXPLANATION

The WHOQOL-BREF, the short form of the WHO quality-of-life instrument, scores four domains: physical health, psychological, social relationships and environment (WHOQOL Group, 1998). A quality-of-life questionnaire with content validity must touch each domain its framework defines. Leaving one out under-represents the construct, however well the other items are written.

ILLUSTRATION / EXAMPLE

A "quality of life" questionnaire that asks only about physical symptoms, with nothing on social relationships or psychological well-being, lacks content validity. It measures something closer to "physical symptom burden".

MUST-KNOW NOTES (MD-LEVEL)

- Assessed by expert judgement: a panel reviews each item against a defined content domain (taken from the literature or a theoretical framework) and rates its relevance and the overall coverage.
- It is judgement-based evidence, gathered during questionnaire design, before data collection.
- It is necessary but not sufficient. An item's presence does not guarantee it works once fielded.

16.2.2 Face Validity

DEFINITION

Face validity is the degree to which an instrument *appears*, on surface inspection by ordinary (often non-expert) reviewers, to measure what it claims to.

SIMPLIFIED EXPLANATION

Face validity asks only whether the questionnaire *looks* right to someone glancing at it. It is judged by intuition, not by a systematic or statistical process. It still matters: respondents who feel the questions make sense are more likely to answer honestly and completely.

ILLUSTRATION / EXAMPLE

A "Diabetes Knowledge Assessment" that visibly asks about blood sugar, insulin and diet has strong face validity, even before any expert review or statistical testing.

MUST-KNOW NOTES (MD-LEVEL)

- The weakest validity evidence: surface appearance, not systematic expert review (content validity) or data (criterion and construct validity).
- Poor face validity can reduce cooperation. Respondents who cannot see why a question is asked may answer carelessly or refuse.
- Content validity is a systematic, expert check of domain coverage. Face validity is "does this look reasonable?", and anyone can judge it.

CHECKPOINT 16.1

1. A new depression screening questionnaire for elderly rural patients is adapted word-for-word from a validated questionnaire for urban working-age adults, without any local review. Which type of validity is most directly at risk, and why?

2. A clinician glances at a new anxiety scale and says "these questions look like they measure anxiety." Which more limited type of validity does this comment support?

(Answers in the Appendix at the end of the book.)

16.2.3 Criterion Validity

DEFINITION

Criterion validity is the degree to which scores from an instrument agree with an external, independently measured standard (the **criterion**) for the same or a closely related construct.

SIMPLIFIED EXPLANATION

Criterion validity asks a direct, testable question: does the new score line up with an already-trusted measurement of the same thing? The statistic depends on the data. For two continuous scores, use a correlation (recall §11.2 for Pearson's r, §11.3 for Spearman's rank correlation). When the criterion is a yes/no diagnosis, use sensitivity, specificity and the area under the ROC curve (recall §5.5.1 and §15.5.1). When the question is whether a new method can *replace* the old one, use agreement statistics: kappa for categories (§16.3.3.1, which corrects agreement for chance) or limits of agreement for continuous readings (§16.3.5, the range expected to contain 95% of the differences between the two methods).

ILLUSTRATION / EXAMPLE

Scores on a quick depression screening questionnaire are compared with the diagnosis from a long, established clinical interview given to the same patients. Close agreement is criterion validity evidence for the quick tool.

MUST-KNOW NOTES (MD-LEVEL)

- It needs a trusted criterion. For many new constructs none exists, and criterion validity cannot be assessed.
- It comes in two forms, set apart by timing: concurrent (§16.2.3.1) and predictive (§16.2.3.2).

16.2.3.1 Concurrent Validity

DEFINITION

Concurrent validity is criterion validity with the criterion measured at about the same time as the new instrument.

SIMPLIFIED EXPLANATION

It asks whether a quicker, cheaper or newer instrument agrees with the gold standard now, on the same patients. No waiting for a future outcome is needed.

ILLUSTRATION / EXAMPLE

A brief anxiety screening tool and a long clinical anxiety interview are given to the same 100 patients on the same day, and the two sets of scores are correlated.

MUST-KNOW NOTES (MD-LEVEL)

- The usual evidence when a shorter, cheaper or less invasive alternative to a burdensome gold standard is being developed.

16.2.3.2 Predictive Validity

DEFINITION

Predictive validity is criterion validity with the criterion (typically a future outcome) measured later.

SIMPLIFIED EXPLANATION

It asks whether a score measured today predicts something that happens later. That is the whole purpose of a screening or risk-prediction tool.

ILLUSTRATION / EXAMPLE

A screening tool is given to 200 pregnant women at their first antenatal visit to predict postpartum depression. Three months after delivery, 30 have postpartum depression on a clinical interview. Of these 30, 24 had screened positive. Of the 170 who did not develop it, 34 had screened positive.

WORKED CALCULATION

1. Sensitivity = 24 ÷ 30 = **0.80** (recall §5.5.1).
2. Specificity = (170 − 34) ÷ 170 = 136 ÷ 170 = **0.80**.
3. Of the 58 women who screened positive, 24 developed depression: PPV = 24 ÷ 58 = **0.41**. Of the 142 who screened negative, 136 did not: NPV = 136 ÷ 142 = **0.96**.

The binary criterion calls for sensitivity and specificity, not r. If the tool gives a continuous score, the ROC curve and its AUC (recall §15.5.1) summarise predictive validity across every cut-off.

MUST-KNOW NOTES (MD-LEVEL)

- It needs longitudinal follow-up, so it costs more time and money than concurrent validation.
- With a binary future outcome, report sensitivity, specificity, predictive values or AUC. A relative risk (high scorers against low scorers) is also used. Correlation is used only when the future criterion is continuous.

16.2.4 Construct Validity

DEFINITION

Construct validity is the degree to which an instrument's scores behave as the theory of the construct predicts: they relate to other measures, differ between groups and hang together internally in the ways the theory expects.

SIMPLIFIED EXPLANATION

Without a single gold standard, you test the instrument against a web of predictions made *before* looking at the data. Terwee et al. (2007) put it this way: construct validity "should be assessed by testing predefined hypotheses concerning the concepts that are being measured." The COSMIN consensus places three strands under it: structural validity (does the factor structure match the theory; §16.5), hypotheses testing, and cross-cultural validity (Mokkink et al., 2010). Hypotheses testing has two common forms:

- **Correlational hypotheses**: strong correlation with measures of the same or related constructs, weaker correlation with measures of related but distinct constructs (§16.2.4.1).
- **Known-groups hypotheses**: the instrument "can discriminate between two groups known to differ on the variable of interest" (Davidson, 2014).

ILLUSTRATION / EXAMPLE

A new Hindi health-literacy questionnaire is tested against four predefined hypotheses. (1) It will correlate strongly ($r > 0.6$) with an established health-literacy tool. (2) It will correlate only moderately (r 0.3–0.5) with a general reading-ability test, because reading is related to health literacy but distinct from it. (3) ASHA workers will score higher than women from the same villages who have no health training. (4) Principal component analysis (§16.5: a method that finds the few weighted combinations of items that carry most of their shared variation) will show one dominant dimension. Each confirmed hypothesis adds evidence. A correlation of 0.9 with reading ability would count *against* the instrument: it would suggest the questionnaire measures reading, not health literacy.

MUST-KNOW NOTES (MD-LEVEL)

- Construct validity is built up across several studies, not settled by one test.
- Write the hypotheses down, with expected directions and magnitudes, before analysis. Post hoc "confirmation" is weak evidence.
- In the modern view (COSMIN), content, criterion and construct evidence all serve one question: do the scores mean what we claim?

16.2.4.1 Convergent and Discriminant Validity

DEFINITION

Convergent validity is evidence that an instrument correlates strongly with other measures of the same or closely related constructs. **Discriminant (divergent) validity** is evidence that it correlates only weakly, or less strongly, with measures of *related but distinct* constructs. They are two strands of construct validity. Known-groups and structural evidence are others.

SIMPLIFIED EXPLANATION

Convergent validity is the "should correlate" half. Discriminant validity is the "should not correlate as much" half. The test of discriminant validity is only informative against a construct that *could* plausibly be confused with yours. A health-literacy score that does not correlate with height proves nothing, because nobody thought it might. A health-literacy score that correlates much less with general reading ability than with another health-literacy tool shows that the instrument is not just a reading test.

ILLUSTRATION / EXAMPLE

A new medication adherence scale correlates at r = 0.62 with pharmacy refill records (convergent). It correlates at r = 0.25 with a general health-beliefs scale, a related but distinct construct that it might have been confused with (discriminant). Together these support the claim that the scale measures adherence, not a broader attitude to health.

MUST-KNOW NOTES (MD-LEVEL)

- Convergent evidence alone cannot rule out the possibility that the instrument measures some broader characteristic that also correlates with the criterion.
- Choose comparison constructs for discriminant validity that a critic could reasonably confuse with yours.

CHECKPOINT 16.2

1. A new physical activity scale correlates at r = 0.85 with an existing validated physical activity scale, and at r = −0.30 with a sedentary-behaviour screen (a related but distinct construct). Which strand of validity does each finding support? Why is the second comparison more informative than a correlation with shoe size would be?
2. A new scale to predict medication adherence correlates well with adherence measured six months later. Which form of validity is this, and how does it differ from correlating with adherence measured on the same day?
3. A tool predicting low birthweight is given at 20 weeks to 250 women. Of 40 babies later born under 2.5 kg, 30 mothers had screened positive; of 210 others, 42 had screened positive. Compute the sensitivity and specificity. Why are these reported instead of r?

(Answers in the Appendix at the end of the book.)

16.3 Reliability of a Measurement Instrument

DEFINITION

Reliability is the degree to which an instrument gives consistent, reproducible results: across occasions, across items meant to measure the same thing, or across raters. It says nothing about whether the instrument measures the right thing.

In classical test theory each observed score X is a true score T plus a random error E, uncorrelated with T:

$X = T + E$, so $\sigma^2_X = \sigma^2_T + \sigma^2_E$

The **reliability coefficient** is the share of observed-score variance that is true-score variance:

$r_{xx} = \sigma^2_T \div \sigma^2_X = \sigma^2_T \div (\sigma^2_T + \sigma^2_E)$

SIMPLIFIED EXPLANATION

The miscalibrated scale of §16.2 reads 2 kg too high every time. That consistency *is* reliability. Reliability of 0.80 means that 80% of the spread in observed scores reflects real differences between people and 20% reflects noise. It runs from 0 (all noise) to 1 (no noise).

ILLUSTRATION / EXAMPLE

A questionnaire given twice, a week apart, to the same stable patient should give nearly the same score both times if it is reliable, whether or not that score reflects the intended construct.

MUST-KNOW NOTES (MD-LEVEL)

- Reliability is consistency, not correctness. §16.4 shows that it caps validity.
- Four types follow, each asking "consistent across what?": items within one administration (internal consistency, §16.3.1); the same instrument on two occasions (test-retest, §16.3.2); different raters (inter-rater, §16.3.3); two equivalent versions (parallel forms, §16.3.4). §16.3.5 then asks a related question: do two measurement methods agree?
- Reliability is a property of scores in a particular population, not of the instrument alone. The same instrument is less reliable in a population where true scores vary less, because σ^2_T is smaller.

16.3.1 Internal Consistency Reliability

DEFINITION

Internal consistency reliability is the degree to which the items of one scale, all meant to measure the same construct, correlate with each other in a single administration.

SIMPLIFIED EXPLANATION

If a 10-item satisfaction scale measures one construct, someone who agrees that "staff were respectful" should tend to agree that "I would recommend this facility". Internal consistency checks whether the items hang together as they should.

ILLUSTRATION / EXAMPLE

If half the items of a satisfaction scale correlate strongly with each other but poorly with the other half, either the scale measures two constructs (a dimensionality problem that principal component analysis, §16.5, can show) or some items are badly worded.

MUST-KNOW NOTES (MD-LEVEL)

- It needs only one administration. Test-retest needs two.

- The **item-total correlation** is the correlation between one item and the total score. The **corrected** item-total correlation (SPSS "corrected item-total correlation", R `r.drop`) correlates the item with the total of the *other* items, so the item does not correlate with itself. A value near zero or negative marks an item that does not belong or was not reverse-coded. Tavakol & Dennick (2011) advise that items whose correlation with the total is low, "approaching zero", are deleted.
- Low internal consistency can mean the scale is not unidimensional, items are poorly worded, or there are too few items. Each needs a different fix.

16.3.1.1 Cronbach's Alpha

DEFINITION

Cronbach's alpha estimates internal consistency reliability from the item variances and the variance of the total score. For a scale of k items, with item variances s_i^2 and total-score variance s_T^2:

$$\alpha = [k \div (k - 1)] \times (1 - \Sigma s_i^2 \div s_T^2)$$

(Bland & Altman, 1997). **Standardised alpha** uses the mean inter-item correlation $\bar{r}$ instead:

$$\alpha_std = k\,\bar{r} \div [1 + (k - 1)\,\bar{r}]$$

SIMPLIFIED EXPLANATION

The total score's variance is the sum of the item variances *plus* twice every covariance between items. If the items measure one construct, they covary positively, the covariances make up a large share of the total variance, and $\Sigma s_i^2 \div s_T^2$ is small, so α is high. If the items are unrelated, the covariances are near zero, $s_T^2 \approx \Sigma s_i^2$, and α is near 0. If the items covary *negatively*, s_T^2 is smaller than Σs_i^2, the bracket goes negative, and so does α.

DERIVATION

1. $T = X_1 + X_2 + \ldots + X_k$. Expanding $\mathrm{Var}(T) = E[(T - E[T])^2]$ gives the variance of a sum: $\mathrm{Var}(T) = \Sigma \mathrm{Var}(X_i) + 2\,\Sigma_{(i<j)}\,\mathrm{Cov}(X_i, X_j)$ (covariance recalled from §11.2).
2. So $1 - \Sigma s_i^2 \div s_T^2 = (2\,\Sigma \text{ covariances}) \div s_T^2$, the share of total variance that comes from items moving together.
3. There are $k(k-1)/2$ covariances but k variances. The factor $k/(k-1)$ rescales so that $\alpha = 1$ when all items are perfectly correlated with equal variances.
4. With standardised items (each variance 1, each covariance $\bar{r}$): $s_T^2 = k + k(k-1)\,\bar{r}$. Substituting gives $\alpha_std = k\,\bar{r} \div [1 + (k-1)\,\bar{r}]$. This is the **Spearman–Brown formula**, the reliability of a total of k parallel items each correlating $\bar{r}$.

ILLUSTRATION / EXAMPLE

Five patients answer three satisfaction items on a 1–5 scale:

Patient	Item 1	Item 2	Item 3	Total
1	4	5	4	13

Patient	Item 1	Item 2	Item 3	Total
2	5	4	5	14
3	3	4	2	9
4	2	2	3	7
5	1	2	1	4

WORKED CALCULATION

1. Item means 3.0, 3.4, 3.0. Sums of squared deviations 10.0, 7.2, 10.0, so the variances (÷ 4) are s_1^2 = 2.5, s_2^2 = 1.8, s_3^2 = 2.5. Σs_i^2 = **6.8**.
2. Totals 13, 14, 9, 7, 4: mean 9.4, sum of squared deviations 69.2, s_T^2 = 69.2 ÷ 4 = **17.3**.
3. Check: covariances 1.75, 2.25 and 1.25, and 6.8 + 2 × (1.75 + 2.25 + 1.25) = 6.8 + 10.5 = 17.3.
4. α = (3 ÷ 2) × (1 − 6.8 ÷ 17.3) = 1.5 × (1 − 0.393) = 1.5 × 0.607 = **0.91**.
5. Inter-item correlations are 0.825, 0.900 and 0.589, so r̄ = 0.771. α_std = 3 × 0.771 ÷ (1 + 2 × 0.771) = 2.314 ÷ 2.543 = **0.91**. It matches raw α here because the item variances are similar.
6. **Negative alpha.** Suppose item 3 were a reverse-worded item ("I felt ignored") entered without recoding, so each score is 6 minus the value shown. Its variance stays 2.5, but its covariances with items 1 and 2 become −2.25 and −1.25. Then s_T^2 = 6.8 + 2 × (1.75 − 2.25 − 1.25) = 3.3, smaller than Σs_i^2, and α = 1.5 × (1 − 6.8 ÷ 3.3) = 1.5 × (−1.061) = **−1.59**. A negative α is a red flag, usually for an item that was not reverse-coded.

Alpha and scale length. Hold r̄ at 0.3. Spearman–Brown gives α_std = 0.56 for 3 items, 0.81 for 10 and 0.90 for 20. Alpha rises with k *only while* r̄ is held. Add five items that correlate only 0.05 with the rest to a 10-item scale (r̄ = 0.3). The mean inter-item correlation of the 15 items falls to (45 × 0.3 + 60 × 0.05) ÷ 105 = 0.157, and α_std falls from 0.81 to **0.74**.

IN R (OPTIONAL)

```
library(psych)
sat ← data.frame(item1 = c(4, 5, 3, 2, 1),
                 item2 = c(5, 4, 4, 2, 2),
                 item3 = c(4, 5, 2, 3, 1))
alpha(sat)
```

```
  raw_alpha std.alpha G6(smc) average_r S/N   ase mean  sd median_r
      0.91      0.91    0.95      0.77  10 0.072  3.1 1.4     0.82
...
 Reliability if an item is dropped:
      raw_alpha std.alpha G6(smc) average_r  S/N alpha se var.r med.r
item1      0.74      0.74    0.59      0.59  2.9    0.232    NA  0.59
item2      0.95      0.95    0.90      0.90 18.0    0.047    NA  0.90
item3      0.90      0.90    0.82      0.82  9.4    0.088    NA  0.82

 Item statistics
      n raw.r std.r r.cor r.drop mean  sd
item1 5  0.99  0.99  0.99   0.97  3.0 1.6
item2 5  0.86  0.87  0.82   0.73  3.4 1.3
item3 5  0.91  0.90  0.88   0.79  3.0 1.6
```

`raw_alpha` 0.91 is step 4 and `std.alpha` 0.91 is step 5, with `average_r` 0.77 as $\bar{r}$. The **alpha-if-item-deleted** column recomputes α with each item left out. Dropping item 2 would *raise* α to 0.95, and item 2 has the lowest corrected item-total correlation (`r.drop` 0.73). That makes it the first candidate for review, though with 5 respondents this is only a teaching example.

MUST-KNOW NOTES (MD-LEVEL)

- Alpha usually lies between 0 and 1, but it has no lower bound. Streiner (2003) lists "restricted to the range of 0 to 1" among four myths about alpha. A negative value means the items covary negatively on average: check reverse-coding first.
- Conventional thresholds (conventions, not laws): α of 0.7 to 0.8 is "satisfactory" for comparing groups, and at least 0.90 is wanted for decisions about individual patients (Bland & Altman, 1997). Terwee et al. (2007) rate internal consistency positively when α lies between 0.70 and 0.95. A very high α (above about 0.90 to 0.95) may mean redundant items that ask the same question in different words (Tavakol & Dennick, 2011).
- Holding the mean inter-item correlation constant, α rises with the number of items (Spearman–Brown). A long scale can reach a high α with weak items. Adding poorly correlated items can *lower* α.
- Alpha assumes one dimension. A scale made of two unrelated sub-scales can still show a respectable α (Tavakol & Dennick, 2011). Check dimensionality first (§16.5) and report α for each sub-scale.
- Alpha belongs to the scores in one sample, not to the instrument. Compute it again in each new study population.

16.3.2 Test-Retest Reliability

DEFINITION

Test-retest reliability is the degree to which the same instrument, given to the same unchanged respondents on two occasions, gives consistent scores. For continuous scores it is quantified by an **intraclass correlation coefficient (ICC)**. For categorical scores it is quantified by kappa (agreement corrected for chance, §16.3.3.1).

SIMPLIFIED EXPLANATION

Internal consistency looks for agreement *within* one administration, across items. Test-retest looks for agreement *across time* with the same instrument, assuming nothing real has changed. The ICC (recall §15.2.1, where it measured clustering) is the share of total variance due to real differences between subjects:

ICC = σ^2_subjects ÷ (σ^2_subjects + σ^2_other)

The forms differ in what counts as "other" (Shrout & Fleiss, 1979; the documentation of R's psych package maps them to SPSS labels):

Form	Design	What counts as error	SPSS label
ICC(1)	One-way: each subject measured by different, random occasions or raters	All within-subject variation, including any systematic shift	One-way random, absolute agreement
ICC(2,1)	Two-way random: occasions or raters are a sample from a wider set	Random error **plus** systematic shift between occasions	Two-way random, absolute agreement
ICC(3,1)	Two-way mixed: these particular occasions or raters are fixed	Random error only; a constant shift is ignored	Two-way mixed, consistency

"k" versions (ICC(1,k), ICC(2,k), ICC(3,k)) give the reliability of the *average* of k measurements.

DERIVATION

Lay out n subjects × k occasions and partition the sums of squares as in two-way ANOVA (recall §9.5.1 for mean squares and §9.5.4 for the two-way layout). MS_S is the between-subjects mean square, MS_O the between-occasions mean square, MS_E the residual, and MS_W the pooled within-subject mean square (occasion plus residual):

$$ICC(1) = (MS_S - MS_W) \div [MS_S + (k - 1)\, MS_W]$$

$$ICC(2,1) = (MS_S - MS_E) \div [MS_S + (k - 1)\, MS_E + k\,(MS_O - MS_E) \div n]$$

$$ICC(3,1) = (MS_S - MS_E) \div [MS_S + (k - 1)\, MS_E]$$

Only ICC(2,1) and ICC(1) carry the occasion effect in the denominator. A systematic shift between occasions lowers them. It leaves ICC(3,1) untouched.

ILLUSTRATION / EXAMPLE

A knowledge questionnaire is given to 5 ASHA workers twice, a week apart. Between the two sessions, their supervisor unexpectedly discussed the topic with them.

Worker	Week 1	Week 2	Difference
1	10	14	4
2	14	19	5
3	18	22	4
4	22	27	5
5	26	30	4

WORKED CALCULATION

1. Grand mean 20.2. Subject means 12, 16.5, 20, 24.5, 28. Occasion means 18.0 and 22.4.
2. $SS_S = 2 \times \Sigma(\text{subject mean} - 20.2)^2 = 2 \times 160.3 = 320.6$, df 4, so $MS_S = \mathbf{80.15}$.
3. $SS_O = 5 \times [(18.0 - 20.2)^2 + (22.4 - 20.2)^2] = 5 \times 9.68 = 48.4$, df 1, so $MS_O = \mathbf{48.4}$.
4. $SS_total = 369.6$, so $SS_E = 369.6 - 320.6 - 48.4 = 0.6$, df 4, and $MS_E = \mathbf{0.15}$. $MS_W = (48.4 + 0.6) \div 5 = \mathbf{9.8}$.
5. $ICC(3,1) = (80.15 - 0.15) \div (80.15 + 0.15) = 80.0 \div 80.3 = \mathbf{0.996}$.
6. $ICC(2,1) = 80.0 \div [80.3 + 2 \times (48.4 - 0.15) \div 5] = 80.0 \div (80.3 + 19.3) = 80.0 \div 99.6 = \mathbf{0.80}$.
7. $ICC(1) = (80.15 - 9.8) \div (80.15 + 9.8) = 70.35 \div 89.95 = \mathbf{0.78}$.

8. Pearson's r between the weeks is **0.996**.

Every worker scored 4 or 5 points higher in week 2. The ranking was preserved, so r and the consistency ICC(3,1) are almost 1. The absolute-agreement ICC(2,1) drops to 0.80 because a respondent's week-1 score does *not* reproduce their week-2 score. For test-retest reliability, where a score should repeat, absolute agreement is the relevant question.

IN R (OPTIONAL)

```
library(psych)
retest ← data.frame(week1 = c(10, 14, 18, 22, 26),
                    week2 = c(14, 19, 22, 27, 30))
cor(retest$week1, retest$week2)
ICC(retest)
```

```
[1] 0.996271
...
                         type  ICC     F df1 df2     p lower bound upper bound
Single_raters_absolute   ICC1 0.78   8.2   4   5 2e-02      0.0508        0.97
Single_random_raters     ICC2 0.80 534.3   4   4 1e-05     -0.0022        0.98
Single_fixed_raters      ICC3 1.00 534.3   4   4 1e-05      0.9647        1.00
...
```

The three single-measure rows reproduce steps 5–7 (ICC3 prints as 1.00 after rounding 0.996). The confidence interval for ICC2 runs from about 0 to 0.98: with 5 subjects almost nothing is pinned down.

MUST-KNOW NOTES (MD-LEVEL)

- Choose the interval with care. Too short, and respondents remember and repeat their answers, which *inflates* the estimate. Too long, and real change in the construct (or, as here, an intervening event) *deflates* it.
- Use only for constructs expected to stay stable over the interval. Retesting momentary mood measures real change, not unreliability.
- Always name the ICC form (model, type, single or average). Koo & Li (2016) note that there are 10 forms of ICC and ask authors to report "model", "type" and "definition". They also propose bands based on the 95% CI: below 0.5 poor, 0.5–0.75 moderate, 0.75–0.9 good, above 0.9 excellent (a convention). Terwee et al. (2007) rate reliability positively when the ICC or weighted kappa (kappa with partial credit for near misses, §16.3.3.1) is at least 0.70 in at least 50 patients.
- **Within-subject SD and repeatability.** When there is no systematic shift, the within-subject SD is $s_w = \sqrt{MS_E}$. The difference between two measurements of one person has SD $\sqrt{2} \times s_w$. So 95% of repeat differences fall within $1.96 \times \sqrt{2} \times s_w =$ **2.77 s_w**, the **repeatability coefficient**.

16.3.3 Inter-Rater Reliability

DEFINITION

Inter-rater reliability is the degree to which two or more independent raters, assessing the same subjects with the same instrument, agree.

SIMPLIFIED EXPLANATION

The attention moves from the instrument to the people using it. A well-designed instrument can still give inconsistent results if one nurse grades a wound "mild" and another grades the same wound "moderate".

ILLUSTRATION / EXAMPLE

Two community health workers independently classify the nutritional status of the same 50 children with the same criteria. How often, and how closely, do they agree?

MUST-KNOW NOTES (MD-LEVEL)

- Test-retest compares the same instrument on two occasions. Inter-rater compares different raters on the same subjects at one time.
- Categorical ratings: Cohen's kappa (§16.3.3.1). Continuous ratings: the ICC, in the forms of §16.3.2, with raters in place of occasions.
- Poor agreement usually points to training, instructions or unclear scoring criteria.

16.3.3.1 Cohen's Kappa

DEFINITION

Cohen's kappa measures agreement between two raters on categorical ratings, corrected for the agreement expected by chance:

$\kappa = (p_o - p_e) \div (1 - p_e)$

where p_o is the observed proportion of agreement and p_e is the proportion expected if the two raters rated independently at their own observed rates.

SIMPLIFIED EXPLANATION

Two raters agree some of the time by chance alone, especially with few categories. Kappa removes that part. The numerator is agreement beyond chance. The denominator is the most agreement beyond chance that was possible.

DERIVATION

If the raters rate independently, the probability that both say "anaemic" is (A's proportion anaemic) × (B's proportion anaemic), by the multiplication rule for independent events (recall §5.2.2). The same holds for "not anaemic". Adding over categories:

$p_e = \Sigma$ (row proportion × column proportion)

ILLUSTRATION / EXAMPLE

Two medical officers independently classify 100 women attending an antenatal clinic as anaemic or not anaemic, by clinical pallor alone.

	B: anaemic	B: not anaemic	Row total
A: anaemic	11	10	21
A: not anaemic	5	74	79
Column total	16	84	100

WORKED CALCULATION

1. $p_o = (11 + 74) \div 100 =$ **0.85**.
2. A calls 21% anaemic and B calls 16%. Chance agreement on "anaemic" = 0.21 × 0.16 = 0.0336. On "not anaemic" = 0.79 × 0.84 = 0.6636.
3. $p_e = 0.0336 + 0.6636 =$ **0.697** (about 0.70).
4. $\kappa = (0.85 - 0.697) \div (1 - 0.697) = 0.153 \div 0.303 =$ **0.50** (0.505 to 3 dp).

An 85% raw agreement shrinks to κ = 0.50. Most of the agreement is on "not anaemic", which both raters say most of the time, so chance alone would produce much of it.

Interpreting kappa. Landis & Koch (1977, p. 165) proposed these labels, reproduced in the Stata manual:

κ	Strength of agreement
below 0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

On this scale κ = 0.50 is "moderate". The bands are a convention. The 95% CI, 0.29 to 0.72, comes from a large-sample SE of κ that is usually left to software; the rough hand version, $SE \approx \sqrt{[p_o(1 - p_o)/n]} \div (1 - p_e) = 0.118$, gives 0.27 to 0.74. The interval runs from "fair" to "substantial", so the label is uncertain.

The prevalence paradox. Keep $p_o = 0.85$ and change only how common the condition is:

Table (a / b / c / d)	A "yes"	B "yes"	p_e	κ
45 / 10 / 5 / 40 (balanced)	0.55	0.50	0.50	**0.70**
11 / 10 / 5 / 74 (this clinic)	0.21	0.16	0.697	**0.50**
80 / 10 / 5 / 5 (very common)	0.90	0.85	0.78	**0.32**

The same 85% agreement gives κ from 0.70 to 0.32. When one category dominates, p_e is high and kappa is low: "a high value of p_0 can be drastically lowered by a substantial imbalance in the table's marginal totals" (Feinstein & Cicchetti, 1990).

The bias paradox. Keep $p_o = 0.85$ and the clinic's 11 and 74 agreements, but move all 15 disagreements to one side (A says anaemic, B not: 11 / 15 / 0 / 74). A now calls 26% anaemic and B 11%. $p_e = 0.26 \times 0.11 + 0.74 \times 0.89 = 0.687$, and $\kappa = (0.85 - 0.687) \div 0.313 =$ **0.52**. Kappa *rises* even though the raters now disagree systematically. Byrt, Bishop & Carlin (1993) recommend reporting "quantitative indicators of bias and prevalence" alongside kappa, such as the two raters' marginal proportions. A related index, the **prevalence-adjusted bias-adjusted kappa** (PABAK), recomputes kappa as if p_e were 0.5 (balanced categories, no bias): $\text{PABAK} = (p_o - 0.5) \div 0.5 = 2p_o - 1$, here $2 \times 0.85 - 1 = 0.70$.

Weighted kappa for ordered categories. Plain kappa treats "normal vs severe" as no worse than "normal vs moderate". **Weighted kappa** (Cohen, 1968) gives partial credit to near misses. Two anganwadi supervisors grade 60 children as normal, moderate acute malnutrition (MAM) or severe acute malnutrition (SAM):

	B: normal	B: MAM	B: SAM	Total
A: normal	30	4	0	34
A: MAM	3	12	2	17
A: SAM	1	2	6	9
Total	34	18	8	60

1. Unweighted: $p_o = 48 \div 60 = 0.80$. Expected diagonal counts (row × column ÷ 60) are 19.27, 5.10 and 1.20, so $p_e = 25.57 \div 60 = 0.426$, and $\kappa = (0.80 - 0.426) \div 0.574 =$ **0.65**.
2. Linear agreement weights: 1 on the diagonal, 0.5 one category apart, 0 two apart. Observed weighted agreement = $(48 + 0.5 \times 11 + 0 \times 1) \div 60 = 53.5 \div 60 = 0.892$.
3. Expected: the one-apart cells have expected counts 10.2, 9.63, 2.27 and 2.70 (total 24.8). $p_e(w) = (25.57 + 0.5 \times 24.8) \div 60 = 37.97 \div 60 = 0.633$.
4. $\kappa_w = (0.892 - 0.633) \div (1 - 0.633) =$ **0.70**.

Eleven of the twelve disagreements are only one grade apart, so the weighted kappa is higher. With quadratic weights (the default in R's psych package) κ_w is 0.76. State which weights you used.

IN R (OPTIONAL)

```
library(psych)
k ← read.csv("data/kappa_anaemia.csv")
tab ← table(rater_A = k$rater_A, rater_B = k$rater_B)
tab
cohen.kappa(tab)

whz ← matrix(c(30, 4, 0,
               3, 12, 2,
               1, 2, 6), nrow = 3, byrow = TRUE)
cohen.kappa(whz, w = abs(outer(1:3, 1:3, "-")) / 2)   # linear weights
```

```
              rater_B
rater_A       anaemic not anaemic
  anaemic          11          10
  not anaemic       5          74
...
                 lower estimate upper
unweighted kappa  0.29      0.5  0.72
...
                 lower estimate upper
unweighted kappa  0.48     0.65  0.82
weighted kappa    0.54     0.70  0.87
```

The first table is the 2×2 above, and the kappa estimate 0.5 (95% CI 0.29 to 0.72) is step 4. For the malnutrition grades, the unweighted 0.65 and linearly weighted 0.70 match steps 1–4. The `w` argument holds *disagreement* weights (0 on the diagonal, 0.5 and 1 off it), the complement of the agreement weights used by hand.

MUST-KNOW NOTES (MD-LEVEL)

- $\kappa = 1$ is perfect agreement, 0 is chance-level agreement, and negative values mean agreement worse than chance. The lower limit depends on the marginals and is not always −1.
- Report p_o, both raters' marginal proportions and κ with its CI, not κ alone (Byrt et al., 1993).
- More than two raters: **Fleiss' kappa** extends the chance correction to more than two raters. The raters need not be the same people for every subject (Stata `kappa` manual).
- Kappa and McNemar's test (recall §10.8) answer different questions about the same 2×2 table. McNemar asks whether one rater says "yes" more often than the other (bias, using only the discordant cells b and c). Kappa asks how well they agree beyond chance.

16.3.4 Parallel-Forms Reliability

DEFINITION

Parallel-forms reliability is the correlation between scores on two different but equivalent versions of an instrument given to the same respondents.

SIMPLIFIED EXPLANATION

Retesting with the *same* items lets respondents remember answers and practise, which inflates test-retest reliability. Two forms with different but equivalent items remove memory and practice effects. They also expose error from **item sampling**: how much a score depends on which particular items were chosen.

ILLUSTRATION / EXAMPLE

Two equivalent 20-item knowledge tests (Form A and Form B) are given to the same medical students. A high correlation supports parallel-forms reliability. The two forms can then be used for pre- and post-training tests without students simply recalling Form A.

MUST-KNOW NOTES (MD-LEVEL)

- Building two equivalent forms is a heavy development burden, which limits the use of this design.

CHECKPOINT 16.3

1. A researcher gives the same questionnaire to the same patients twice, two weeks apart, and finds a high correlation. Which type of reliability is this? Name one threat that would inflate the estimate and one that would deflate it, even for a well-designed instrument.
2. A four-item scale has item variances 1.2, 1.0, 1.5 and 1.3 and total-score variance 12.0. Compute Cronbach's alpha. A ten-item scale has $\alpha = 0.97$. What might this very high value signal?
3. Two nurses classify 200 children as "fever" or "no fever" by touch. Both say fever for 30, neither for 140, nurse A only for 20, nurse B only for 10. Compute p_o, p_e and κ, and give the Landis & Koch label.

(Answers in the Appendix at the end of the book.)

16.3.5 Agreement Between Two Methods – Bland–Altman

DEFINITION

A **Bland–Altman analysis** asks whether two methods of measuring the same continuous quantity agree well enough to be used interchangeably. For each subject, compute the difference d between the two methods. Report the **bias** (the mean difference, $\bar{d}$) and the **95% limits of agreement**:

$$\bar{d} \pm 1.96 \times s_d$$

where s_d is the SD of the differences (Bland & Altman, 1986). The limits give the range within which about 95% of differences between the two methods are expected to lie. The **Bland–Altman plot** graphs each difference against the mean of the two readings.

SIMPLIFIED EXPLANATION

A PHC wants to replace its mercury sphygmomanometer with a digital device. The question is not "are the readings related?" (they must be: both measure the same blood pressure). The question is "by how much can a digital reading differ from the mercury reading for the same person?" If that range is clinically unimportant, the methods can be swapped. That judgement is clinical, not statistical: you decide beforehand how large a difference would matter.

DERIVATION

1. Let d_i = digital − mercury for subject i. If the differences are roughly Normal with mean μ_d and SD σ_d, then 95% of them lie within $\mu_d \pm 1.96\sigma_d$ (recall the Normal distribution, §5.7.2.2).
2. Estimate μ_d by $\bar{d}$ and σ_d by s_d. This gives the limits $\bar{d} \pm 1.96$ s_d.
3. $\bar{d}$ is a sample mean, so its SE is $s_d \div \sqrt{n}$ and its 95% CI is $\bar{d} \pm t_{0.975,n-1} \times s_d \div \sqrt{n}$ (recall §7.2.1). If the CI excludes 0, the bias is systematic, which is the same as a paired t-test (recall §9.2.3).

4. **Why r is the wrong statistic** (recall §11.2): r measures how closely points lie along *any* straight line. If the digital reading were always mercury + 10 mmHg, r would be 1 while every reading was 10 mmHg wrong. r also rises with the spread of the true values in the sample, which has nothing to do with agreement.

ILLUSTRATION / EXAMPLE

Thirty adults at a PHC have systolic BP measured with both a mercury sphygmomanometer and a digital device (`data/bp_methods.csv`). Mean SBP is 129.0 mmHg by mercury and 133.5 mmHg by digital.

WORKED CALCULATION

1. Differences (digital − mercury): $\bar{d}$ = **+4.53 mmHg**, s_d = **4.95 mmHg**.
2. Limits of agreement: 4.53 − 1.96 × 4.95 = **−5.2 mmHg** and 4.53 + 1.96 × 4.95 = **14.2 mmHg**.
3. SE($\bar{d}$) = 4.95 ÷ √30 = 0.903. With $t_{0.975,29}$ = 2.045, the 95% CI for the bias is 4.53 ± 2.045 × 0.903 = **2.69 to 6.38 mmHg**. It excludes 0, so the digital device reads systematically higher.
4. Pearson's r between the methods is **0.95**, although individual readings differ by up to 17 mmHg.
5. Reading: for a given adult, the digital device may read from about 5 mmHg lower to 14 mmHg higher than mercury. A 14 mmHg over-reading can move a person across the systolic threshold of 140 mmHg at which WHO diagnoses hypertension (on two different days; WHO, 2025), so the two methods are not interchangeable for diagnosis without correction.

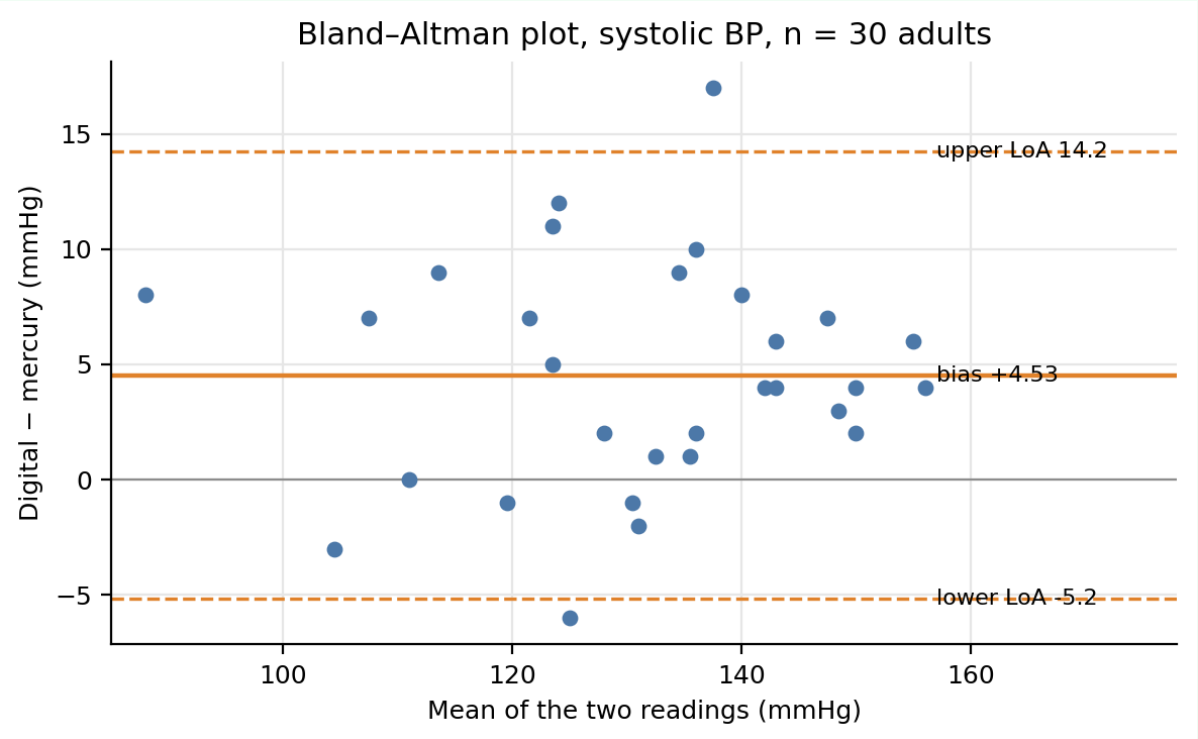


Figure 16.1. Bland–Altman plot for 30 adults. Each point is one adult: the difference between the digital and mercury readings against their mean. The solid line is the bias (+4.53 mmHg), the dashed lines the 95% limits of agreement (−5.2 and 14.2 mmHg). Two of the 30 points lie outside the limits, about the 5% expected.

IN R (OPTIONAL)

```
library(psych)
bp ← read.csv("data/bp_methods.csv")
d ← bp$sbp_digital - bp$sbp_mercury
m ← (bp$sbp_digital + bp$sbp_mercury) / 2
c(bias = mean(d), sd = sd(d))
mean(d) + c(-1.96, 1.96) * sd(d)
t.test(d)$conf.int
cor(bp$sbp_mercury, bp$sbp_digital)
cor(d, m)
ICC(bp[, c("sbp_mercury", "sbp_digital")])
plot(m, d)
abline(h = mean(d) + c(-1.96, 0, 1.96) * sd(d), lty = c(2, 1, 2))
```

```
    bias       sd
4.533333 4.946147
[1] -5.161114 14.227781
[1] 2.686412 6.380255
[1] 0.9530262
[1] 0.0519762
...
                         type  ICC  F df1 df2       p lower bound upper bound
Single_raters_absolute   ICC1 0.92 23  29  30 1.0e-13        0.83        0.96
Single_random_raters     ICC2 0.92 41  29  29 7.2e-17        0.58        0.97
Single_fixed_raters      ICC3 0.95 41  29  29 7.2e-17        0.90        0.98
...
```

The first four lines are steps 1–4. `cor(d, m)` = 0.05 shows that the differences do not grow with the level of BP, so a single pair of limits fits the whole range. (If they did grow, you would analyse log-transformed readings or ratios.) The ICCs repeat the lesson of §16.3.2: the consistency ICC(3,1), 0.95, ignores the 4.5 mmHg bias, while the absolute-agreement ICC(2,1), 0.92, is lower and has a much wider lower confidence bound (0.58). No ICC tells you, in mmHg, how far apart two readings can be. Only the limits of agreement do.

MUST-KNOW NOTES (MD-LEVEL)

- The limits of agreement are estimates too. $\mathrm{Var}(\bar{d}) = s_d^2/n$, and the SD itself has sampling variance of about $s_d^2/2n$ (proof beyond this book), so each limit has SE $\approx \sqrt{(s_d^2/n + 1.96^2 \times s_d^2/2n)} \approx \sqrt{(3s_d^2/n)}$. Here that is $\sqrt{(3 \times 4.95^2 \div 30)} = 1.56$ mmHg: the upper limit's 95% CI is about 14.2 ± 2.045 × 1.56, or 11.0 to 17.4 mmHg. Thirty subjects pin the limits down only roughly.
- Always look at the plot. A funnel shape (differences widening with the mean) or a slope means one pair of limits misdescribes the agreement.
- A regression of one method on the other and a paired t-test each answer only part of the question: the paired t-test detects bias but says nothing about the spread of individual differences.
- A method can be compared with itself: two readings by the same method give the repeatability coefficient of §16.3.2.

CHECKPOINT 16.4

1. A new haemoglobinometer and the laboratory cyanmethaemoglobin method are compared in 40 pregnant women. Mean difference (new − lab) is −0.3 g/dL, SD of differences 0.6 g/dL. Compute the 95% limits of agreement. The study defines anaemia as Hb below 11 g/dL: interpret the limits for a woman whose laboratory value is 11.2 g/dL.
2. The same study reports r = 0.96 between the two methods and concludes "excellent agreement". What is wrong with this conclusion?

(Answers in the Appendix at the end of the book.)

16.4 Relationship Between Validity and Reliability

DEFINITION

Reliability is necessary but not sufficient for validity, in the psychometric sense this chapter uses (validity of individual scores). A reliable instrument can be consistently wrong. An unreliable instrument cannot give individual scores that are valid, because random error dominates each one. The **reliability ceiling** makes this exact:

$r_{xy} \le \sqrt{(r_{xx} \times r_{yy})}$

where r_{xy} is the observed correlation between instrument X and criterion Y, and r_{xx} and r_{yy} are their reliabilities.

SIMPLIFIED EXPLANATION

Picture three shooters aiming at a bullseye, the true value of the construct (Figure 16.2). Shooter 1's arrows cluster tightly but far from the bullseye: consistently wrong, reliable but not valid, like the scale that reads 2 kg high. Shooter 2's arrows scatter widely, but their average lies near the bullseye. Shooter 3's arrows cluster tightly on the bullseye: reliable and valid.

Shooter 2 is where two traditions of vocabulary part. In epidemiology, **validity** usually means the absence of *systematic* error and **precision** (or reliability) the absence of *random* error. The two are treated as separate properties, "threats to validity" and "assessing precision" (Rothman, Greenland & Lash, 2008). On that reading Shooter 2 is "valid but imprecise": accurate on average, unbiased, useful for estimating a *group* mean. In psychometrics, validity is a property of each person's score. Shooter 2's individual arrows land far from the bullseye, so no single score can be trusted, and the scores are not valid. This chapter uses the psychometric sense. When you read "valid" in a paper, check which sense is meant.

DERIVATION

1. $X = T_X + E_X$ and $Y = T_Y + E_Y$, with errors uncorrelated with each other and with the true scores (§16.3).
2. Then $\mathrm{Cov}(X, Y) = \mathrm{Cov}(T_X, T_Y)$, because every covariance involving an error term is 0.
3. $r_{xy} = \mathrm{Cov}(T_X, T_Y) \div (\sigma_X \sigma_Y) = \rho_T \times \sigma_{TX} \sigma_{TY} \div (\sigma_X \sigma_Y)$, where ρ_T is the correlation between the true scores.
4. $\sigma_{TX} \div \sigma_X = \sqrt{(\sigma^2_{TX} \div \sigma^2_X)} = \sqrt{r_{xx}}$, and likewise for Y. So $r_{xy} = \rho_T \times \sqrt{(r_{xx}\, r_{yy})}$.

5. Since $|\rho_T| \le 1$, $|r_{xy}| \le \sqrt{(r_{xx}\, r_{yy})}$. Rearranged, $\rho_T = r_{xy} \div \sqrt{(r_{xx}\, r_{yy})}$: this is Spearman's **correction for attenuation** (Spearman, 1904), the correlation that would be seen if both measures were perfectly reliable.

ILLUSTRATION / EXAMPLE

A new stigma scale has test-retest reliability 0.70. The criterion, a structured interview, has reliability 0.80. The highest correlation the scale can show with the interview, even if both measure the same construct perfectly, is $\sqrt{(0.70 \times 0.80)} = \sqrt{0.56} = \mathbf{0.75}$. An observed $r = 0.45$ corrects to $0.45 \div 0.748 = \mathbf{0.60}$. Part of the apparently weak validity is measurement noise.

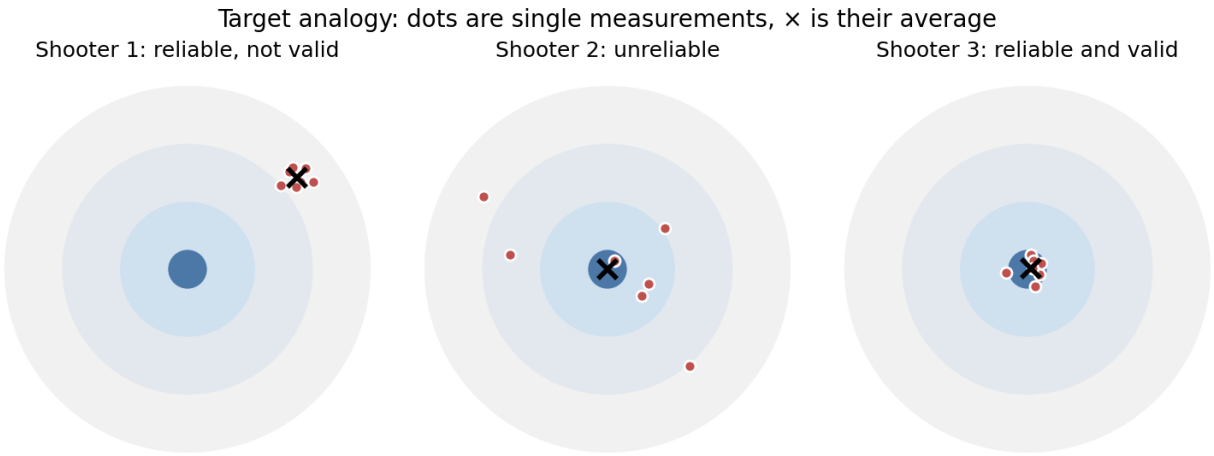


Figure 16.2. Three shooters. Red dots are single measurements, and the black × marks their average. Shooter 1 is reliable but not valid. Shooter 2 is unreliable; the average lies near the bullseye but single shots do not. Shooter 3 is reliable and valid.

MUST-KNOW NOTES (MD-LEVEL)

- A reliable instrument can still be invalid. An unreliable one cannot give valid individual scores, and its validity coefficients are capped at $\sqrt{(r_{xx}\, r_{yy})}$.
- Disattenuated correlations are estimates of what *could* be seen with perfect measurement. Report them beside, not instead of, the observed r.
- In practice, check reliability statistically before interpreting criterion or construct validity (§16.7). Content and face validity come earlier still, at the design stage.

16.5 Principal Component Analysis

DEFINITION

Principal component analysis (PCA) replaces a set of p correlated variables with p new variables, the **principal components**. Each component is a weighted sum of the standardised originals. The components are uncorrelated with each other, and they are ordered so that the first captures as

much of the total variance as possible, the second as much of what remains, and so on. Usually only the first few are kept.

- An **eigenvalue** is the variance of one component. With p standardised items, total variance = p, so the proportion of variance explained by a component is its eigenvalue ÷ p.
- A **loading** is the correlation between an original item and a component. An item "loads on" a component when this correlation is large. For standardised items, loading = eigenvector weight × √eigenvalue, and the squared loadings on a component add up to its eigenvalue.

SIMPLIFIED EXPLANATION

If 20 items are strongly correlated because they tap a few underlying ideas, much of the information is redundant. PCA finds the directions in which the items vary together most and compresses them into a few components. Each component is a weighted combination of the items, so it can be scored for every respondent. (Components are called **orthogonal** because they are uncorrelated.)

DERIVATION

1. With standardised items, $\mathrm{Var}(a_1z_1 + \ldots + a_pz_p) = \Sigma_i\Sigma_j\, a_i a_j r_{ij}$, the variance of a sum (recall §16.3.1.1, Derivation step 1) with correlations in place of covariances. In matrix shorthand this is written $a^{T}Ra$, where R is the p × p table (matrix) of correlations and a the list (vector) of weights.
2. Maximise $a^{T}Ra$ subject to $a^{T}a = a_1^{\,2} + \ldots + a_p^{\,2} = 1$, to stop the weights growing without limit. Setting the derivative of $a^{T}Ra - \lambda(a^{T}a - 1)$ to zero (the Lagrange-multiplier method; the calculus is beyond this book) gives $Ra = \lambda a$.
3. So the weights are an **eigenvector** of R and the component's variance is $a^{T}Ra = \lambda$, the matching **eigenvalue**. The largest eigenvalue gives the first component. The eigenvalues of R sum to its diagonal total (the trace), which is p ones, so p.

ILLUSTRATION / EXAMPLE

Household asset indices, such as the DHS wealth index, are the classic use in public health. Filmer & Pritchett (2001), using Indian data, constructed "a linear index from asset ownership indicators, using principal-components analysis to derive weights." The illustration below uses a correlation matrix of five binary asset items (electricity, television, mobile phone, motorcycle, bicycle) in 500 households. It is a teaching matrix, not survey data.

WORKED CALCULATION

1. **Eigenvalues**: 2.24, 1.06, 0.69, 0.58, 0.43 (sum 5). The first component explains 2.24 ÷ 5 = **45%** of the variance. The first two explain 66%.
2. **Loadings on component 1**: electricity 0.77, television 0.82, mobile 0.74, motorcycle 0.66, bicycle −0.04. Squared and summed: 0.585 + 0.674 + 0.542 + 0.432 + 0.001 = 2.24, the first eigenvalue. The four "modern" assets load together. The bicycle does not: in many villages poorer households own bicycles too.
3. **How many components?** The **Kaiser criterion** keeps components with eigenvalue above 1, a common convention (Vyas & Kumaranayake, 2006). Here it keeps two. The second component

(eigenvalue 1.06) is essentially the bicycle alone: its loading is 0.96, and no other item loads above 0.33 in absolute value. A lone item forming its own "component" is noise, not a construct. The **scree plot** (Figure 16.3) shows a sharp bend after component 1. Cattell's scree test reads the number of components off that bend. Parallel analysis (Horn, 1965) keeps only components whose eigenvalue beats the matching eigenvalue from random data of the same size. Here parallel analysis returns one component: the random-data second eigenvalue is about 1.05–1.06, the same as the observed 1.06. Both methods point to one component.

4. **Is the matrix worth analysing?** The **Kaiser–Meyer–Olkin (KMO)** measure of sampling adequacy compares the correlations with the partial correlations. Values near 1 mean the items share common variance. Kaiser's own labels, quoted in the documentation of R's psych package, run from "unacceptable" below 0.5 through "miserable" (0.5s), "mediocre" (0.6s), "middling" (0.7s) and "meritorious" (0.8s) to "marvelous" above 0.9. Here overall KMO = **0.72** ("middling"). For the bicycle alone it is 0.37, a sign to drop it. **Bartlett's test of sphericity** tests H_0: the correlation matrix is an identity matrix (no correlations at all). $\chi^2 = 439.6$ on 10 df, $p < 0.001$, so there is correlation to summarise. With large samples it is usually significant, so it is a low bar.

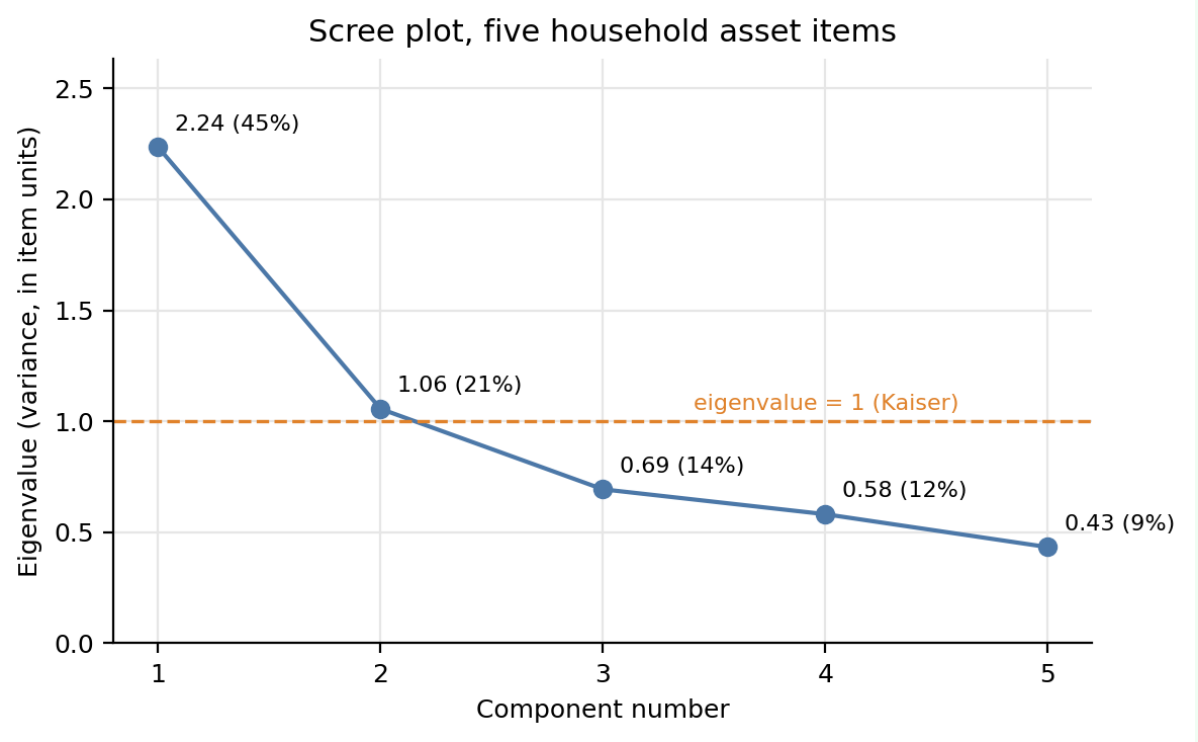


Figure 16.3. Scree plot for the five asset items. The first eigenvalue (2.24, 45% of variance) stands well above the rest. The second (1.06) only just crosses the Kaiser line at 1, and the bend after component 1 suggests one component.

IN R (OPTIONAL)

```
library(psych)
items ← c("electricity", "television", "mobile", "motorcycle",
          "bicycle")
R ← matrix(c(1.00, 0.55, 0.40,  0.30,  0.05,
             0.55, 1.00, 0.45,  0.40,  0.00,
             0.40, 0.45, 1.00,  0.35,  0.10,
             0.30, 0.40, 0.35,  1.00, -0.10,
             0.05, 0.00, 0.10, -0.10,  1.00),
           nrow = 5, dimnames = list(items, items))
round(eigen(R)$values, 2)
KMO(R)
cortest.bartlett(R, n = 500)
principal(R, nfactors = 1, n.obs = 500)$loadings
```

```
[1] 2.24 1.06 0.69 0.58 0.43
...
Overall MSA =  0.72
MSA for each item =
electricity  television      mobile  motorcycle     bicycle
       0.72        0.69        0.77        0.76        0.37
$chisq
[1] 439.6392
$p.value
[1] 3.385033e-88
$df
[1] 10

Loadings:
            PC1
electricity 0.765
television  0.821
mobile      0.736
motorcycle  0.657
bicycle

                 PC1
SS loadings    2.236
Proportion Var 0.447
```

`eigen` gives step 1. `KMO` and `cortest.bartlett` give step 4. `principal` prints the loadings of step 2. The bicycle's loading (−0.04) is left blank because R hides loadings below 0.1. "SS loadings" is the sum of squared loadings (the eigenvalue, 2.236), and "Proportion Var" is 2.236 ÷ 5 = 0.447.

MUST-KNOW NOTES (MD-LEVEL)

- **Binary items.** PCA was designed for continuous variables, yet asset indices routinely run it on binary items (Filmer & Pritchett, 2001). Categorical variables are recoded into binary indicators (Vyas & Kumaranayake, 2006). Critics note that the approach "is statistically inappropriate for use with discrete data" (Howe, Hargreaves & Huttly, 2008). Alternatives are PCA on tetrachoric or polychoric correlations or multiple correspondence analysis.
- **Rotation.** When two or more components are kept, their loadings can be rotated to make them easier to interpret: each item loads highly on one component and near zero on the others. **Varimax** keeps the components uncorrelated (orthogonal rotation). **Promax** and oblimin let them correlate (oblique rotation), which suits constructs that are expected to be related, such as

sub-domains of quality of life. Rotation redistributes the variance explained among the kept components. It does not change the total.

- **PCA versus factor analysis.** PCA summarises *total* variance and makes no model of where it comes from. **Exploratory factor analysis** (EFA) models only the *shared* (common) variance as caused by latent factors and treats the rest as item-specific error. For scale development, EFA matches the idea of a construct causing its indicators more closely. In practice the two often give similar loadings.
- PCA checks dimensionality before α is reported (§16.3.1.1). If items split into several components, report α for each sub-scale.
- Software names saved component or factor scores automatically. SPSS, for example, names them FAC1_1, FAC2_1 and so on for the first set of saved scores (IBM SPSS FACTOR documentation).

16.6 Building Composite Variables

DEFINITION

A **composite variable** is a single score built by combining several items or indicators, intended to represent a construct more completely than any single item.

SIMPLIFIED EXPLANATION

Once internal consistency (§16.3.1.1) and dimensionality (§16.5) have been checked, item responses are combined into one score. That score is the variable used in the substantive analysis.

ILLUSTRATION / EXAMPLE

A 10-item satisfaction scale with acceptable α and one dominant component becomes a satisfaction score for each patient: the sum or mean of the items. It then becomes the outcome when mean satisfaction is compared between two PHCs.

MUST-KNOW NOTES (MD-LEVEL)

- The simplest composite is an unweighted sum or mean, giving every item equal weight.
- A weighted composite uses the component or factor weights, so items with higher loadings count more. An asset-based wealth index is built this way, with weights from PCA (Filmer & Pritchett, 2001; §16.5).
- The proper order is: content validity at design; collect data; check internal consistency and dimensionality; *then* build the composite; *then* assess criterion and construct validity.
- A sum of ordinal items is often treated as approximately continuous in later analysis. This is a working approximation, not a promotion to an interval scale (recall §1.1.2). Whether it is reasonable depends on the measurement model, the items, the dimensionality and the validation evidence, not on the arithmetic of adding.

16.7 Questionnaire Validation Methodology

DEFINITION

Questionnaire validation methodology is the staged process of developing or adapting an instrument and gathering evidence that its scores are valid and reliable in the target population.

SIMPLIFIED EXPLANATION

This section places the chapter's tools in the order a validation study uses them. The COSMIN checklist (Mokkink et al., 2010) lists the measurement properties such a study should address: "internal consistency, reliability, measurement error, content validity (including face validity), construct validity (including structural validity, hypotheses testing and cross-cultural validity), criterion validity, responsiveness". Use it to plan a study and to appraise published ones.

ILLUSTRATION / EXAMPLE

A team develops a Hindi "CBHI trust" scale for rural Chhattisgarh.

1. **Define the construct** and its domains from the literature and qualitative work.
2. **Draft items; content validity** by an expert panel (§16.2.1).
3. **Translation and cultural adaptation** when adapting an existing instrument. Beaton et al. (2000) set out five stages: at least two independent forward translations; a synthesis of the translations; back-translation into the source language by a translator "totally blind to the original version"; review by an expert committee; and a pretest of the prefinal version in the target population.
4. **Face validity and cognitive pretest** in a small group of target respondents (§16.2.2).
5. **Field the draft** to a validation sample.
6. **Structure and internal consistency.** Use PCA or EFA to examine dimensionality (§16.5), then α per sub-scale with corrected item-total correlations and alpha-if-deleted (§16.3.1.1). Revise or remove items. **Confirmatory factor analysis** (CFA) then tests a pre-specified structure, ideally in a new sample. It is fitted as a structural equation model and is beyond this book.
7. **Test-retest reliability** in a subsample (ICC(2,1) for the score, §16.3.2; weighted kappa for single ordinal items, §16.3.3.1), plus measurement error such as the repeatability coefficient.
8. **Build the composite score** (§16.6).
9. **Construct validity** by predefined hypotheses: convergent and discriminant correlations (§16.2.4.1) and **known-groups** comparisons, such as CBHI members against non-members (§16.2.4). **Criterion validity** if a gold standard exists (§16.2.3).
10. **Responsiveness**, if the scale will measure change: "the ability of a questionnaire to detect clinically important changes over time, even if these changes are small" (Terwee et al., 2007). It is tested by giving the scale before and after an intervention expected to change trust.

MUST-KNOW NOTES (MD-LEVEL)

- **Order.** Statistical reliability checks (steps 6–7) should precede the interpretation of criterion and construct validity (step 9), because unreliability caps validity (§16.4). Design-stage evidence (content and face validity, steps 2–4) comes before any data.
- **Sample size.** Terwee et al. (2007) consider "a sample size of at least 50 patients adequate" for assessing reliability and validity. For factor analysis, "rules-of-thumb vary from four to 10

subjects per variable, with a minimum number of 100 subjects". These are conventions. A 20-item scale would need 80–200 respondents for EFA, and at least 100.

- **Iteration.** Poor performance at any step (weak item-total correlations, low α, failed hypotheses) usually sends you back to revise specific items, not to abandon the instrument.
- **Validity is population-specific.** Evidence applies to the population, language and setting in which it was gathered. An instrument validated elsewhere (for example, a Western quality-of-life scale) needs cross-cultural adaptation and fresh reliability and validity testing before use in rural Chhattisgarh.

CHECKPOINT 16.5

1. A team runs PCA on 20 items meant to measure one construct, "quality of life". Three components have eigenvalues above 1 (4.1, 1.3 and 1.1), explaining 20.5%, 6.5% and 5.5% of the variance. What do the percentages mean? What should the team check before concluding that the scale has three sub-dimensions?
2. Explain why a highly reliable instrument is not automatically valid. A scale has reliability 0.64 and its criterion has reliability 0.81. What is the largest correlation the scale could show with the criterion?
3. A Hindi translation of an English depression scale is to be used in Raipur. List the translation steps, and name two measurement properties you would test afresh, with the statistic for each.

(Answers in the Appendix at the end of the book.)

Threats to Study Validity

Chapter 16 asked whether one *instrument* measures what it claims to. This chapter asks the same question of a whole *study*: even with a perfect instrument and a correct calculation, can the design or conduct of the study give a wrong answer? It sets out the three alternative explanations for an observed association (§17.1), the systematic errors of selection and measurement with a catalogue of the named biases and Bradford Hill's viewpoints (§17.2), confounding and what design can do about it, with quantitative bias analysis and the E-value (§17.3). It closes with the two skills that turn a valid study into a usable paper: reporting it to the standard of the reporting guidelines (§17.4), and knowing enough qualitative and mixed methods to work alongside them (§17.5).

17.1 Validity of a Study

DEFINITION

The **validity of a study** is the degree to which its design and conduct support correct conclusions. It has two parts: **internal validity** (are the conclusions correct for the people studied?) and **external validity** (do they carry over to other people, places or times?). Before an observed association can be read as an effect, three alternative explanations must be weighed: **chance** (random variation from studying a sample), **bias** (a systematic error in how people were selected or measured) and **confounding** (a mixing of the exposure's effect with the effect of a third variable).

SIMPLIFIED EXPLANATION

Recall §11.5, which listed these three and a fourth, reverse causation. Reverse causation is a question of which came first; the design settles it (a cohort measures exposure before outcome, recall §1.5), so this chapter concentrates on the other three.

The three differ in kind. Chance is random: it shrinks as the sample grows, and the confidence interval and p-value measure it (recall §7.2 and §8.5). Bias and confounding are systematic: they push the estimate one way and do not shrink with sample size (recall §6.4). A larger study removes chance and leaves bias and confounding exactly where they were, now with a narrower confidence interval around the wrong answer.

The word "validity" also has two meanings in this book. Instrument validity (recall §16.2) asks whether a tool measures the right construct. Study validity asks whether the whole study supports its conclusion; in this epidemiological sense validity is the absence of systematic error and precision the absence of random error (recall §16.4). A valid instrument used in a badly designed study still gives an invalid result.

ILLUSTRATION / EXAMPLE

A trial of a new oral rehydration regimen (hypothetical figures), run with care in a tertiary hospital, finds fewer treatment failures in the intervention arm, RR 0.70 (95% CI 0.55 to 0.89). The three questions in order: Chance? The interval excludes 1, so chance alone is an unlikely explanation. Bias? Were outcomes assessed by staff who knew the arm, and were children lost from one arm more than the other? Confounding? Randomisation balanced the arms on average. If all three are answered, the result is internally valid for hospitalised children. Whether it holds for children treated at a PHC is the separate question of external validity.

MUST-KNOW NOTES (MD-LEVEL)

- Ask the questions in order: chance, bias, confounding, then generalisability. A p-value answers only the first.
- A statistical technique can be applied with perfect arithmetic to a biased or confounded comparison. Statistical rigour and design rigour are separate requirements.

17.1.1 Internal Validity

DEFINITION

Internal validity is the degree to which the association observed among the study participants reflects the true effect of the exposure in those participants, with chance, bias and confounding ruled out as far as possible.

SIMPLIFIED EXPLANATION

Within *this* study, among *these* people, is the comparison fair and the measurement true? Each systematic threat in the rest of this chapter is a way internal validity fails: selection bias (§17.2.1), information bias (§17.2.2) and confounding (§17.3).

ILLUSTRATION / EXAMPLE

In an RCT the investigator assigns the exposure at random, which makes the arms alike, on average, in measured and unmeasured characteristics (recall §1.5; §15.9 develops the design). Confounding is then answered by design. An observational cohort or case-control study (recall §1.5) must instead measure the confounders and adjust for them (recall §11.7 and §12.4), and can do nothing about the ones it did not measure (§17.3).

MUST-KNOW NOTES (MD-LEVEL)

- Internal validity is a precondition for external validity. A result that is wrong for the people studied has nothing correct to generalise, however representative the sample.
- Randomisation protects against confounding. It does not protect against bias: an unblinded outcome assessor or unequal loss to follow-up can bias a randomised trial.

17.1.2 External Validity

DEFINITION

External validity (generalisability) is the degree to which a result that is internally valid for the study participants holds for another target population, setting or time.

SIMPLIFIED EXPLANATION

External validity is a question about whether the *effect* is the same elsewhere, not about whether the sample looks like the population. A sample that differs from the target population in age, severity or setting limits generalisability only if the effect itself differs across those characteristics (effect modification, recall §11.7). This is why a non-representative sample does not by itself make a study invalid: it is a threat to external validity, and it becomes a threat to internal validity only when selection depends on both exposure and outcome (selection bias, §17.2.1).

ILLUSTRATION / EXAMPLE

A nutrition trial run in urban, well-resourced hospital patients shows a genuine effect. Rural PHC patients differ in baseline nutrition, access to follow-up and causes of malnutrition. If the intervention works through a mechanism that depends on those features (for example, it corrects a deficiency that is commoner in rural children), the effect may be larger or smaller there. Whether it transfers is a judgement about effect modification, informed by biology and by trials in other settings.

MUST-KNOW NOTES (MD-LEVEL)

- Strict eligibility criteria and intensive monitoring strengthen internal validity and can narrow the population to which the result applies. This trade-off is a deliberate design choice.
- Probability sampling from the target population (recall §6.2) supports generalisation of descriptive estimates such as a prevalence. For an effect estimate, the stronger evidence is consistency across settings.

CHECKPOINT 17.1

1. A rigorous RCT conducted only at a single urban tertiary hospital shows that a treatment works, with excellent internal validity. A district health officer wants to apply the finding in rural PHCs. What should she consider first, and what kind of evidence would answer it?
2. Explain why internal and external validity can pull against each other in study design.
3. Study A (n = 120) reports RR 1.9 (95% CI 0.9 to 4.0). Study B (n = 12,000), using the same exposure questionnaire with the same systematic error, reports RR 1.35 (95% CI 1.25 to 1.46). Which alternative explanation has Study B's size dealt with, and which has it not?

(Answers in the Appendix at the end of the book.)

17.2 Bias in Research

DEFINITION

Bias, in the design sense (recall §6.4), is a systematic error in the design, conduct or analysis of a study that makes the estimate of the exposure–outcome association differ from the true value in a consistent direction. Two families are recognised: **selection bias**, from how people enter or stay in the analysis (§17.2.1), and **information bias**, from how exposure, outcome or other variables are measured (§17.2.2).

SIMPLIFIED EXPLANATION

Bias is not noise. A larger sample of the same flawed kind gives a more precisely wrong answer. The two families correspond to the two places a study can go wrong: who is counted, and what is recorded about them.

Two designs recur in this section (recall §1.5). A **cohort** study starts from exposure and follows people forward to count new cases. A **case-control** study starts from the outcome, selecting people with the disease (cases) and without it (controls), and looks back for past exposure. Each design has its characteristic biases: the cohort loses people to follow-up, the case-control study must choose its controls well and depends on reports of the past.

ILLUSTRATION / EXAMPLE

A case-control study of oral cancer takes cases from a cancer hospital and controls from the same hospital's general wards, then asks both groups about tobacco chewing. Two things can go wrong. If the controls were admitted for conditions that are themselves caused by tobacco, they chew more than the population that produced the cases, and the odds ratio is pulled down (selection bias). If cases, having thought hard about the cause of their cancer, report past chewing more completely than controls, the odds ratio is pushed up (information bias).

MUST-KNOW NOTES (MD-LEVEL)

- In the traditional classification, bias is an error in how the study was run and confounding is a feature of the world the study failed to account for. Many epidemiologists treat confounding as a third kind of bias ("confounding bias"): failing to measure a known confounder *is* a design flaw.
- The structural view (Hernán et al., 2004) sharpens the line: confounding comes from **common causes** of exposure and outcome; selection bias comes from **conditioning on a common effect** (a collider, recall §11.6.1). §17.2.1 uses this.
- Most bias is best prevented at the design stage. Some can be partly corrected afterwards, by weighting for loss to follow-up, by validation sub-studies or by quantitative bias analysis (recomputing the estimate under stated error rates, §17.3.1).

17.2.1 Selection Bias

DEFINITION

Selection bias is a distortion of the exposure–outcome association that arises because the probability of entering the study, or of remaining in the analysis, depends on **both** the exposure (or a cause of it) **and** the outcome (or a cause of it). The association among the people analysed then differs from the association in the population they came from.

SIMPLIFIED EXPLANATION

Keep two questions apart.

1. **Is the sample like the target population?** If not, *descriptive* estimates such as a prevalence or a mean weight are biased whenever selection is related to the outcome (recall §6.4 and Checkpoint 6.3, where a clinic sample overstated children's mean weight). For an *association*, a non-representative sample limits generalisability (§17.1.2) but does not, by itself, bias the estimate.
2. **Does selection depend on both exposure and outcome?** Only then is the association among those studied distorted. That is selection bias in the epidemiological sense. (Selection on the outcome alone, as in every case-control study, changes the risks but leaves the odds ratio intact; recall §1.5 and §10.6.4.)

In the causal-diagram language of §11.6.1, selection S is a collider: arrows run into it from both the exposure E and the outcome D (E → S ← D). Studying only the selected people conditions on S and opens the path between E and D, creating or distorting an association (Hernán et al., 2004). Recall §11.6.1 step 5: restricting to hospital deliveries opened such a path.

ILLUSTRATION / EXAMPLE

Recall the cohort of 200 infants in §10.6.4: 33 of 80 not exclusively breastfed (not EBF) and 14 of 120 EBF infants had diarrhoea, RR 3.54.

Selection on exposure only. Suppose the study had enrolled every not-EBF infant but only half of the EBF infants, chosen without regard to later diarrhoea. The expected table is 33/47 (not EBF) and 7/53 (EBF). The sample is now unrepresentative: 57% of it is not EBF (80 of 140), against 40% in the source population. Yet the risks are 33/80 = 0.412 and 7/60 = 0.117, and the RR is still **3.54**. Unrepresentative, unbiased.

Selection on exposure and outcome. Now suppose 30% of the not-EBF infants who stayed well moved away and were lost, while every infant with diarrhoea stayed (sick infants keep returning to the clinic). The not-EBF row becomes 33 cases and 47 × 0.7 = 32.9 non-cases. Risk = 33 ÷ 65.9 = 0.501, and RR = 0.501 ÷ 0.117 = **4.29**, against a true 3.54. Loss depended on exposure and outcome together, and the association is biased away from the null.

WORKED CALCULATION

Berkson's bias as collider stratification (an illustrative example with hypothetical counts, after the diabetes and gallbladder disease example in Berkson, 1946). In a population of 10,000 adults, 30% have diabetes (E) and 10% have gallbladder disease (D), and the two are independent. A hospital

admits people with probabilities 0.05 (neither condition), 0.20 (diabetes only), 0.30 (gallbladder disease only) and 0.40 (both).

1. Population cells: E and D 300; E only 2,700; D only 700; neither 6,300. Population OR = (300 × 6,300) ÷ (2,700 × 700) = **1.00**.
2. Admitted: 300 × 0.40 = 120; 2,700 × 0.20 = 540; 700 × 0.30 = 210; 6,300 × 0.05 = 315.
3. Hospital case-control study (gallbladder cases versus other in-patients as controls): OR = (120 × 315) ÷ (540 × 210) = 37,800 ÷ 113,400 = **0.33**.
4. Among admitted cases, 120 of 330 (36.4%) have diabetes; among admitted controls, 540 of 855 (63.2%) do; in the population, 30%.

Result: two independent conditions appear strongly *inversely* associated in hospital data, because admission depends on both. With other admission probabilities the spurious association can point the other way; its direction is not predictable in general.

MUST-KNOW NOTES (MD-LEVEL)

- Test for selection bias with one question: could being in the analysis depend on both the exposure and the outcome (or their causes)? If on exposure alone, risks and ratios are unbiased; if on outcome alone, the odds ratio is unbiased (the basis of the case-control design). Generalisability may still suffer.
- **Loss to follow-up bias** is selection bias at exit. If loss depends on exposure and outcome together, the analysed cohort is biased even when enrolment was perfect.
- Prevention is at the design stage: population-based cases and controls, controls from the population that produced the cases, incident rather than prevalent cases, and active tracing to minimise loss. §17.2.3 lists the named forms.

17.2.2 Information Bias

DEFINITION

Information bias is a systematic error in how exposure, outcome or covariates are measured or recorded. For a categorical variable it takes the form of **misclassification**: people placed in the wrong category. Misclassification is **non-differential** when the probability of being misclassified on one variable is the same in the groups defined by the other (for exposure: the same in people with and without the outcome), and **differential** when it differs (Catalogue of Bias, misclassification bias).

SIMPLIFIED EXPLANATION

Selection bias concerns who is counted; information bias concerns what is recorded about them. The accuracy of a classification is described by the sensitivity and specificity of §5.5.1. For exposure misclassification, sensitivity (Se) is the proportion of truly exposed people recorded as exposed, and specificity (Sp) the proportion of truly unexposed people recorded as unexposed.

Non-differential misclassification of a binary exposure mixes the two exposure groups. Some exposed people sit among the unexposed and some unexposed among the exposed, so the two groups look more

alike than they are. The expected result is a bias **towards the null**: "Non-differential (random) misclassification of measures ... will tend to lead to an underestimation of effect" (Catalogue of Bias, information bias). Differential misclassification can move the estimate in either direction.

An example of differential measurement is **interviewer bias**: an interviewer who knows a participant's outcome asks about exposure differently. In a case-control study of neonatal tetanus, an interviewer who knows which mothers lost a baby may probe harder about unclean cord care in those mothers, recording more exposure among cases than among controls.

DERIVATION

Why non-differential misclassification pulls a binary exposure towards the null. Among n people with the outcome, let a be truly exposed. The number *recorded* as exposed is the truly exposed who are detected plus the truly unexposed wrongly recorded:

$a^* = Se \times a + (1 - Sp) \times (n - a)$

The same formula, with the same Se and Sp, holds among people without the outcome. So each recorded exposure group is a blend of the true exposed and the true unexposed. When Se + Sp > 1 (the classification is better than a coin toss), the blend shrinks the difference between the recorded groups towards zero, and a ratio measure moves towards 1. Solving the equation for a gives the correction used in §17.3.1:

$a = [a^* - (1 - Sp) \times n] \div (Se + Sp - 1)$

ILLUSTRATION / EXAMPLE

Suppose the feeding status of the 200 infants of §10.6.4 had been recorded with Se = 0.80 and Sp = 0.90 for "not EBF", and suppose three versions of the error.

Scenario	Recorded not EBF: diarrhoea / none	Recorded EBF: diarrhoea / none	Recorded RR	OR
True classification	33 / 47	14 / 106	3.54	5.32
Non-differential (Se 0.80, Sp 0.90 in all)	27.8 / 48.2	19.2 / 104.8	2.36	3.15
Differential: Se 0.95 in mothers of cases, 0.80 in others	32.8 / 48.2	14.2 / 104.8	3.38	5.00
Differential: as above, and Sp 0.80 in mothers of cases	34.1 / 48.2	12.9 / 104.8	3.80	5.78

(Cell counts are expected values, which is why they have decimals.)

WORKED CALCULATION

The non-differential row.

1. Among the 47 infants with diarrhoea (33 truly not EBF, 14 EBF): recorded not EBF = 0.80 × 33 + 0.10 × 14 = 26.4 + 1.4 = **27.8**; recorded EBF = 47 − 27.8 = 19.2.
2. Among the 153 without diarrhoea (47 not EBF, 106 EBF): recorded not EBF = 0.80 × 47 + 0.10 × 106 = 37.6 + 10.6 = **48.2**; recorded EBF = 104.8.
3. Recorded risks: 27.8 ÷ 76.0 = 0.366 (not EBF) and 19.2 ÷ 124.0 = 0.155 (EBF).

4. Recorded RR = 0.366 ÷ 0.155 = **2.36**, against a true 3.54.

Result: the same error rate in both outcome groups cut the RR by a third, from 3.54 to 2.36, and hid nearly half of the excess (RR − 1 fell from 2.54 to 1.36). The two differential rows show that when mothers of infants with diarrhoea report feeding more completely, the error can leave the RR below the truth (3.38) or push it above (3.80). The direction depends on the error rates, which is why differential misclassification cannot be assumed to be conservative.

MUST-KNOW NOTES (MD-LEVEL)

- "Bias towards the null" holds on average for a *binary* variable with non-differential, independent errors and Se + Sp > 1. It can fail for exposures with three or more categories, when errors in exposure and outcome are correlated, and in any single study by chance. It is an expectation, not a guarantee.
- A null or weak result from a study with poor measurement is weak evidence of no effect: non-differential misclassification can mask a real association.
- Reduce information bias with objective records, standardised instruments (Chapter 16), neutral scripted questions, training, and blinding of interviewers and assessors to the other variable.

17.2.2.1 Recall Bias

DEFINITION

Recall bias is differential misclassification of past exposure because people with the outcome remember or report it differently, usually more completely, than people without it: "systematic error due to differences in accuracy or completeness of recall to memory of past events or experiences" (Catalogue of Bias, recall bias).

SIMPLIFIED EXPLANATION

The error comes from the participant's memory, not from the interviewer. A serious outcome prompts people to search their past for causes; people without the outcome have no such reason. The asymmetry, not any true difference in exposure, can create an association.

ILLUSTRATION / EXAMPLE

In a case-control study of a birth defect, mothers of affected children, having spent months looking for an explanation, recall and report minor pregnancy exposures (a medicine, an unusual food, a fever) more completely than mothers of healthy children. The odds ratio for each such exposure is inflated.

MUST-KNOW NOTES (MD-LEVEL)

- A concern wherever exposure is reported after the outcome: above all in case-control studies. Any case-control study relying on self-report should prompt the question.

- Reduce it by using records made before the outcome (antenatal cards, MCP cards, pharmacy records), by choosing controls with a different condition that would prompt similar reflection, and by prospective designs that record exposure first.

17.2.2.2 Observer Bias

DEFINITION

Observer bias is systematic error in measuring or classifying data that arises because the observer knows the participant's exposure, treatment or outcome, or holds expectations about the hypothesis. It is related to, but distinct from, **detection bias**: a systematic difference between groups in how often or how thoroughly the *outcome is looked for* ("systematic differences between groups in how outcomes are determined", Catalogue of Bias, detection bias).

SIMPLIFIED EXPLANATION

In recall bias the participant's memory is distorted; in observer bias the observer's judgement is. Detection bias is a third mechanism: the exposure brings people under medical attention, so their outcome is found more often, whatever its true frequency. Horwitz and Feinstein (1978) showed this for oestrogen and endometrial cancer: oestrogen caused uterine bleeding, bleeding led to investigation, and investigation found cancers that would otherwise have gone undetected. With conventional case and control selection the odds ratio was 11.98; with controls who had also been investigated for bleeding it was 1.7 (Horwitz & Feinstein, 1978). "Ascertainment bias" is used loosely in the literature for several of these, and for selection problems too, so define the mechanism rather than rely on the label.

ILLUSTRATION / EXAMPLE

A researcher measuring wound healing who knows which patients received a new dressing may rate borderline wounds as "healed" more often in the new-dressing arm. No dishonesty is needed; expectation is enough.

MUST-KNOW NOTES (MD-LEVEL)

- The countermeasure for observer bias is blinding of the outcome assessor. In a systematic review of 21 trials with both blinded and non-blinded assessment of the same binary outcome, non-blinded assessors exaggerated the odds ratio by 36% on average (ratio of odds ratios 0.64, 95% CI 0.43 to 0.96), an effect driven by misclassifying a median of 3% of patients (Hróbjartsson et al., 2012).
- The countermeasure for detection bias is equal surveillance: the same schedule of examinations and tests in exposed and unexposed groups.

CHECKPOINT 17.2

1. A cohort has 100 exposed people (30 develop the outcome) and 100 unexposed (10 develop it), so the true RR is 3.0. Exposure is recorded with Se = 0.90 and Sp = 0.80, equally in people with and without the outcome. Compute the recorded 2 × 2 table and the recorded RR. What kind of misclassification is this, and in which direction did it move the estimate?

2. In a case-control study of low birthweight, (a) interviewers who know which mothers had a low-birthweight baby ask them more follow-up questions about tobacco chewing, and (b) mothers of low-birthweight babies, looking for a reason, report chewing more completely. Name each bias and give one way to prevent each.
3. Women taking an oral contraceptive visit a doctor more often than other women, and more hypertension is recorded among them. Is this observer bias or detection bias? How would you design the study to prevent it?

(Answers in the Appendix at the end of the book.)

17.2.3 A Catalogue of Named Biases

DEFINITION

The **named biases** are recurring, recognisable patterns of selection or information bias that arise in particular designs. Each has a mechanism, a usual direction and a design fix. The table below lists those most often met in community medicine research and examinations. Sackett (1979) catalogued the biases of analytic research; the Catalogue of Bias of the Centre for Evidence-Based Medicine, Oxford, gives current definitions.

SIMPLIFIED EXPLANATION

Memorise the mechanism, not the name. Every row of the table is one of two things: selection that depends on both exposure and outcome (§17.2.1), or a measurement that errs differently in different groups (§17.2.2). If you can say which, you can predict the direction and the fix.

ILLUSTRATION / EXAMPLE

Bias	Family	Mechanism	Usual direction	Fix
Berkson's (admission rate) bias	Selection	Hospital admission depends on both exposure and disease, so hospital cases and controls are a collider-selected group: "a distortion … which arises because the subjects studied had been admitted to hospital" (Catalogue of Bias)	Either way (worked example in §17.2.1: OR 0.33 from a true 1.00)	Population controls; incident cases
Neyman (prevalence–incidence) bias	Selection	Prevalent cases exclude those who died quickly or recovered quickly; if exposure affects survival or duration, exposure among surviving cases is not exposure among all cases	Either way; misses exposures that cause rapidly fatal disease	"Using incident cases rather than prevalent cases" (Catalogue of Bias)

Bias	Family	Mechanism	Usual direction	Fix
Healthy-worker effect	Selection (some authors: confounding)	Employed people are healthier than the general population, because the sick are not hired or leave; "Workers usually exhibit lower overall death rates than the general population" (Last, quoted in Shah, 2009)	Towards the null for a harmful exposure when workers are compared with the general population (in Doll's gas workers, the unexposed had a standardised mortality ratio, recall §7.4.3, of 84)	Internal comparison group (exposed vs unexposed workers); exclude the first years of follow-up (Shah, 2009)
Volunteer (self-selection) bias	Selection	Volunteers differ from others; "Volunteers tend to be more educated, come from high social class" and adhere more (Catalogue of Bias)	Volunteers are often healthier and more adherent; biases an association only if volunteering depends on exposure and outcome	Random sampling; compare volunteers with non-volunteers on known variables
Non-response bias	Selection	"Systematic differences between responders and non-responders" (Catalogue of Bias)	Unpredictable	Repeat visits, short instruments, compare early and late responders; report the response rate
Loss to follow-up (attrition) bias	Selection	Leaving the cohort or trial depends on exposure and outcome (worked example in §17.2.1)	Either way (RR 4.29 from a true 3.54 in §17.2.1)	Tracing; intention-to-treat analysis (§15.9); multiple imputation and worst-case sensitivity analysis (§15.4.3, §17.3.1)
Lead-time bias	Information (timing of diagnosis)	Screening brings diagnosis forward, so survival measured from diagnosis lengthens even if death is not delayed: "overestimating the apparent time surviving with a disease caused by bringing forward the time of its diagnosis" (Catalogue of Bias)	Favours screening	Compare mortality in randomised screened and unscreened populations, not survival from diagnosis
Length-time bias	Selection (by screening)	Periodic screening preferentially finds slow-growing disease, which has a better prognosis anyway (Catalogue of Bias)	Favours screening	Intention-to-screen analysis of all randomised people, including cancers that surface between screens
Overdiagnosis	Selection (by screening)	Screening finds "cancer" that "would otherwise not go on to cause symptoms or death" (Welch & Black, 2010), inflating incidence and survival	Favours screening	Randomised trials with mortality end-points and long follow-up

Bias	Family	Mechanism	Usual direction	Fix
Interviewer bias	Information	The interviewer asks or records differently according to the participant's outcome or exposure (§17.2.2)	Either way (differential)	Scripted neutral questions, training, interviewer blinded to case status
Recall bias	Information	Cases remember or report past exposure differently from controls (§17.2.2.1)	Either way (differential)	Records made before the outcome; similar-condition controls
Observer / detection bias	Information	Assessor's knowledge of exposure or unequal search for the outcome (§17.2.2.2)	Non-blinded assessment exaggerated trial effects on average (Hróbjartsson et al., 2012); detection bias inflated the oestrogen OR (Horwitz & Feinstein, 1978)	Blinded assessment; equal surveillance
Hawthorne effect	Information (reactivity)	People change behaviour "in response to their awareness of being observed" (Catalogue of Bias)	Can change the measured behaviour of any observed group; distorts an uncontrolled before–after study	A concurrent control group observed in the same way; unobtrusive measures
Immortal-time bias	Time-related (misclassified person-time)	Follow-up during which, because of how exposure is defined, the outcome could not occur is credited to the exposed (recall §14.6.3; Suissa, 2008)	Favours the exposure (spurious protection)	Enter exposure that starts after time zero as a time-varying covariate (§14.6.3)

Evidence on the Hawthorne effect is mixed: a systematic review found that most studies reported some effect but that "little can be securely known about the conditions under which they operate, their mechanisms of effects, or their magnitudes" (McCambridge et al., 2014).

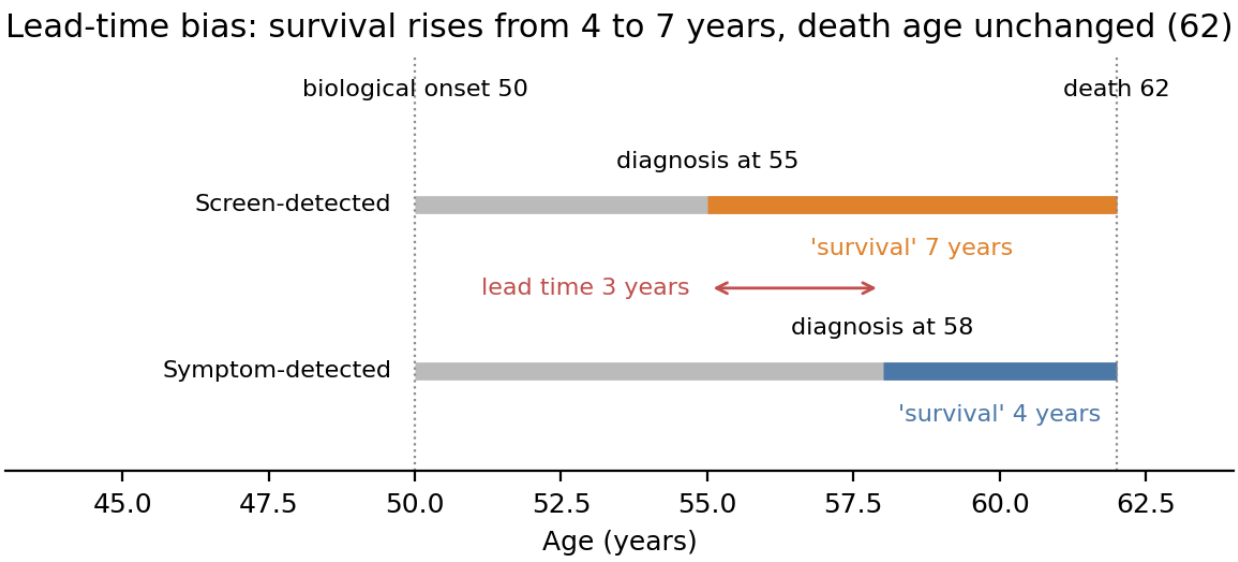


Figure 17.1. Lead-time bias. Screening moves the diagnosis three years earlier; the patient dies at 62 either way, yet "survival" rises from 4 to 7 years.

WORKED CALCULATION

Lead-time bias with the numbers of Figure 17.1.

1. Screen-detected: survival = 62 − 55 = **7 years**. Symptom-detected: survival = 62 − 58 = **4 years**.
2. Lead time = 58 − 55 = **3 years**, exactly the difference in survival.
3. Age at death is 62 in both. Screening has changed the start of the clock, not the end.

Result: comparing survival from diagnosis rewards screening even when it achieves nothing. Comparing death rates in populations randomised to screening or not removes the problem, because the clock no longer starts at diagnosis.

Bradford Hill's viewpoints on causation

When chance, bias and confounding have been considered and an association remains, how strong is the case for cause? Recall §11.5: Hill (1965) listed nine viewpoints "from all of which we should study association before we cry causation". He called them viewpoints, not criteria, and wrote that "None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a *sine qua non*" (Hill, 1965).

Viewpoint	Hill's words (Hill, 1965)	What it adds	Its limit
Strength	"First upon my list, I would put the strength of the association."	A large RR is harder to explain away by bias or unmeasured confounding (§17.3.1 quantifies this with the E-value, the confounder strength needed to explain an association away)	Weak associations can be causal
Consistency	"the consistency of the observed association"	Repeated in different places, people and designs	Studies can share the same bias

Viewpoint	Hill's words (Hill, 1965)	What it adds	Its limit
Specificity	"the specificity of the association"	One exposure, one disease	"We must not, however, over-emphasise the importance of the characteristic": tobacco causes many diseases
Temporality	"which is the cart and which the horse?"	The cause must come first	The only viewpoint that is logically necessary
Biological gradient	"biological gradient, or dose-response curve"	Risk rises with dose (recall the trend test, §10.6.5)	Thresholds exist; a confounder can also have a gradient
Plausibility	"It will be helpful if the causation we suspect is biologically plausible."	A mechanism	"What is biologically plausible depends upon the biological knowledge of the day."
Coherence	"should not seriously conflict with the generally known facts of the natural history and biology of the disease"	Fits what is known	Knowledge changes
Experiment	"it is possible to appeal to experimental, or semi-experimental, evidence"	Removing the exposure reduces the disease	Often impossible or unethical
Analogy	"it would be fair to judge by analogy"	A similar exposure is known to cause a similar effect	Analogies are easy to find, so it adds little on its own

MUST-KNOW NOTES (MD-LEVEL)

- For any named bias, say whether it is selection (depends on both exposure and outcome) or information (differential or non-differential), then predict the direction.
- Screening evaluations face three biases that all favour screening (lead time, length time, overdiagnosis). Only a randomised comparison of mortality avoids all three.
- Hill's viewpoints are aids to judgement, not a checklist to score. Only temporality is indispensable.

CHECKPOINT 17.3

1. Name the bias in each: (a) a study of coal miners finds their all-cause mortality lower than the state's; (b) a case-control study of tuberculosis and diabetes recruits cases from patients currently on treatment at a DOTS centre; (c) a free cancer screening camp finds that participants have lower cancer mortality than the district; (d) women in a diet study eat more vegetables in the weeks they know they are being observed.
2. An oral cancer screening programme detects a cancer at age 52 that would have caused symptoms at 55; the patient dies at 58. Compute survival from diagnosis with and without screening and the lead time. What outcome measure should the programme's evaluation use?
3. Which of Hill's viewpoints is logically necessary for any causal claim, and why would a lack of specificity be a weak argument against tobacco as a cause of lung cancer?

(Answers in the Appendix at the end of the book.)

17.3 Confounding

DEFINITION

Confounding is a distortion of the exposure–outcome association by a third variable that is a common cause of both. Recall §11.6: a variable is a confounder if it (1) is associated with the exposure in the source population, (2) is a risk factor for the outcome (or a proxy for one) independently of the exposure, that is, among the unexposed, and (3) is not an effect of the exposure (not on the causal path). §11.6.1 refines this with the backdoor criterion.

SIMPLIFIED EXPLANATION

§11.6 and §11.7 built confounding and its removal by stratification. This section places it among the threats to validity and answers the design question left open there: what can a study do about confounding *before* any data are analysed?

Design-stage control. Three tools act before analysis:

1. **Randomisation**: the investigator assigns exposure by chance, so measured and unmeasured confounders are balanced on average (recall §1.5; §15.9). Only available when the exposure can be assigned.
2. **Restriction**: admit only people in one category of the confounder (for example, only rural PHCs, or only non-smokers). The confounder cannot vary, so it cannot confound. The cost is fewer eligible people and a narrower population for generalisation.
3. **Matching**: choose comparison subjects with the same confounder values as each index subject (individually or by frequency). In a case-control study matching does not by itself remove confounding; the analysis must respect it (Rothman, Greenland & Lash, 2008; recall §10.8; §13.2.9 gives conditional logistic regression, a regression method for matched sets).

Analysis-stage control of *measured* confounders: stratification with Mantel–Haenszel (recall §11.7), standardisation (recall §7.4.2), multiple regression (recall §12.4; Chapters 13–14) and propensity scores (§15.3).

ILLUSTRATION / EXAMPLE

Recall §11.7: in the infant cohort, PHC confounded the association between not being EBF and diarrhoea (crude RR 3.54; MH RR 2.93). Three design alternatives would have handled it before analysis. Restriction to the rural PHC gives the stratum RR, 2.88. Frequency matching would have enrolled EBF and not-EBF infants in the same PHC proportions. Randomisation is not possible here: feeding cannot ethically be assigned. What none of these can do is remove a confounder nobody thought of, such as household wealth if it was never recorded. §17.3.1 asks how strong such a confounder would have to be.

MUST-KNOW NOTES (MD-LEVEL)

- **Unmeasured confounding** comes from a confounder never recorded; no analysis of the data at hand removes it. **Residual confounding** is what remains *after* adjusting for a measured

confounder, because it was measured with error or grouped too coarsely (recall §11.6; Fewell et al., 2007). Better measurement shrinks residual confounding; only new data, or a sensitivity analysis (§17.3.1), address unmeasured confounding.

- Confounding can push an estimate in either direction, including towards the null or across it (Simpson's paradox, recall §11.7). A null result can hide a real effect.
- A randomised trial can still be confounded by chance imbalance in a small trial, and can still be biased (§17.1.1).

17.3.1 Sensitivity and Quantitative Bias Analysis; the E-value

DEFINITION

A **sensitivity analysis** repeats an analysis under different plausible assumptions to see whether the conclusion changes. **Quantitative bias analysis** is the form that models a specific systematic error: it states the bias parameters (for example, the sensitivity and specificity of exposure measurement, or the strength of an unmeasured confounder) and computes the estimate those parameters imply. It "provides a quantitative estimate of the direction, magnitude and uncertainty arising from systematic errors" (Lash et al., 2014). The **E-value** is "the minimum strength of association, on the risk ratio scale, that an unmeasured confounder would need to have with both the treatment and the outcome to fully explain away a specific treatment-outcome association, conditional on the measured covariates" (VanderWeele & Ding, 2017).

SIMPLIFIED EXPLANATION

A confidence interval describes chance only. Bias analysis puts a number on the other two explanations. Two simple tools cover most needs.

- **Misclassification.** If you can state the likely Se and Sp of your exposure measure (from a validation sub-study or the literature), the formula of §17.2.2 recovers the table you would have seen with perfect measurement.
- **Unmeasured confounding.** Describe a hypothetical confounder U by two risk ratios: RR_EU, how much more common U is in the exposed than the unexposed, and RR_UD, how much U raises the risk of the outcome. The **bounding factor** B = RR_EU × RR_UD ÷ (RR_EU + RR_UD − 1) is the largest factor by which such a confounder could have inflated the observed RR (Ding & VanderWeele, 2016). Divide the observed RR and its confidence limits by B to get the smallest true RR compatible with that confounder. The E-value is the value of RR_EU = RR_UD that makes B equal to the observed RR.

DERIVATION

The E-value formula. Set RR_EU = RR_UD = x and ask when the bounding factor equals the observed RR:

1. $x^2 \div (2x - 1) = RR$
2. $x^2 - 2RR \cdot x + RR = 0$
3. $x = RR \pm \sqrt{(RR^2 - RR)}$. The larger root (the one above RR) is the relevant one:

$$\text{E-value} = RR + \sqrt{[RR \times (RR - 1)]}$$

For an RR below 1, first take its inverse; "If the original risk ratio is below 1, then one first takes the inverse before applying the E-value formula" (Mathur et al., 2018). Compute it for the point estimate and for the confidence limit closer to the null (VanderWeele & Ding, 2017); if the interval already includes 1, the E-value for the limit is 1.

ILLUSTRATION / EXAMPLE

The infant cohort, after adjustment for PHC: RR_MH 2.93 (95% CI 1.60 to 5.35) (recall §11.7). A reviewer asks whether household wealth, which was not recorded, could explain the association.

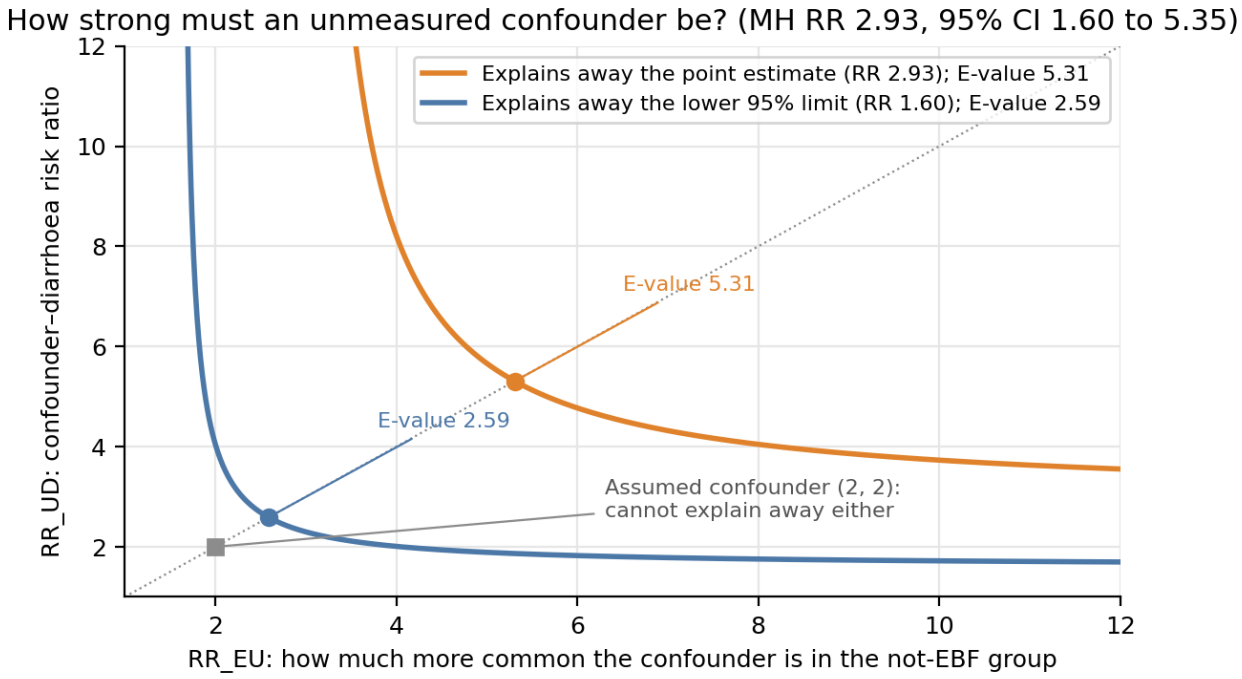


Figure 17.2. Every pair (RR_EU, RR_UD) on or above the orange curve could explain away the MH RR of 2.93; on or above the blue curve, the lower confidence limit of 1.60. The E-values are where each curve crosses the diagonal.

WORKED CALCULATION

1. E-value for the point estimate: 2.930 + √(2.930 × 1.930) = 2.930 + √5.655 = 2.930 + 2.378 = **5.31**.
2. E-value for the lower limit (unrounded 1.604): 1.604 + √(1.604 × 0.604) = 1.604 + 0.984 = **2.59**.
3. Reading: to explain away the association entirely, an unmeasured confounder would need to be about 5.3 times as common among not-EBF infants *and* raise diarrhoea risk about 5.3-fold, beyond PHC. To pull the lower limit down to 1, both associations would need to reach about 2.6.
4. A specific confounder: suppose poverty is twice as common among not-EBF infants (RR_EU = 2) and doubles diarrhoea risk (RR_UD = 2). B = (2 × 2) ÷ (2 + 2 − 1) = 4 ÷ 3 = 1.33. Corrected RR ≥ 2.930 ÷ 1.333 = **2.20**, lower limit ≥ 1.604 ÷ 1.333 = **1.20**. Such a confounder could not explain the association away.
5. Misclassification: had the recorded table been the non-differential one of §17.2.2 (27.8/48.2 and 19.2/104.8) and the Se and Sp known to be 0.80 and 0.90, the correction gives a = (27.8 − 0.10 × 47) ÷ 0.70 = 23.1 ÷ 0.70 = **33.0** and b = (48.2 − 0.10 × 153) ÷ 0.70 = 32.9 ÷ 0.70 = **47.0**: the true table, and RR 3.54, are recovered. With wrong assumed values (Se 0.90, Sp 0.95) the same calculation

gives RR 2.77. The correction is only as good as the Se and Sp you feed it, which is why bias analyses report a range of values.

IN R (OPTIONAL)

```
d ← read.csv("data/ebf_diarrhoea_cohort.csv")
library(epiR)
tab ← xtabs(~ factor(not_ebf, levels = c(1, 0)) +
              factor(diarrhoea, levels = c(1, 0)) + phc, data = d)
mh ← epi.2by2(tab, method = "cohort.count")$massoc.detail$RR.mh.wald
mh

# E-value (VanderWeele & Ding, 2017): RR + sqrt(RR * (RR - 1))
evalue ← function(rr) {
  if (rr < 1) rr ← 1 / rr
  rr + sqrt(rr * (rr - 1))
}
c(point = evalue(mh$est), ci_limit = evalue(mh$lower))

# Bounding factor for an assumed confounder (Ding & VanderWeele, 2016)
bf ← function(rr_eu, rr_ud) rr_eu * rr_ud / (rr_eu + rr_ud - 1)
round(unlist(mh) / bf(2, 2), 2)

# Simple bias analysis: back-correct a misclassified exposure
correct ← function(x_obs, n, se, sp)
  (x_obs - (1 - sp) * n) / (se + sp - 1)
# observed 2x2 (Se 0.80, Sp 0.90)
obs ← c(a = 27.8, c = 19.2, b = 48.2, d = 104.8)
a ← correct(obs["a"], 47, 0.80, 0.90)
b ← correct(obs["b"], 153, 0.80, 0.90)
rr ← (a / (a + b)) / ((47 - a) / (47 - a + 153 - b))
c(a = unname(a), b = unname(b), RR = unname(rr))
```

```
       est    lower    upper
1 2.930233 1.604305 5.352015
   point ci_limit
5.308474 2.588931
  est lower upper
 2.20  1.20  4.01
        a         b        RR
33.000000 47.000000  3.535714
```

The first block is the MH RR with its Wald interval, 2.93 (1.60 to 5.35), as in §11.7. The E-values 5.31 and 2.59 are the confounder strengths needed to explain away the estimate and the lower limit. Dividing by the bounding factor 1.33 for an assumed (2, 2) confounder leaves 2.20 (1.20 to 4.01), still above 1. The last line recovers the true cells 33 and 47 and RR 3.54 from the misclassified table. The EValue and episensr packages automate these calculations but are not installed here; the base-R functions above are the whole method.

MUST-KNOW NOTES (MD-LEVEL)

- An E-value is not a test and not a probability. A large E-value says only that a strong confounder would be needed; whether one plausibly exists is a subject-matter judgement. Compare it with the strength of known confounders: here the rural PHC was 2.5 times as common among not-

EBF infants (62.5% against 25.0%) and raised diarrhoea risk 1.67-fold among EBF infants (16.7% against 10.0%), both well short of 5.31.

- The E-value is on the RR scale. It "can also be used for hazard ratios or odds ratios with outcomes that are rare at the end of follow-up" (Mathur et al., 2018); for an OR with a common outcome, such as the diarrhoea OR of 5.32 here (risk 23.5% overall), use the RR, or an approximate conversion.
- Pre-specify sensitivity analyses in the protocol: complete-case against multiple imputation (recall §15.4.3), per-protocol against intention-to-treat (§15.9), and alternative definitions of exposure and outcome. Report them whether or not they agree.
- Good practice for bias analysis is to state the bias model, justify the parameter values from validation data or literature, and present results over a range (Lash et al., 2014).

CHECKPOINT 17.4

1. Two studies each find no association between a workplace exposure and a health outcome. Study X sampled workers well but did not measure smoking, a known confounder. Study Y measured everything but recruited from a clinic that sees mainly workers who fell ill. Explain why these problems differ, how each could be addressed, and how each could have *hidden* a real effect.
2. An adjusted RR is 1.8 (95% CI 1.2 to 2.7). Compute the E-values for the estimate and the confidence limit. Would a confounder with RR_EU = RR_UD = 2 explain the association away? (For a protective RR of 0.5, what is the E-value?)
3. A cohort adjusts for maternal education recorded only as "literate / not literate", and an association remains. A colleague calls what remains "unmeasured confounding". Correct her. Then name the three design-stage methods of controlling confounding and one limitation of each.

(Answers in the Appendix at the end of the book.)

17.4 Reporting Guidelines and Writing the Statistical Methods

DEFINITION

A **reporting guideline** is a checklist, often with a flow diagram, that states the minimum information a report of a given study design must contain so that readers can judge its validity and reproduce it. The EQUATOR Network ("Enhancing the QUAlity and Transparency Of health Research") collects them. The **statistical methods** section states, for each analysis, what was done, with what software, to which data, and why.

SIMPLIFIED EXPLANATION

Everything in this chapter is invisible to a reader unless you report it: how people were selected, how many were lost, how exposure was measured, which confounders were adjusted for and why. The guidelines are the lists of what reviewers and readers need. Choose the guideline by design at the pro-

tocol stage, not after the paper is written, because some items (a registered protocol, a sample size calculation, a flow diagram) cannot be produced afterwards.

ILLUSTRATION / EXAMPLE

Guideline by design (EQUATOR Network):

Design	Guideline	Notes
Randomised trial	CONSORT	2025 update, "a 30-item checklist and a flow diagram" (CONSORT website); extensions for cluster and non-inferiority trials
Trial protocol	SPIRIT	
Observational (cohort, case-control, cross-sectional)	STROBE	"a checklist of 22 items" (Vandenbroucke et al., 2007)
Diagnostic accuracy	STARD	
Prediction model	TRIPOD	TRIPOD+AI (Collins et al., 2024) supersedes TRIPOD 2015 (recall §15.5.2)
Systematic review and meta-analysis	PRISMA	PRISMA 2020: a 27-item checklist and a flow diagram (Page et al., 2021; recall §15.7.4)
Case report	CARE	
Qualitative research	SRQR, COREQ	§17.5
Economic evaluation	CHEERS	

A statistical-methods paragraph template. Fill each bracket; delete what does not apply.

> [Design] of [population, setting, dates]. The sample size of [n] was calculated to [estimate / detect] [quantity] with [precision / power] at α = [value], allowing [x]% non-response. Continuous variables are summarised as mean (SD) or median (IQR) according to their distribution, categorical variables as n (%). The primary outcome, [outcome], was compared between [groups] by [measure: risk ratio, mean difference ...] with 95% confidence intervals. [Confounders], chosen in advance from [a causal diagram / the literature], were adjusted for by [method]. [Missing data: extent and handling]. [Sensitivity analyses]. Analyses used [software, version, packages].

Filled in for the infant cohort (recall §10.6.4 and §11.7): "Prospective cohort of 200 infants registered at two PHCs ... Risk of diarrhoea was compared between infants not exclusively breastfed and exclusively breastfed infants by the risk ratio with 95% confidence interval. PHC, identified in advance as a potential confounder, was adjusted for by Mantel–Haenszel stratification; homogeneity across PHCs was assessed by Woolf's test. The robustness of the adjusted estimate to unmeasured confounding was assessed with the E-value. Analyses used R 4.x with the epiR package."

Table 1 conventions. Table 1 describes the participants by group (exposure groups in a cohort, arms in a trial, cases and controls in a case-control study): numbers, then each characteristic as n (%), mean (SD) or median (IQR).

Characteristic	Not EBF (n = 80)	EBF (n = 120)
Registered at PHC Rural, n (%)	50 (62.5)	30 (25.0)
Registered at PHC Urban, n (%)	30 (37.5)	90 (75.0)

In a randomised trial, do not add a column of p-values testing baseline differences (recall §15.9). According to the CONSORT statement such testing "should not be performed" (de Boer et al., 2015). In an observational study, the imbalance in Table 1 is information about confounding (here, PHC); describe it rather than test it.

Results: effect and interval before p. The ICMJE asks authors to "quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals)" and to "Avoid relying solely on statistical hypothesis testing, such as P values, which fail to convey important information about effect size and precision of estimates", and to "Specify the statistical software package(s) and versions used" (ICMJE).

- Weak: "Not being exclusively breastfed was significantly associated with diarrhoea ($p < 0.001$)."
- Better: "Diarrhoea occurred in 41.3% of infants not exclusively breastfed and 11.7% of exclusively breastfed infants (RR 3.54, 95% CI 2.02 to 6.17); after adjustment for PHC, RR 2.93 (1.60 to 5.35)."

MUST-KNOW NOTES (MD-LEVEL)

- Many of reviewers' statistical objections are predictable; check before submission: sample size not justified; flow of participants and losses not shown; confounders chosen by significance rather than reasoning (recall §11.7); p-values without effect sizes; baseline p-values in a trial; subgroup claims without an interaction test (recall §11.7); missing data unmentioned (§15.4); multiple comparisons unaddressed (recall §8.5).
- Report numbers to a precision the data support; this book gives risk ratios to two decimals and percentages to one.
- Attach the completed checklist when you submit; many journals require it.

17.5 Qualitative and Mixed Methods – What a Quantitative Researcher Must Know

DEFINITION

Qualitative research studies meanings, experiences and processes through words and observation rather than numbers, asking *how* and *why* rather than *how many*. **Mixed-methods research** combines qualitative and quantitative data in one study and integrates them at the design, methods or interpretation stage (Fetters et al., 2013).

SIMPLIFIED EXPLANATION

A survey can tell you that 40% of infants are not exclusively breastfed. It cannot tell you why mothers give water in the first month, who advises them, or what an ASHA's message sounds like at the doorstep. Those questions need qualitative methods, and the answers often decide whether a quantitative intervention works.

The logic of sampling differs. A quantitative sample is chosen to represent a population so that estimates generalise. A qualitative sample is chosen for information: **purposive sampling** selects people who can speak to the question (recall §6.3.2), and **theoretical sampling** chooses the next participants to test ideas emerging from the analysis so far. Sampling stops at **saturation**, when new data stop yielding new codes

or themes. In a systematic review of empirical tests, saturation was reached within 9–17 interviews or 4–8 focus group discussions, "particularly those with relatively homogenous study populations and narrowly defined objectives" (Hennink & Kaiser, 2022). Treat these as a planning range, not a rule.

ILLUSTRATION / EXAMPLE

Three data-collection methods recur in community medicine:

Method	What it is	Use it for
In-depth interview (IDI)	One participant, open questions from a topic guide	Personal experience, sensitive topics
Focus group discussion (FGD)	A small group of similar participants led by a moderator	Shared norms, community language, range of views
Key informant interview (KII)	One person chosen for position or expertise (a medical officer, an anganwadi supervisor)	How a programme or system works

Thematic analysis is "a method for identifying, analysing and reporting patterns (themes) within data" (Braun & Clarke, 2006). Its six phases: (1) familiarising yourself with your data; (2) generating initial codes; (3) searching for themes; (4) reviewing themes; (5) defining and naming themes; (6) producing the report (Braun & Clarke, 2006). A theme is not decided by counting: "The 'keyness' of a theme is not necessarily dependent on quantifiable measures – but rather on whether it captures something important in relation to the overall research question" (Braun & Clarke, 2006).

Rigour is judged by trustworthiness rather than validity and reliability: "Quality criteria for all qualitative research are credibility, transferability, dependability, and confirmability" (Korstjens & Moser, 2018). Credibility (are the findings believable? checked by triangulation, comparing data sources or methods, and member checking, taking findings back to participants), transferability (thick description so readers can judge fit to their setting), dependability (an audit trail of decisions) and confirmability (findings grounded in data, with the researcher's reflexivity, an account of how their own position may have shaped the data, stated).

Mixed-methods designs (Fetters et al., 2013):

- **Convergent**: collect both at the same time, analyse separately, compare.
- **Explanatory sequential**: quantitative first, then qualitative to explain the results (a survey finds low IFA adherence; IDIs with non-adherent women explain why).
- **Exploratory sequential**: qualitative first, then quantitative to build or test (FGDs generate items for a new questionnaire, then validated as in Chapter 16).

MUST-KNOW NOTES (MD-LEVEL)

- Report qualitative studies with COREQ (a "32-item checklist for interviews and focus groups", Tong et al., 2007) or SRQR (21 items, O'Brien et al., 2014).
- Qualitative findings are not generalised statistically; their transferability is judged by the reader from the description of setting and participants.
- Do not add p-values to qualitative counts, and do not treat an FGD's majority view as a prevalence.

CHECKPOINT 17.5

1. Name the reporting guideline for each: (a) a cluster-randomised trial of weekly iron–folic acid in schools; (b) a case-control study of neonatal sepsis; (c) the accuracy of a digital haemoglobin meter against the laboratory; (d) a systematic review of yoga for hypertension; (e) FGDs with ASHAs on home visits; (f) a risk score predicting pre-eclampsia.
2. Rewrite for a results section: "Not exclusively breastfed infants had significantly more diarrhoea ($\chi^2 = 23.37$, $p < 0.001$)." Use the numbers of §10.6.4.
3. A district survey finds that only 35% of pregnant women took 100 IFA tablets. The team then interviews 15 women who did not, choosing them to include different ages, parities and blocks, and stops when new interviews add no new reasons. Name the mixed-methods design, the sampling approach and the stopping rule.

(Answers in the Appendix at the end of the book.)

Checkpoint Answers

Try each checkpoint from memory before reading the answer below it. The model answers show the working for numeric questions; your own wording will often differ, and that is fine as long as the reasoning matches.

Chapter 1: Statistics – Foundations of Data

Checkpoint 1.1

1. Explain, without using the word "variable", why "sex" might be a variable in one dataset but a constant in another.

In a general OPD register some records say "male" and others "female", so sex tells you something that differs from one person to the next. In a dataset of pregnant women every record says "female". There, sex is identical for every unit, so it carries no information that distinguishes one record from another: it is a constant in that dataset.

2. A field worker records "type of latrine used" and later reports its "mean" across households. What is wrong with that statement, and what should be reported instead?

Type of latrine is qualitative (nominal): "pit latrine" and "flush toilet" are categories, not amounts, so averaging them (or their numeric codes) is meaningless. Report the frequency and percentage of households in each category, and the mode if one typical value is needed.

3. The clinic register holds 100 children and 11 columns. How many data points (cells) does it contain? Classify "age in completed months" and "number of diarrhoeal episodes in the past month" as discrete or continuous, with a reason.

Cells = 100 × 11 = 1,100. Number of diarrhoeal episodes is discrete: it is a count (0, 1, 2 ...) and 1.5 episodes is impossible. Age is continuous in nature (time passes continuously); "completed months" is a recording choice that rounds it down to whole months. As recorded, it takes only separate values (12, 13 ... 36), so treating it as discrete is also acceptable if you state that it is age in completed months.

Checkpoint 1.2

1. Without naming the four scales, describe the one extra "permission" each step up the measurement ladder unlocks.

First level: only same or different. Second: adds ordering (higher or lower). Third: adds meaningful subtraction, because gaps are equal. Fourth: adds meaningful division, because zero means none of the quantity.

2. A colleague says "0°C means there is no heat." Which scale is she confusing Celsius with? Convert 10°C and 20°C to kelvin and compute the true ratio of the two temperatures.

She is treating Celsius (interval: arbitrary zero at the freezing point of water) as a ratio scale. Kelvin starts at absolute zero, so it is a ratio scale:

1. 10°C = 10 + 273.15 = 283.15 K.
2. 20°C = 20 + 273.15 = 293.15 K.
3. Ratio = 293.15 ÷ 283.15 = 1.035.

20°C is about 3.5% "hotter" than 10°C in absolute terms, not twice as hot.

3. In a dataset of 10 children, sex is coded Male = 1, Female = 2, and the "mean sex" comes out as 1.4. What, if anything, does 1.4 tell you?

As a measurement, nothing: the codes are labels. But because sex is binary, the mean of the codes has a hidden meaning. Mean = 1 + (proportion coded 2), so 1.4 = 1 + 0.4: 40% of the children (4 of 10) are girls. The correct report is "4 girls (40%) and 6 boys (60%)"; coding 0/1 would give 0.4 directly as the proportion.

Checkpoint 1.3

1. A colleague says "if you sampled correctly, the sample mean should equal the population mean exactly." Is that true?

No. Only a census (sample = population) guarantees that. Any particular random sample gives a statistic ($\bar{x}$) that usually differs from the parameter (μ) by sampling error, even when the sampling was done perfectly. Correct sampling makes the error unbiased and quantifiable (Chapters 6–7), not zero.

2. Give a population and a sample for a study you might run as a Community Medicine resident, and name which of your numbers is the parameter and which the statistic.

Population: all pregnant women attending antenatal care in the district this year. Sample: the 200 women you interview. The true mean haemoglobin of all of them is the parameter (μ); the mean haemoglobin of your 200 women is the statistic ($\bar{x}$).

3. In the CBHI survey, 150 of 500 sampled households are members. Compute the membership proportion, give its symbol, and say whether it is a parameter or a statistic. Is the survey dataset primary or secondary data for a resident who reanalyses it five years later?

$\hat{p}$ = 150 ÷ 500 = 0.30 (30%). It is computed from a sample, so it is a statistic, written $\hat{p}$; it estimates the population proportion p (written π or P in some texts). For the original survey team the data were primary; for a resident reanalysing them later they are secondary data, collected by someone else for another purpose.

Checkpoint 1.4

1. List, in order, the steps that turn a raw, unsorted list of 200 haemoglobin readings into something you could read a percentile from.

1. Start from the raw list. (2) Sort it and choose class intervals with unambiguous limits (e.g. 7.0–7.9, 8.0–8.9 … or half-open classes). (3) Count the frequency in each class (grouped, because 200 readings to 0.1 g/dL have many distinct values). (4) Add relative frequency and cumulative frequency (and cumulative relative frequency). (5) Read or interpolate the percentile from the cumulative relative frequency column, using class boundaries.

2. Why group haemoglobin readings but not "number of children per household"?

Haemoglobin is continuous with many distinct decimal values; listing each would give a table almost as long as the data. Number of children per household takes a few small whole numbers (0, 1, 2, 3 …), so an ungrouped table stays short and loses nothing.

3. Using the diarrhoea table in §1.4.2 (0: 41, 1: 35, 2: 16, 3: 5, 4: 3), compute the cumulative frequency up to 1 episode and the percentage of children with 2 or more episodes.

1. Cumulative frequency up to 1 episode = 41 + 35 = 76 children.
2. Children with 2 or more = 16 + 5 + 3 = 24 (equivalently 100 − 76 = 24).
3. Percentage = 24 ÷ 100 = 24%.

Checkpoint 1.5

1. Why can a hospital-based cancer registry describe the pattern of cancers among its patients but not the incidence of cancer in the district?

Incidence needs new cases divided by the population they arose from. An HBCR records the patients who came to that hospital, from an undefined catchment and filtered by referral, distance and cost; it has no population denominator. It can say what share of *its patients* have each cancer site. A population-based registry, which tries to record every new case in a defined area with a census-based population, can give incidence rates.

2. A district's IDSP data show acute diarrhoeal disease cases doubling this month. Give three explanations other than a true doubling of disease.

1. Better reporting: more sub-centres or facilities reported this month (completeness rose), or a new private facility started reporting. (2) More care-seeking: a campaign, a new facility or free ORS distribution brought more people to health workers. (3) Artefacts: late reports from last month entered this month, duplicate entries, or a change in case definition or form. Check completeness and timeliness before reading the trend.

3. In a month, 50 primary health centres (PHCs) should report; 44 reports arrive, 36 of them by the deadline. Compute completeness and timeliness (as a share of expected reports).

1. Completeness = 44 ÷ 50 = 88%.
2. Timeliness = 36 ÷ 50 = 72% of expected reports (36 ÷ 44 = 81.8% of reports received).

Checkpoint 1.6

1. A resident surveys 400 randomly chosen women in a block on one day, recording current tobacco use and current hypertension. Name the design and the measure of disease frequency it can give. Why can it not show that tobacco use preceded hypertension?

Cross-sectional survey; it gives the prevalence of hypertension (and of tobacco use) at that time. Exposure and outcome are measured at the same moment, so the survey cannot show which came first; a woman might, for example, have changed her tobacco use after being told she had hypertension.

2. A case-control study of 80 neonatal tetanus cases uses 240 controls. What proportion of the study sample are cases? Why is this not the incidence of neonatal tetanus?

Proportion = 80 ÷ (80 + 240) = 80 ÷ 320 = 25%. The investigator chose 3 controls per case; with 1 control per case the figure would be 50%. The proportion reflects the sampling ratio, not the number of new cases among all births, so it estimates neither incidence nor prevalence.

3. Districts with more mobile phones per household have lower infant mortality. Name the design and the error of concluding that a mother's phone lowers her own infant's risk of death.

Ecological study (units are districts). Concluding an individual-level effect from a group-level association is the ecological fallacy: district phone ownership may simply mark wealth, roads and facilities, and within districts the infants who died may have had phones as often as those who survived.

Checkpoint 1.7

1. Write codebook rows (name, type, values, missing code) for "ever breastfed (yes/no)" and "age at first solid food in months" in an infant-feeding survey.

Variable	Label	Type (scale)	Values / range	Missing
ever_bf	Ever breastfed	binary (nominal)	1 = yes, 0 = no	empty cell
age_solids_mo	Age at first solid food	continuous (ratio), recorded in completed months	0–24 months	empty if not recorded; "not yet started" (not applicable) recorded in a separate variable (e.g. solids_started = 1 yes / 0 no), never as 0 or a code in this column

2. In a double-entry comparison of 250 forms with 40 fields each, 30 cells disagree. What percentage of cells disagree, and what do you do with each disagreeing cell?

Cells = 250 × 40 = 10,000. Disagreement = 30 ÷ 10,000 = 0.3%. For each of the 30 cells, pull the original paper form, read the true value, and correct whichever entry is wrong, recording the correction in the cleaning log or script. Cells where both typists made the same error are not caught, which is why range and consistency checks follow.

3. (Optional, R) Why does `table(cl$nutritional_status)` list "moderate wasting" first if you skip the `factor(...)` line?

Without the `factor(..., levels = ..., ordered = TRUE)` line, the column is plain text, and R sorts text alphabetically: "moderate wasting" < "normal" < "severe wasting". The clinical order (normal < moderate < severe) exists only in your head until you declare it as an ordered factor.

Chapter 2: Descriptive Statistics – Central Tendency

Checkpoint 2.1

1. A district has three villages. Village P has 1,000 eligible children with 60% measles vaccination coverage; villages Q and R have 100 eligible children each, both with 95% coverage. Compute the unweighted mean of the three coverages and the district's true coverage, and explain which weight you used.

Unweighted mean = (60 + 95 + 95) ÷ 3 = 83.3%. This gives the two small villages two-thirds of the say, although they hold only 200 of the 1,200 eligible children.

Coverage is a per-child quantity, so weight each village by its number of eligible children: vaccinated = 0.60 × 1000 + 0.95 × 100 + 0.95 × 100 = 600 + 95 + 95 = 790; district coverage = 790 ÷ 1200 = **65.8%**.

The unweighted mean overstates coverage by about 17.5 percentage points. This is total numerator ÷ total denominator, the weighted mean of §2.2.2.

2. A population grew 10% one year and 20% the next. Compute the average annual growth rate with the geometric mean and compare it with the simple average of 10% and 20%.

Growth compounds: the population is multiplied by 1.10 and then by 1.20, a total factor of 1.32. GM = $\sqrt{(1.10 \times 1.20)} = \sqrt{1.32} = 1.1489$, so average growth is **14.9% per year**. Check: $1.1489^2 = 1.32$. The simple average is 15%, and $1.15^2 = 1.3225$ overshoots the true 1.32. The gap is small here because the two rates are close. It widens as the rates differ more (GM ≤ AM, with equality only when all values are equal).

3. An ASHA walks 12 km to a sub-centre at 4 km/h and returns the same 12 km by shared auto at 12 km/h. What is her average speed for the round trip, and which mean gives it directly?

Time out = 12 ÷ 4 = 3 h; time back = 12 ÷ 12 = 1 h. Average speed = 24 km ÷ 4 h = **6 km/h**. The distances (the numerators of the speeds) are equal, so the harmonic mean gives it directly: HM = 2 ÷ (1/4 + 1/12) = 2 ÷ (4/12) = 6 km/h. The arithmetic mean, 8 km/h, is wrong because she spends three times as long at the slower speed.

Checkpoint 2.2

1. Hospital length of stay is right-skewed. In one ward the mean stay is 6.2 days and the median is 4 days. Which would you report as the typical stay, and why could the mean alone mislead?

Report the median, 4 days (with the IQR). A few very long stays pull the mean up to 6.2 days, above what most patients experience: at least half of them stay 4 days or less. The mean remains useful for a different question, total bed-days (mean × number of patients), but it does not describe a typical patient.

2. For a dataset of blood groups (A, B, AB, O), which single measure of central tendency is valid, and why do the other two fail?

Only the mode. Blood group is nominal: its categories have no numeric values to average and no order to find a middle in, so the mean and median are undefined. The mode just reports the commonest category.

3. Haemoglobin of 50 pregnant women at a PHC is grouped as 7 to <9 g/dL (6 women), 9 to <11 (14), 11 to <13 (20) and 13 to <15 (10). Compute the grouped mean, median and mode. Then check the mode against the empirical relation mode ≈ 3 × median − 2 × mean.

Mid-points 8, 10, 12, 14; h = 2; N = 50; cumulative frequencies 6, 20, 40, 50.

Mean = (6×8 + 14×10 + 20×12 + 10×14) ÷ 50 = (48 + 140 + 240 + 140) ÷ 50 = 568 ÷ 50 = **11.36 g/dL**.

Median: N/2 = 25 falls in 11 to <13 (cumulative 40). L = 11, cf = 20, f = 20. Median = 11 + [(25 − 20) ÷ 20] × 2 = 11 + 0.5 = **11.5 g/dL**.

Mode: modal class 11 to <13 ($f_1 = 20$), $f_0 = 14$, $f_2 = 10$. Mode = 11 + [(20 − 14) ÷ (40 − 14 − 10)] × 2 = 11 + (6 ÷ 16) × 2 = **11.75 g/dL**.

Empirical relation: 3 × 11.5 − 2 × 11.36 = 34.5 − 22.72 = 11.78 g/dL, within 0.03 of the formula mode. The order mean < median < mode usually indicates, in single-peaked data, a slight left skew: a longer tail toward low haemoglobin.

Chapter 3: Descriptive Statistics – Dispersion

Checkpoint 3.1

1. Two datasets have the same range. Explain how they could still look completely different, and which measure from §3.3–3.4 would catch the difference.

The range uses only the two extremes. One dataset can be tightly clustered with one value at each end (9, 11, 11, 11, 13) while the other is spread evenly (9, 10, 11, 12, 13); both have range 4. The interquartile range, built from the middle 50% of the data, catches the difference: by the type 7 rule the first dataset has Q1 = Q3 = 11 and IQR = 0, and the second has IQR = 12 − 10 = 2.

2. A child's weight is "at the 10th percentile" on a growth chart. Explain what that means without using the word "percentile".

Among children of the same age and sex in the reference population, 10% weigh less than this child and 90% weigh more. The child is on the light side of the reference, but not extremely so.

3. For six women's haemoglobin values, 9.0, 10.5, 11.0, 11.5, 12.0 and 14.0 g/dL, find Q1 and Q3 by the type 7 rule, the IQR, the quartile deviation and the coefficient of quartile deviation.

n = 6, already sorted. Q1 position = 1 + 5 × 0.25 = 2.25 → 10.5 + 0.25 × (11.0 − 10.5) = **10.625 g/dL**. Q3 position = 1 + 5 × 0.75 = 4.75 → 11.5 + 0.75 × (12.0 − 11.5) = **11.875 g/dL**. IQR = 11.875 − 10.625 = **1.25 g/dL**. QD = 1.25 ÷ 2 = **0.625 g/dL**. Coefficient of QD = 1.25 ÷ (11.875 + 10.625) = 1.25 ÷ 22.5 = **0.056**.

Checkpoint 3.2

1. Explain why variance squares the deviations instead of averaging them directly, and name one property squared deviations have that absolute deviations lack.

Raw deviations from the mean always sum to zero ($\Sigma(x - \bar{x}) = 0$), so their average is zero for every dataset and says nothing about spread. Squaring removes the signs. Unlike absolute values, squares split cleanly: $\Sigma(x - c)^2 = \Sigma(x - \bar{x})^2 + n(\bar{x} - c)^2$ for any c. This identity shows that the mean minimises the sum of squares, gives the proof that $E[\Sigma(x - \bar{x})^2] = (n - 1)\sigma^2$, and is the start of the sums-of-squares partitions used in ANOVA and regression.

2. For the six haemoglobin values in Checkpoint 3.1 (9.0, 10.5, 11.0, 11.5, 12.0, 14.0 g/dL), compute the mean, the sum of squares, s^2 and s. Then compute SS ÷ n and say why it is not the usual estimate of σ^2.

$\Sigma x = 68.0$; mean = 68.0 ÷ 6 = **11.33 g/dL**. $\Sigma x^2 = 784.5$; $SS = \Sigma x^2 - (\Sigma x)^2 \div n = 784.5 - 4624 \div 6 = 784.5 - 770.67 =$ **13.83**. (Directly: deviations −2.33, −0.83, −0.33, 0.17, 0.67, 2.67; squares 5.44, 0.69, 0.11, 0.03, 0.44, 7.11; total 13.83.) $s^2 = 13.83 \div 5 =$ **2.77 (g/dL)²**; $s = \sqrt{2.77} =$ **1.66 g/dL**. SS ÷ n = 13.83 ÷ 6 = 2.31. Deviations are measured from $\bar{x}$, the point that minimises squared distance, and on average over repeated samples $E[SS] = (n - 1)\sigma^2$. So SS ÷ n underestimates σ^2 on average by the factor $(n - 1)/n = 5/6$. Dividing by n − 1 removes that bias.

3. A student writes: "SD = 1.66 g/dL, so on average each woman's haemoglobin differs from the mean by 1.66 g/dL." Correct the sentence, using the mean deviation of the same data.

The average absolute distance from the mean is the mean deviation: (2.33 + 0.83 + 0.33 + 0.17 + 0.67 + 2.67) ÷ 6 = 7.0 ÷ 6 = **1.17 g/dL**. A correct sentence: "The SD is 1.66 g/dL, the root-mean-square distance from the mean. It gives the two women farthest from the mean (9.0 and 14.0 g/dL) extra weight, so it exceeds the mean deviation of 1.17 g/dL."

Checkpoint 3.3

1. You want to compare how variable height is with how variable weight is in the same children. Why can't you compare raw SDs, and what do you compare instead? Compute it for heights with mean 85 cm, SD 4 cm and weights with mean 11.8 kg, SD 1.75 kg.

The SDs are in different units (cm and kg) and on different scales, so their sizes cannot be compared. Compare coefficients of variation, which are unit-free. Height: CV = 4 ÷ 85 × 100 = **4.7%**. Weight: CV = 1.75 ÷ 11.8 × 100 = **14.8%**. Relative to its own mean, weight is about three times as variable as height.

2. Describe a right-skewed distribution of hospital length of stay, and state the usual order of mean, median and mode. In one ward the mean stay is 6 days, the median 5 days and the SD 4 days: compute Pearson's coefficient of skewness and the mode implied by the empirical relation.

Most patients cluster at short stays, and a long tail stretches toward a few very long stays. For such a unimodal right-skewed curve the order is usually mode < median < mean: the mean is pulled toward the tail. This is a rule of thumb that can fail, especially for discrete data such as whole days. Pearson Sk = 3(mean − median) ÷ SD = 3 × (6 − 5) ÷ 4 = **0.75** (positive: right skew). Empirical mode ≈ 3 × median − 2 × mean = 15 − 12 = **3 days**.

3. A colleague says a platykurtic distribution "must have a flat top". Is that right? Explain what kurtosis actually measures.

No. Kurtosis, $b_2 = m_4 \div m_2^2$, is dominated by the fourth powers of the largest deviations, so it measures tail extremity: how prone the distribution is to extreme values relative to a Normal distribution ($b_2 = 3$). A platykurtic distribution (excess kurtosis < 0) has lighter tails than Normal. It often looks flatter near the centre, but it need not; the peak's shape contributes little to kurtosis (Westfall, 2014).

Chapter 4: Data Visualisation

Checkpoint 4.1

1. The five reasons for not joining CBHI (112, 98, 67, 46 and 27 households) are to go in your thesis. You want the reader to compare the reasons with one another. Would a bar diagram or a pie diagram serve better, and why?

A bar diagram. The purpose is to compare categories with each other, and the eye compares lengths on a common baseline more accurately than angles or areas (Cleveland & McGill, 1984). The number of categories is not the deciding point: five exhaustive categories (they include "other" and add to 350) could be drawn as a pie, and a pie would suit a message about the share of one whole. For "which reason is most common, and by how much?", bars ordered from largest to smallest are clearer.

2. Why must a bar diagram's value axis start at zero? Use the drawn-ratio formula to show what happens if it starts at 50 for two bars of 60 and 80.

A bar encodes its value by its length, so lengths are in the right ratio only if they are measured from zero. From baseline c = 50: drawn ratio = (80 − 50) ÷ (60 − 50) = 30 ÷ 10 = 3. True ratio = 80 ÷ 60 = 1.33. The truncated chart makes the second bar look three times the first, when it is only one-third larger.

3. In Figure 4.2, Q3 has 15 normal, 4 moderately wasted and 2 severely wasted children. What height does the moderate-wasting segment have in the percentage bar?

Q3 total = 15 + 4 + 2 = 21. Moderate wasting = 4 ÷ 21 × 100 = 19.0%. The segment runs from 71.4% (the top of the normal segment, 15/21) to 71.4 + 19.0 = 90.5%.

Checkpoint 4.2

1. Name the one visual feature that tells you at once whether you are looking at a histogram or a bar diagram, and say why that feature exists.

Histogram bars adjoin, with no gaps, because the x-axis is a continuous numeric scale and one class runs straight into the next. Bar-diagram bars stand apart, because each bar is a separate category (ordered or not) with nothing in between.

2. Using the "less than" ogive points for today's 30 weights, estimate the first quartile (the weight with 7.5 children below it).

The count 7.5 lies between the points (10, 5) and (12, 17). Interpolating: Q1 ≈ 10 + (7.5 − 5) ÷ (17 − 5) × 2 = 10 + 2.5/12 × 2 = 10.42 kg. (The Q1 of the 30 raw weights by R's default method is 10.57 kg; the ogive estimate assumes values are spread evenly within the class.)

3. A histogram of haemoglobin uses classes 7 to <9, 9 to <11 and 11 to <15 g/dL, with 10, 30 and 40 women. What height should each bar have, and which class is the most crowded per g/dL?

The widths are unequal (2, 2 and 4 g/dL), so plot frequency density = frequency ÷ width: 10 ÷ 2 = 5, 30 ÷ 2 = 15 and 40 ÷ 4 = 10 women per g/dL. The 9 to <11 g/dL class is the most crowded (15 per g/dL), even though the 11 to <15 class holds the most women. Drawn at height 40, the last bar would wrongly look the densest.

Checkpoint 4.3

1. A box plot's median line sits much closer to the top of the box than the bottom. What does that tell you about the shape of the middle half of the data?

The distance from the median to Q3 is short and the distance from Q1 to the median is long, so the lower half of the middle 50% is more spread out. That is a sign of left (negative) skew within the IQR. Check the whiskers and outliers too before describing the whole distribution.

2. For CBHI members, Q1 = 3.5 km and Q3 = 9.2 km, and distances run from 0.2 to 14.7 km. Compute the Tukey fences and say whether any member household would be drawn as a separate point.

IQR = 9.2 − 3.5 = 5.7 km; 1.5 × IQR = 8.55 km. Lower fence = 3.5 − 8.55 = −5.05 km; upper fence = 9.2 + 8.55 = 17.75 km. The minimum (0.2 km) and maximum (14.7 km) both lie inside, so no household is drawn separately, and the whiskers reach the minimum and the maximum. (With the unrounded Q3 of 9.175 km the fences are −5.01 and 17.69 km; the conclusion is the same.)

3. Why might you prefer a stem-and-leaf plot to a histogram for a small dataset?

It shows the shape of the distribution and still keeps every value, so you can read back any observation and count the median and quartiles straight off it. A histogram loses the individual values once they are grouped. With a small dataset the individual values still carry information, and the plot stays short enough to read.

Checkpoint 4.4

1. A district's measles cases rose from 50 to 200 in five years, while a neighbouring district's rose from 500 to 1,000. On a semi-log graph, which line is steeper? Give each district's annual growth factor.

First district: 200 ÷ 50 = 4-fold in 5 years; annual factor = 4^(1/5) = 1.32 (32% a year). Second district: 1,000 ÷ 500 = 2-fold; annual factor = 2^(1/5) = 1.15 (15% a year). On a semi-log graph the slope is log(annual factor), so the first district's line is steeper ($\log_{10}$ 1.32 = 0.120 vs $\log_{10}$ 1.15 = 0.060 per year), even though the second district added more cases (500 vs 150).

2. In a point-source outbreak the first case fell ill 30 hours after a community lunch and the last 90 hours after it. The agent's incubation period is taken as 24 to 96 hours. Using the first- and last-case rules, where does the exposure window fall relative to the lunch?

Latest possible exposure = first onset − minimum incubation = 30 − 24 = 6 hours after the lunch. Earliest possible exposure = last onset − maximum incubation = 90 − 96 = −6, that is 6 hours before the lunch. The window runs from 6 hours before to 6 hours after the lunch, so it contains the lunch, which is consistent with the lunch being the source. Food-specific attack rates are the next step.

3. An epi curve rises steadily for three weeks and stays high. Which of the three patterns does this suggest, and what kind of source would you look for?

A continuous common-source outbreak: the curve "rises gradually and might plateau" (CDC, n.d.-a). Look for a source that keeps exposing people, such as a contaminated well or water supply, rather than a single meal (which would give a sharp rise and fall within one incubation period) or person-to-person spread (successive peaks one incubation period apart).

Checkpoint 4.5

1. Ward P had 40 dengue cases in a population of 50,000; ward Q had 15 cases in 6,000. Which has more dots on a spot map, and which is darker on a choropleth map? Give both rates per 10,000.

Ward P has more dots (40 vs 15). Rates: P = 40 ÷ 50,000 × 10,000 = 8.0 per 10,000; Q = 15 ÷ 6,000 × 10,000 = 25.0 per 10,000. Ward Q is darker on the choropleth map: its risk is about three times ward P's.

2. A district's population is 34% aged 0–14, 60% aged 15–64 and 6% aged 65+. Compute its age dependency ratio. Is its pyramid more likely to be expansive or constrictive?

Dependency ratio = (34 + 6) ÷ 60 × 100 = 66.7 dependents per 100 working-age people (young-age ratio 34 ÷ 60 × 100 = 56.7; old-age ratio 6 ÷ 60 × 100 = 10.0). A third of the population is under 15 and the burden is almost all children, so the pyramid is more likely expansive (broad base).

3. A poster shows PHC coverage of 60% by a syringe 3 cm tall and 1 cm wide, and 90% by a syringe 4.5 cm tall and 1.5 cm wide. What ratio does the poster's area suggest, and what is the true ratio?

Areas: 3 × 1 = 3 cm^2 and 4.5 × 1.5 = 6.75 cm^2. The poster suggests 6.75 ÷ 3 = 2.25 times the coverage. The true ratio is 90 ÷ 60 = 1.5. Scaling both dimensions by 1.5 multiplied the area by 1.5 × 1.5 = 2.25. Draw more symbols, or bars, instead.

Chapter 5: Foundations of Probability

Checkpoint 5.1

1. A clinician says "I think this patient has about a 70% chance of responding to treatment", based on years of experience with similar cases. Which approach to probability is this, and why can it not be classical probability?

Subjective probability: a degree of belief built from judgement and experience. It cannot be classical, because classical probability needs a finite set of equally likely outcomes that can be counted, and a single patient's response has no such symmetric structure. (Nor is it strictly empirical: no count of repeated trials was made, although the clinician's experience is informal evidence of that kind.)

2. Write down the sample space for "number of ANC visits a woman makes, capped at 4 or more". Which of the three axioms tells you the probabilities of its outcomes must add to 1?

S = {0, 1, 2, 3, 4 or more}. The outcomes do not overlap and together make up S, so by additivity (axiom 3) their probabilities add to P(S), which is 1 by axiom 2 (total probability one).

3. In a village survey, 54 of 600 adults report a diagnosis of hypertension. What is the empirical probability? Would a second survey give exactly the same number? Which law describes what happens as the survey grows?

54 ÷ 600 = **0.09**. A second survey of 600 would almost certainly give a slightly different proportion, because each sample gives its own statistic for the same parameter. The (weak) law of large numbers says that as the number of people surveyed grows, the chance that the observed proportion is far from the true probability shrinks towards zero.

Checkpoint 5.2

1. Ten independent tests, each with a 5% chance of an abnormal result. Probability that at least one is abnormal?

P(one normal) = 0.95. P(all ten normal) = 0.95^{10} = 0.599. P(at least one abnormal) = 1 − 0.599 = **0.401**. The direct route would add P(exactly 1) through P(exactly 10), ten separate calculations; the complement needs one.

2. Two events are mutually exclusive. Explain, without using the word "exclusive", why the addition rule simplifies.

They can never happen together, so no outcome belongs to both and P(A and B) = 0. There is no double-counted overlap to subtract, and P(A or B) = P(A) + P(B).

3. P(A) = 0.2, P(B) = 0.3, A and B independent. Find P(A or B).

Independence gives P(A and B) = 0.2 × 0.3 = 0.06. Then P(A or B) = 0.2 + 0.3 − 0.06 = **0.44**.

Checkpoint 5.3

1. Explain the difference between "mutually exclusive" and "independent" using a concrete pair of events.

Mutually exclusive events cannot both happen: one person's ABO blood group being A and being B. Independent events can both happen but do not change each other's probability: two unrelated patients in

different villages each testing positive for malaria. Mutually exclusive events with non-zero probabilities are in fact dependent: knowing the blood group is A tells you with certainty it is not B.

2. How does conditional probability define independence?

A and B are independent exactly when P(A|B) = P(A): learning that B happened leaves the probability of A unchanged. Equivalently P(A and B) = P(A) × P(B).

3. P(underweight) = 0.25, P(recent diarrhoea) = 0.20, P(both) = 0.08. Find P(underweight | diarrhoea). Independent?

P(U|D) = P(U and D) ÷ P(D) = 0.08 ÷ 0.20 = **0.40**. This is not equal to P(U) = 0.25, so the events are **not independent**. The product check agrees: P(U) × P(D) = 0.25 × 0.20 = 0.05 ≠ 0.08. Underweight is more common among children with recent diarrhoea than overall. (This is an association; the data do not say which causes which.)

Checkpoint 5.4

1. Why can most positive results be false alarms when prevalence is very low?

The healthy group is far larger than the diseased group. Even a small false-positive rate, applied to that large group, produces more false positives in absolute numbers than the small diseased group produces true positives. At 1% prevalence with Se 95% and Sp 90%, 990 of 1,085 positives per 10,000 people (91%) are false.

2. Se 95%, Sp 90%, prevalence 10%. 2×2 table for 10,000 and PPV; then with odds and LR+.

Diseased 1,000: TP = 0.95 × 1,000 = 950, FN = 50. Healthy 9,000: TN = 0.90 × 9,000 = 8,100, FP = 900.

	Disease	No disease	Total
Positive	950	900	1,850
Negative	50	8,100	8,150
Total	1,000	9,000	10,000

PPV = 950 ÷ 1,850 = **51.4%** (NPV = 8,100 ÷ 8,150 = 99.4%).

Odds route: pre-test odds = 0.10 ÷ 0.90 = 0.111. LR+ = 0.95 ÷ 0.10 = 9.5. Post-test odds = 0.111 × 9.5 = 1.056. Post-test probability = 1.056 ÷ 2.056 = **0.514**, the same PPV.

3. Why can an LR be carried across settings when a PPV cannot, and what limits this?

The LR is built only from Se and Sp, each computed within one true-status group, so the arithmetic contains no prevalence; you apply it to any setting's pre-test odds. The PPV is the output of Bayes' theorem for one particular prevalence and must be recomputed for another. The limit: Se and Sp themselves vary with disease spectrum, severity and setting (Whiting et al., 2013; Bossuyt et al., 2015), so an LR from a referral hospital may not hold in community screening.

Checkpoint 5.5

1. Why is P(one exact value) zero for a continuous random variable, and what do you calculate instead?

Probability is spread over infinitely many values in any interval, so a single, infinitely precise value receives none of it. You calculate the probability of a range, as the area under the probability density curve between two points.

2. One discrete and one continuous random variable from clinical work.

Discrete: number of ANC visits, or number of children with a positive stool test out of 5 (countable values). Continuous: haemoglobin in g/dL, or systolic BP (any value in a range, limited only by the instrument). The test is whether the values are countable steps or points on a continuum.

3. For the ANC PMF, find $P(X \leq 1)$ and check $E(X) = 2.9$.

$P(X \leq 1) = 0.05 + 0.10 =$ **0.15**. $E(X) = 0 \times 0.05 + 1 \times 0.10 + 2 \times 0.15 + 3 \times 0.30 + 4 \times 0.40 = 0 + 0.10 + 0.30 + 0.90 + 1.60 =$ **2.9**.

Checkpoint 5.6

1. Binomial as a sum of Bernoulli trials; mean and variance for $n = 20$, $p = 0.8$.

Code each patient 1 (recovers) or 0 (does not): each is Bernoulli(0.8). The number who recover is the sum of the 20 codes, and that sum is Binomial(20, 0.8) when the patients are independent with the same p. Expected values add: mean $= 20 \times 0.8 =$ **16**. Variances of independent variables add: variance $= 20 \times 0.8 \times 0.2 =$ **3.2** (SD 1.79).

2. Poisson for cancer diagnoses per month but not recoveries out of 20; why are outbreak counts overdispersed?

New cancer diagnoses in a district occur one at a time, roughly independently, at a fairly steady rate, with no fixed number of trials: the Poisson assumptions. Recoveries out of 20 have a fixed number of trials (20) and a success probability per patient, which is the binomial set-up, and the count cannot exceed 20. Measles cases are contagious: each case raises the chance of further cases, so cases arrive in bursts. Some weeks have many and others none, and the variance of weekly counts exceeds the mean (overdispersion), breaking the independence assumption.

3. $\lambda = 2$ road-traffic deaths per month. P(none) and P(at least one).

$P(X = 0) = e^{(-2)} =$ **0.135**. $P(X \geq 1) = 1 - 0.135 =$ **0.865**.

Checkpoint 5.7

1. What does $z = -2$ tell you, and what does it not?

The value lies 2 standard deviations below the mean of its reference distribution. If that distribution is Normal, about 2.3% of the reference population lie lower. It does not by itself mean disease or that "something is wrong": that needs clinical context and, where relevant, an agreed cut-off (such as the WHO −2 growth cut-offs of §5.7.2.2.3).

2. Percentage beyond ±2 SD; exact multiplier for 5%.

Empirical rule: about $100 - 95 = 5\%$. Exactly: $100 - 95.45 =$ **4.55%** (2.275% in each tail). Exactly 5% lies beyond **±1.96** SD.

3. Height mean 165 cm, SD 7 cm. Percentage taller than 179 cm?

$z = (179 - 165) \div 7 = 2$. $P(Z > 2) = 1 - 0.9772 =$ **0.0228**, about 2.3%.

Checkpoint 5.8

1. Boy, 36 months, 88.0 cm; median 96.0835, SD 3.7069. HAZ, classification, centile.

HAZ = (88.0 − 96.0835) ÷ 3.7069 = −8.0835 ÷ 3.7069 = **−2.18**. Below −2 but not below −3: **stunted (moderate), not severely stunted**. Centile: P(Z < −2.18) = 0.0146, about the **1.5th centile**.

2. Why can a healthy person fall outside the ALT "normal range", and why may μ ± 1.96σ be wrong?

A reference interval is set at the 2.5th and 97.5th centiles of healthy people, so 5% of healthy people fall outside it by construction. μ ± 1.96σ gives those centiles only if the analyte is Normal in healthy people. Many analytes are skewed in healthy people (Ozarda, 2016); if ALT is, the laboratory should read the centiles directly or transform the values (for example, take logarithms), calculate the limits on that scale and transform back (Ozarda, 2016).

3. 20% anaemic; sample of 100; P(30 or more) by Normal approximation with continuity correction.

Condition: np = 20 and n(1 − p) = 80, both ≥ 5. Mean = 20; SD = √(100 × 0.2 × 0.8) = 4.0. P(X ≥ 30) ≈ P(Z > (29.5 − 20) ÷ 4) = P(Z > 2.375) = **0.009**. (The exact binomial value is 0.011: the approximation is close but slightly understates this far tail.)

Checkpoint 5.9

1. Why is t almost Normal for a large sample but not for a small one?

T = Z ÷ √(V ÷ ν), where V ÷ ν comes from the sample SD. With many degrees of freedom, s is a precise estimate of σ, V ÷ ν stays close to 1, and T behaves like Z. With few degrees of freedom, s can be well above or below σ, so the denominator varies a lot; small values of s produce large values of T, and the tails are heavier. For example, the 97.5% point is 2.776 with 4 df, 2.045 with 29 df and 1.984 with 99 df, against 1.960 for Z.

2. What are t, chi-square and F built from, and what is each used for?

All are built from independent standard Normal variables. Chi-square is a sum of ν squared standard Normals; t is a standard Normal divided by √(chi-square ÷ its df); F is the ratio of two independent chi-squares, each divided by its df. t is used for means when σ is estimated (t-tests, confidence intervals for means, regression coefficients); chi-square for tests on categorical counts and for inference on a variance; F for comparing variances and for ANOVA and overall regression tests.

3. Widening of a 95% interval for n = 10; check $2.262^2 = F(1, 9)$.

2.262 ÷ 1.960 = 1.154, so the interval is about **15% wider**. 2.262^2 = **5.117**, the 95% point of F(1, 9): the square of a t with 9 df is an F with (1, 9) df.

Chapter 6: Sampling

Checkpoint 6.1

1. A researcher must survey households across a state but has only a list of villages, not of households. Which design fits, and why would simple random sampling of households fail here?

A cluster design, in practice multistage (§6.2.4–6.2.5): select villages at random (by PPS if their sizes differ), list the households in the selected villages only, then sample households within them. SRS of households needs a complete frame of every household in the state, which does not exist; building one would cost more than the survey. SRS would also scatter the field teams across the whole state.

2. A district is 70% rural and 30% urban. A survey samples 100 rural and 100 urban households and finds that 25% of rural and 10% of urban households use a PHC for deliveries. Give the unweighted and the correctly weighted district estimate.

Unweighted: $(0.25 \times 100 + 0.10 \times 100) \div 200 = 35 \div 200 = 17.5\%$. Weighted by population shares: $0.70 \times 25\% + 0.30 \times 10\% = 17.5\% + 3.0\% = 20.5\%$. The unweighted figure is 3 points too low, because it gives urban households half the weight when they are 30% of the population. Allocation was disproportionate, so weights are required.

3. Stratified and cluster sampling both begin by dividing the population into groups. State the structural difference, and its effect on precision.

Stratified sampling samples *within every* group; cluster sampling samples *some* groups and leaves the rest out entirely. Strata are chosen to be internally alike and different from each other, so stratification removes between-stratum variation from the error and usually increases precision. Clusters are internally alike too, but because only some are sampled, that likeness means each extra unit in a cluster adds little new information, so clustering usually reduces precision (DEFF > 1). Stratification serves precision; clustering serves feasibility.

Checkpoint 6.2

1. In the PPS example, what was the chance of selection for village PV01 (population 1,450), and what is the overall chance that a given resident of PV01 is interviewed if 20 persons are taken in each selected village?

$P(\text{PV01 selected}) = k \times M_i \div M = 3 \times 1{,}450 \div 16{,}000 = 0.272$. $P(\text{resident interviewed} \mid \text{PV01 selected}) = 20 \div 1{,}450 = 0.0138$. Overall $= 0.272 \times 0.0138 = 3 \times 20 \div 16{,}000 = 0.00375$, the same as for a resident of any other village (the design is self-weighting).

2. A cluster survey takes 10 households in each of 30 villages (300 households). The ICC for the outcome is 0.05. Compute the DEFF and the effective sample size.

$\text{DEFF} = 1 + (m - 1)\rho = 1 + 9 \times 0.05 = 1.45$. Effective sample size $= 300 \div 1.45 = 206.9$, about 207 households. The 300 clustered households are worth about 207 households drawn by SRS.

3. Explain why an LQAS area with 12 of 19 children immunised is "flagged" under the rule in this section, and what error the programme risks by using d = 12 instead of d = 13.

The rule is: pass if 13 or more of 19 are fully immunised. $12 < 13$, so the area is flagged as possibly below target. With d = 12, an area whose true coverage is only 50% would still pass with probability $\beta = P(X \geq 12 \mid n = 19, p = 0.5) = 0.180$, about 1 in 5. Under d = 13 this risk is 0.084. The gain from d = 12 is a smaller α (0.023 instead of 0.068: fewer good areas wrongly flagged), but the programme would miss more failing areas.

Checkpoint 6.3

1. State the single structural difference between quota sampling and stratified sampling, and explain the specific bias that snowball sampling risks.

Both fix numbers per subgroup; stratified sampling selects *randomly* within each subgroup, quota sampling fills each quota by convenience. Snowball sampling is biased towards well-connected members of the hidden population: people who are socially isolated, or whose networks do not overlap with the initial contacts, are rarely reached however many referral chains are followed.

2. The true mean weight of under-five children in a district is 11.8 kg. A team weighs 2,000 children attending a well-funded urban clinic and gets a mean of 12.6 kg; its standard error is 0.04 kg. How large is the error of this estimate, how much of it could sampling error plausibly explain, and would weighing 8,000 clinic children fix it?

Total error = 12.6 − 11.8 = 0.8 kg. The SE is the typical size of the sampling error (§6.4): about 0.04 kg. 0.8 kg is 0.8 ÷ 0.04 = 20 times that typical size, so chance explains almost none of it: it is selection bias (the clinic's children are better-off and heavier). With 8,000 clinic children the SE falls to 0.04 ÷ √4 = 0.02 kg (four times the sample halves the SE), but the bias of about 0.8 kg remains. The answer becomes more precisely wrong. Only a better design (a probability sample of the district's children) fixes it.

3. Name three sources of non-sampling error in a household survey and say which of them would also affect a census.

Any three of: coverage (frame) error, such as households missing from an outdated list; non-response (absent or refusing households); measurement error (a faulty scale, a leading question, interviewer effects, recall error); processing error (data entry, coding). All of them can affect a census too; only sampling error is absent from a census.

Checkpoint 6.4

1. A colleague doubles her sample size and expects her standard error to halve. Is she right? What sample-size multiple would halve it?

No. SE = $\sigma/\sqrt{n}$, so doubling n multiplies the SE by $1/\sqrt{2} = 0.71$: a 29% reduction, not 50%. To halve the SE, $\sqrt{n}$ must double, so n must be multiplied by 4.

2. If the clinic survey weighed 120 children instead of 30 and the SD stayed at 1.75 kg, what would the estimated SE of the mean be?

SE = 1.75 ÷ √120 = 1.75 ÷ 10.95 = 0.16 kg. This is half the SE for 30 children (0.32 kg), as expected for four times the sample.

3. Using the rule "variances of independent variables add", show that $\mathrm{Var}(\bar{x}) = \sigma^2/n$. Then state what the Law of Large Numbers adds, and what the Central Limit Theorem adds beyond it.

$\bar{x} = (X_1 + \ldots + X_n)/n$ with independent X_i each of variance σ^2. $\mathrm{Var}(X_1 + \ldots + X_n) = \sigma^2 + \ldots + \sigma^2 = n\sigma^2$ (variances add for independent variables). $\mathrm{Var}(\bar{x}) = (1/n)^2 \times n\sigma^2 = \sigma^2/n$ (a constant factor c scales the variance by c^2). Because $E(\bar{x}) = \mu$ and $\mathrm{Var}(\bar{x}) = \sigma^2/n$ shrinks to 0, $\bar{x}$ settles at μ as n grows: that is the Law of Large Numbers. The LLN says *where* $\bar{x}$ ends up; the CLT adds the *shape* of its sampling distribution for large n, approximately Normal with mean μ and SD $\sigma/\sqrt{n}$, whatever the population's shape. That shape is what allows Normal-based confidence intervals and tests.

Chapter 7: Estimation

Checkpoint 7.1

1. Explain why Bessel's correction (dividing by n − 1) is an application of the unbiasedness property.

An estimator is unbiased if its average over all possible samples equals the parameter. Dividing the sum of squared deviations by n gives a variance whose average is only $(n - 1)/n \times \sigma^2$, because deviations are

measured from $\bar{x}$, which sits closer to the data than μ does. Dividing by $n - 1$ instead scales the average up to exactly σ^2. The correction exists to make s^2 unbiased.

2. For today's 30 weights, $s^2 = 3.07\ \text{kg}^2$. What is the divide-by-n variance, and in which direction is it biased?

Divide-by-n variance = $3.07 \times 29/30 = 2.97\ \text{kg}^2$. It is biased downwards: on average it underestimates σ^2 by the factor 29/30. The bias shrinks as n grows, so the estimator is consistent.

3. A point estimate alone is called incomplete. What is missing from it?

Any statement of precision: how far the estimate might plausibly be from the parameter. A confidence interval supplies it.

Checkpoint 7.2

1. A colleague says, "There is a 95% probability that the true mean lies in this interval." Correct the statement.

The true mean is fixed; the computed interval either contains it or does not. The 95% describes the procedure: if sampling and interval construction were repeated many times, about 95% of the intervals would contain the true mean.

2. For today's 30 children ($\bar{x} = 11.81$ kg, SE = 0.32 kg, t_{29} for 99% = 2.756), compute the 99% CI. Is it wider or narrower than the 95% CI, and why?

Margin = $2.756 \times 0.32 = 0.88$ kg. 99% CI = $11.81 \pm 0.88 = 10.93$ to 12.69 kg. It is wider than the 95% CI (11.16 to 12.47 kg) because capturing μ more often requires going further into the tails: $2.756 > 2.045$.

3. Two independent groups of 50 each have SD 2 kg and means differing by 0.8 kg. Compute SE_diff and the 95% CI for the difference. Does it include 0?

SE_diff = $\sqrt{(2^2/50 + 2^2/50)} = \sqrt{(0.08 + 0.08)} = \sqrt{0.16} = 0.40$ kg. 95% CI = $0.8 \pm 1.96 \times 0.40 = 0.8 \pm 0.78 = 0.02$ to 1.58 kg. It just excludes 0, so the data are incompatible with a zero difference at the 5% level, though the lower limit shows the true difference could be very small.

Checkpoint 7.3

1. Two of 25 children in an anganwadi test positive for a condition. Compute the Wald 95% CI and state what is wrong with it. Which intervals would you report instead?

$\hat{p} = 2/25 = 0.08$. SE = $\sqrt{(0.08 \times 0.92/25)} = 0.054$. Wald CI = $0.08 \pm 1.96 \times 0.054 = -2.6\%$ to 18.6%. The lower limit is an impossible negative proportion; with $n\hat{p} = 2$ the Normal approximation fails. Report the Wilson interval (2.2% to 25.0%) or the Clopper–Pearson exact interval (1.0% to 26.0%).

2. A surgeon reports "no wound infections in 60 caesarean sections". Using the rule of three, what infection risk is still compatible with this record?

Upper 95% limit ≈ 3/60 = 0.05. A true infection risk as high as 5% is compatible with zero infections in 60 operations. (Exact one-sided limit: 1 − 0.05^(1/60) = 4.9%.)

3. For n = 50, $P(X \le 17) = 0.016$ and $P(X \le 18) = 0.032$ for $X \sim \text{Binomial}(50, 0.5)$. Which ordered values form the 95% CI for the median?

Choose r so that $P(X \le r - 1) \le 0.025$. $P(X \le 17) = 0.016$ qualifies but $P(X \le 18) = 0.032$ does not, so $r - 1 = 17$ and $r = 18$. The upper rank is $n + 1 - r = 51 - 18 = 33$. The CI runs from the 18th to the 33rd ordered value, with coverage $1 - 2 \times 0.016 = 96.7\%$.

Checkpoint 7.4

1. You want to halve the margin of error of a planned prevalence survey. Roughly how much larger must the sample be, and why?

Four times larger. $n = Z^2p(1 - p)/d^2$, so $n \propto 1/d^2$. Halving d multiplies n by $2^2 = 4$, because the standard error shrinks only with $\sqrt{n}$.

2. Expected anaemia prevalence among adolescent girls is 10%, with a margin of ±3 percentage points at 95% confidence. Compute n_0. Then adjust for a cluster design with DEFF = 2 and 85% expected response.

$n_0 = 1.96^2 \times 0.10 \times 0.90/0.03^2 = 3.8416 \times 0.09/0.0009 = 384.2$. With the design effect: $384.2 \times 2 = 768.3$. With non-response: $768.3/0.85 = 903.9$, so approach 904 girls.

3. For the same 10%, what n_0 is needed for a relative precision of ±20%? Why is it larger than in question 2?

$d = 0.20 \times 0.10 = 0.02$. $n_0 = Z^2q/(\varepsilon^2p) = 3.8416 \times 0.90/(0.04 \times 0.10) = 864.4 \rightarrow 865$. It is larger because ±20% of 10% is ±2 points, a tighter margin than the ±3 points of question 2.

Checkpoint 7.5

1. A block with a mid-year population of 60,000 registers 1,500 live births, 42 infant deaths (27 of them neonatal), 25 stillbirths, 18 early neonatal deaths and 3 maternal deaths. Compute the CBR, IMR, NMR, PMR and MMR.

- CBR = $1{,}500/60{,}000 \times 1{,}000 = 25$ per 1,000 population.
- IMR = $42/1{,}500 \times 1{,}000 = 28$ per 1,000 live births.
- NMR = $27/1{,}500 \times 1{,}000 = 18$ per 1,000 live births.
- PMR = $(25 + 18)/(1{,}500 + 25) \times 1{,}000 = 43/1{,}525 \times 1{,}000 = 28.2$ per 1,000 births (the denominator includes stillbirths).
- MMR = $3/1{,}500 \times 100{,}000 = 200$ per 100,000 live births.

2. Why is the case fatality rate not a rate, and why can it mislead early in an outbreak?

It is a proportion (deaths among cases ÷ cases) with no time component, so it measures severity, not a rate of occurrence. Early in an outbreak, detected cases are skewed towards severe ones (inflating it), and some current cases have not yet died or recovered (deflating it).

3. Diabetes has a prevalence of 10% in a town, and people live with it for an average of 20 years. Assuming a steady state, what is the approximate annual incidence among people without diabetes?

$P/(1 - P) = I \times D$, so $0.10/0.90 = 0.111 = I \times 20$, giving $I = 0.0056$ per year, about 5.6 new cases per 1,000 people without diabetes per year. The shortcut $P \approx I \times D$ gives $0.10/20 = 5$ per 1,000; it underestimates here because a prevalence of 10% is not small.

Checkpoint 7.6

1. Two districts have age-specific death rates of 2 and 20 per 1,000 (District X) and 3 and 24 per 1,000 (District Y) in the under-60 and 60+ groups. Using a standard population of 60,000 under-60 and 40,000 aged 60+, compute both DSRs.

X: expected deaths in the standard = 2 × 60 + 20 × 40 = 120 + 800 = 920; DSR = 920/100,000 × 1,000 = 9.2 per 1,000. Y: 3 × 60 + 24 × 40 = 180 + 960 = 1,140; DSR = 11.4 per 1,000. Y's standardised mortality is 11.4/9.2 = 1.24 times X's.

2. A PHC catchment records 30 deaths where 20 were expected at state rates. Compute the SMR and its approximate 95% CI. What do you conclude?

SMR = 30/20 = 1.5. SE(ln SMR) = 1/√30 = 0.183. CI = 1.5 × exp(± 1.96 × 0.183) = 1.5 × exp(± 0.358) = 1.05 to 2.15 (exact Poisson: 1.01 to 2.14). Mortality is about 50% above what state rates predict for this age structure; the interval excludes 1, but with only 30 deaths the excess could be anywhere from a few per cent to more than double.

3. Why is indirect rather than direct standardisation used for a small PHC catchment?

Direct standardisation needs the catchment's own age-specific death rates, which rest on very few deaths per age group and are unstable. Indirect standardisation needs only the catchment's age structure and total deaths, borrowing stable age-specific rates from the standard (state) population.

Checkpoint 7.7

1. In a 5-year band, the death rate is 0.02 per person-year and those who die live on average 2.5 years in the band. Compute ${}_5q_x$.

${}_5q_x = n \times M/[1 + (n - a) \times M] = 5 \times 0.02/(1 + 2.5 \times 0.02) = 0.10/1.05 = 0.095$.

2. Using the table above, what is the probability that a 15-year-old survives to age 45?

$l_{45}/l_{15} = 89{,}161/95{,}388 = 0.935$ (93.5%).

3. If the open 60+ band had a death rate of 0.05 per person-year, what would e_{60} be? Why can e_1 exceed e_0?

$e_{60} = 1/M = 1/0.05 = 20$ years. e_1 can exceed e_0 because e_0 averages over everyone born, including the many who die in infancy; a child who survives the first year has passed that high-risk period, so its expected remaining life can be longer than a newborn's (here 68.8 vs 67.7 years).

Chapter 8: Hypothesis Testing – The Framework

Checkpoint 8.1

1. A small study reports "not statistically significant", and adds that its observed power was 30%. Explain why this does not show there is no real effect, and why the observed power adds no information. What should the authors report instead?

Failing to reject H_0 is not proving it (§8.1). A small study has a large standard error, so it can miss a real, clinically important effect (a Type II error). The "observed power" is power computed at the observed effect: $\Phi(|z| - 1.96)$, a one-to-one function of the p-value. Every result with $p > 0.05$ has observed power below about 50%, so "30%" only restates "p was well above 0.05" (Hoenig & Heisey, 2001). The authors

should report the effect estimate with its 95% confidence interval. If the interval includes clinically important effects, the study is inconclusive, not negative. Power belongs to the planning stage, for an effect size chosen in advance.

2. Explain the trade-off between Type I and Type II error using a clinical example of your choosing. What, besides a larger sample, could reduce both?

Example: testing a new antihypertensive against H_0 "no effect". A Type I error approves a useless drug; a Type II error abandons a useful one. For a fixed sample and design, lowering α (say from 0.05 to 0.01) moves the critical values outwards: fewer false approvals, more missed effects (in §8.4 power for a true 1 kg difference fell from 0.69 to 0.46). Both errors fall only if the standard error falls. Besides a larger n, a paired or matched design (each person as their own control) or a more precise measurement (less variability) lowers the SE and so reduces both.

3. For the urban–rural comparison (SE_diff = 0.406 kg, pooled SD 1.891 kg), compute the standardised effect size d and the power for a true difference of 0.5 kg. How many children per group (σ = 2 kg, z-formula) would give 80% power for 0.5 kg?

d = 0.5/1.891 = **0.26**.

Power ≈ $\Phi(\Delta/SE - 1.96) = \Phi(0.5/0.406 - 1.96) = \Phi(1.232 - 1.96) = \Phi(-0.73)$ = **0.23**. The study would miss a true 0.5 kg difference about 77% of the time.

n per group = $2(1.96 + 0.84)^2\sigma^2/\Delta^2 = 2 \times 7.84 \times 4/0.25 = 250.9 \rightarrow$ **251 per group** (with the unrounded z-values, 1.95996 and 0.84162, 251.2 → 252; the t-distribution version, which software computes, gives 252.1 → 253). Halving the effect to be detected quadruples n (63 → about 251).

Checkpoint 8.2

1. Two studies report p = 0.001 and p = 0.049. A colleague says the first is "much more significant", so its effect must be larger. What is right and what is wrong in this?

Right: both are below 0.05, and the smaller p does indicate greater incompatibility of that study's data with H_0 under its statistical model. Wrong: (1) "significant" at α = 0.05 is a yes/no classification, and both studies cross it; (2) a p-value does not measure the size or importance of an effect (ASA, 2016, principle 5). A tiny effect in a very large study can give p = 0.001, and a large effect in a small study p = 0.049. To compare effects, look at each study's effect estimate with its confidence interval.

2. A trial tests a new drug in 8 independent subgroups, each at α = 0.05, and the drug truly does nothing. What is the probability of at least one "significant" subgroup? What per-test threshold would Bonferroni use?

Family-wise error rate = $1 - (1 - 0.05)^8 = 1 - 0.95^8 = 1 - 0.663$ = **0.337**: about one chance in three of a false "finding". Bonferroni tests each subgroup at 0.05/8 = **0.00625**.

3. A district sample of 100 births has mean birthweight 2.9 kg. Test H_0: μ = 3.0 kg with σ = 0.5 kg: compute z and the two-sided p. The team, seeing the result lies below 3.0 kg, switches to a one-tailed test. What p would they report, and why is the switch wrong?

SE = 0.5/√100 = 0.05 kg. z = (2.9 − 3.0)/0.05 = **−2.00**. Two-sided p = 2 × [1 − Φ(2.00)] = **0.046**. One-tailed (H_1: μ < 3.0) p = **0.023**, half the two-sided value. The switch is wrong because the direction was chosen after seeing the data. It halves the p-value. A team that picks the tail the data point to, then tests at 0.05, rejects whenever the two-sided p is below 0.10, so its real Type I error rate is 10%, not the 5% it claims. This is p-

hacking (§8.6). The direction must be fixed at step 1, before the data are seen. Note also that H_0 here is $\mu = 3.0$ kg, not "no effect".

Checkpoint 8.3

1. Name the five questions you ask to choose a statistical test, and say why the design question comes first.

1. Type of outcome (quantitative normal or skewed, ordinal, binary, categorical, count/rate, time to event). (2) What is compared (one group vs a value, two groups, more than two, association of two variables). (3) Independent, paired/matched, or clustered observations. (4) For quantitative outcomes, whether parametric assumptions are reasonable. (5) The study design. The design comes first because it fixes which effect measures are valid: a case-control study, for example, can give only the odds ratio of exposure, never prevalence or risk (§1.5). No choice of test can recover a measure the design cannot supply.

2. Using Tables 8.1–8.4, choose the test and the adjusting model for: (a) haemoglobin before and after iron–folic acid in the same 12 women, with skewed differences; (b) a case-control study of neonatal sepsis with each case matched to a neighbourhood control, exposure "home delivery"; (c) full-immunisation coverage compared between urban and rural children in a 30-cluster survey.

1. Quantitative, two paired measurements, skewed differences: the sign test (§10.1.1). The Wilcoxon signed-rank test (§10.3) needs roughly symmetric differences, so it fits only if the skew is mild. No regression is needed for a single before–after comparison; to adjust, model the differences (linear regression after transformation, §9.4) or use a mixed model (§15.2.4).
2. Binary exposure, matched case-control pairs: matched-pair odds ratio b/c and McNemar's test (§10.8); adjusting model: conditional logistic regression (§13.2.9).
3. Binary outcome, two groups, clustered sample: difference or ratio of proportions (§10.6.4) with an SE corrected for clustering by the design effect (§6.2.6) or a survey analysis (§15.2.6); adjusting model: logistic regression with GEE or survey weights and PSUs (§15.2.5, §15.2.6). An ordinary chi-square would give a p-value that is too small.

3. In a cohort, 30 of 150 exposed and 15 of 150 unexposed people develop the outcome. Following the seven steps of §8.7, compute the risk difference, the pooled SE under H_0, z and the two-sided p, and the 95% CI of the difference (unpooled SE). State the conclusion.

1. H_0: risk exposed = risk unexposed; H_1: they differ (two-tailed).
2. $\alpha = 0.05$.
3. Binary outcome, two independent groups → two-proportion z-test (or chi-square). Pooled risk 45/300 = 0.15; under H_0 each group expects $150 \times 0.15 = 22.5$ events and 127.5 non-events, ample for the normal approximation (§5.7.2.2.4).
4. Risks 30/150 = 0.20 and 15/150 = 0.10; risk difference = **0.10**. $SE_0 = \sqrt{[0.15 \times 0.85 \times (1/150 + 1/150)]} = \sqrt{(0.1275 \times 0.01333)} = \sqrt{0.0017} =$ **0.0412**. $z = 0.10/0.0412 =$ **2.43**.
5. Two-sided $p = 2 \times [1 - \Phi(2.43)] =$ **0.015**.
6. $0.015 < 0.05$: reject H_0.
7. Unpooled SE $= \sqrt{(0.20 \times 0.80/150 + 0.10 \times 0.90/150)} = \sqrt{(0.001067 + 0.000600)} = 0.0408$. 95% CI $= 0.10 \pm 1.96 \times 0.0408 = 0.10 \pm 0.080 =$ **0.020 to 0.180**. Conclusion: the risk was 10 percentage points higher in the exposed (20% vs 10%; 95% CI 2.0 to 18.0 points; p = 0.015). The interval is wide: the data are compatible with a small (2 points) or a large (18 points) excess risk. It is an association; confounding has not been addressed.

Chapter 9: Comparing Groups – Parametric Tests

Checkpoint 9.1

1. In a block, 45 of 60 sampled children are fully immunised. Test against a target of 85%. Check the Normal-approximation condition first, then give z and the two-sided p-value.

Condition: $np_0 = 60 \times 0.85 = 51$ and $n(1 - p_0) = 60 \times 0.15 = 9$, both ≥ 5, so the Normal approximation is acceptable (a convention, §5.7.2.2.4). $\hat{p} = 45/60 = 0.75$. SE under $H_0 = \sqrt{(0.85 \times 0.15 \div 60)} = \sqrt{0.002125} = 0.0461$. $z = (0.75 - 0.85) \div 0.0461 = -2.17$. Two-sided $p = 2 \times P(Z < -2.17) = 0.030$. At $\alpha = 0.05$, coverage is below the 85% target. If the 60 children came from a few villages, the SE should allow for clustering (§6.2.6).

2. ORS use during the last diarrhoea episode was 60/100 in Block A and 45/100 in Block B. Compute the two-proportion z statistic, then the Pearson χ^2 from the expected counts, and show that $\chi^2 = z^2$.

$\hat{p}_A = 0.60$, $\hat{p}_B = 0.45$, difference 0.15. Pooled $\bar{p} = 105/200 = 0.525$. $SE_0 = \sqrt{[0.525 \times 0.475 \times (1/100 + 1/100)]} = \sqrt{0.0049875} = 0.0706$. $z = 0.15 \div 0.0706 = 2.12$ ($p = 0.034$). Expected counts: users $100 \times 105/200 = 52.5$ in each block; non-users 47.5 in each block. Every $|O - E| = 7.5$. $\chi^2 = 7.5^2 \times (1/52.5 + 1/47.5 + 1/52.5 + 1/47.5) = 56.25 \times 0.0802 = 4.51$. $z^2 = 2.124^2 = 4.51 = \chi^2$.

3. Why does the one-proportion test use p_0 in the SE while the confidence interval uses $\hat{p}$?

A test calculates how surprising the data would be if H_0 were true, so every part of the calculation, including the SE, assumes $p = p_0$. A confidence interval makes no assumption about the true value; it estimates the SE from the data, using $\hat{p}$.

Checkpoint 9.2

1. Explain why a paired t-test is generally more powerful than an independent samples t-test for detecting the same size of change.

Each person's difference removes their stable level (their "baseline" BP or haemoglobin). The SD of the differences, s_d, therefore excludes between-person variation and is usually much smaller than the SD of either measurement. A smaller SE gives a larger t for the same mean change. The independent test leaves the between-person variation in the SE.

2. Group 1: n = 12, mean SBP 128 mmHg, SD 10. Group 2: n = 15, mean 120 mmHg, SD 18. Compute the pooled t (with df) and the Welch t (with Welch–Satterthwaite df). Which SE is larger here, and why?

Pooled: $s_p^2 = (11 \times 100 + 14 \times 324) \div 25 = (1{,}100 + 4{,}536) \div 25 = 225.4$; $s_p = 15.01$. $SE = 15.01 \times \sqrt{(1/12 + 1/15)} = 15.01 \times 0.3873 = 5.82$. $t = 8 \div 5.82 = 1.38$, $df = 25$ ($p = 0.18$). Welch: $v_1 = 100/12 = 8.33$, $v_2 = 324/15 = 21.6$. $SE = \sqrt{29.93} = 5.47$. $t = 8 \div 5.47 = 1.46$. $df = 29.93^2 \div (8.33^2/11 + 21.6^2/14) = 896.0 \div (6.31 + 33.33) = 22.6$ ($p = 0.16$). The pooled SE is larger. Pooling weights each variance by its df, so the larger group's larger variance (324) dominates s_p^2, and that inflated value is then applied to the smaller group as well. Here the pooled test is conservative. When the *smaller* group has the larger variance, the pooled SE is too small and the pooled test gives too many false positives. Welch avoids both.

3. Sixteen children have mean haemoglobin 11.2 g/dL, s = 1.6 g/dL. Test against a hypothesised value of 12 g/dL (t_{15}, 97.5% point = 2.131). State t and your conclusion.

$SE = 1.6 \div \sqrt{16} = 0.4$. $t = (11.2 - 12) \div 0.4 = -2.00$, $df = 15$. $|t| < 2.131$, so do not reject H_0 at $\alpha = 0.05$ ($p = 0.064$). The 95% CI, $11.2 \pm 2.131 \times 0.4 = 10.35$ to 12.05 g/dL, includes 12 but also values well below it: the study is inconclusive, not evidence that the mean is 12.

4. Nine pregnant women have haemoglobin measured before and after a month of supervised IFA. The mean change (after − before) is 0.45 g/dL with s_d = 0.60 g/dL. Compute the paired t, its df and the 95% CI (t_8, 97.5% point = 2.306), and state your conclusion.

$SE = s_d \div \sqrt{n} = 0.60 \div \sqrt{9} = 0.20$ g/dL. $t = 0.45 \div 0.20 = 2.25$, $df = 9 - 1 = 8$. $|t| < 2.306$, so do not reject H_0 at $\alpha = 0.05$ ($p = 0.055$). 95% CI: $0.45 \pm 2.306 \times 0.20 = 0.45 \pm 0.46 = -0.01$ to 0.91 g/dL. The data are compatible

with no change but also with a rise of nearly 1 g/dL: nine women are too few to settle it, which is not the same as showing IFA has no effect.

Checkpoint 9.3

1. A dataset fails the Shapiro–Wilk test with n = 2,000. Should you conclude it is unfit for a parametric test? What should you look at instead?

No. With n = 2,000 the test detects departures too small to matter, and the Central Limit Theorem makes the sampling distribution of the mean close to Normal even for moderately non-Normal data. Look at a Q-Q plot (and a histogram) to see the kind and size of the departure, and at skewness. Only marked skew or extreme outliers would call for a transformation or a non-parametric method.

2. A right-skewed variable, household income, is log-transformed before a t-test. The mean log income is 0.18 higher in village X than in village Y, with 95% CI 0.02 to 0.34. What does the difference represent on the original scale? Give the point estimate and CI.

A t-test gives a difference, not a coefficient. The difference of mean logs is the log of the ratio of geometric means. Back-transform: e^0.18 = 1.20, so the geometric mean income in village X is about 20% higher than in village Y. 95% CI: e^0.02 = 1.02 to e^0.34 = 1.40. It cannot be read in rupees, because it is a ratio, and it says nothing directly about the arithmetic means.

Checkpoint 9.4

1. A one-way ANOVA comparing four clinics is significant. A colleague concludes "all four clinics differ from each other". What is wrong, and what should be done next?

A significant F says only that the four means are not all equal: at least one differs from at least one other. It could be one clinic differing from the other three. To find which pairs differ, use a multiple comparison procedure: Tukey's HSD for all six pairs, or Dunnett's test if each clinic is compared with one reference clinic. Report the pairwise differences with their adjusted CIs. (These procedures do not require the F-test to be significant first.)

2. A study tests 8 independent hypotheses at $\alpha = 0.05$, all nulls true. What is the FWER? What is the Bonferroni per-test α?

FWER = $1 - 0.95^8 = 1 - 0.663 = 0.337$: about a one-in-three chance of at least one false positive. Bonferroni per-test $\alpha = 0.05 \div 8 = 0.00625$.

3. A one-way ANOVA has 4 groups and N = 24, with SSB = 6.0 and SSW = 20.0. Complete the ANOVA table and decide at $\alpha = 0.05$ (the 95% point of $F(3, 20)$ is 3.10).

Source	SS	df	MS	F
Between	6.0	4 − 1 = 3	2.0	2.0
Within	20.0	24 − 4 = 20	1.0	
Total	26.0	23		

$F = 2.0 < 3.10$: do not reject H_0 ($p = 0.15$). The data do not show a difference between the group means; they do not show the means are equal.

Checkpoint 9.5

1. Using the table in §9.5.4, compute the supplementation gain at clinic A and at clinic B, and the difference between the two gains. What does a non-zero difference of this kind represent?

Gain at A = 2.85 − 2.80 = 0.05 kg. Gain at B = 3.30 − 2.85 = 0.45 kg. Difference of gains = 0.45 − 0.05 = 0.40 kg. A difference between the effect of one factor at different levels of the other is an interaction: the benefit of supplementation depends on the clinic. Neither main effect can show this, because each averages over the other factor.

2. Haemoglobin is measured in 40 pregnant women at 0, 2 and 4 months; 9 miss the 4-month visit. What does RM-ANOVA do with those 9 women, and name two better approaches.

RM-ANOVA needs complete data, so it drops the 9 women entirely, including their 0- and 2-month values. That loses 22.5% of the sample (9 of 40) and can bias the result if the women who missed the visit differ (for example, sicker or poorer). Better: (a) a summary-measure approach, such as each woman's change per month computed from the visits she attended, analysed with a one-sample or two-sample t-test (Matthews et al., 1990); (b) a linear mixed model (§15.2.4), which uses every available measurement and allows unequal timing.

3. Four children's weights gain 0.20, 0.30, 0.10 and 0.20 kg per month. Using the summary-measure approach, compute the mean gain and its SE.

Mean = 0.80 ÷ 4 = 0.20 kg/month. Deviations 0, 0.10, −0.10, 0; SD = √(0.02 ÷ 3) = 0.0816 kg/month. SE = 0.0816 ÷ √4 = 0.041 kg/month. (A one-sample t-test against 0 would give t = 4.9 on 3 df.)

Chapter 10: Comparing Groups – Non-Parametric Tests

Checkpoint 10.1

1. In a sign test on 10 pairs, 8 differences are positive, 1 is negative and 1 is zero. What are n and X, and what is the two-sided p-value?

The zero difference has no sign and is dropped, so n = 9 and X = 8 positives (1 negative). Under H_0, X ~ Binomial(9, 0.5). P(X ≥ 8) = [C(9, 8) + C(9, 9)] ÷ 2^9 = (9 + 1) ÷ 512 = 0.0195. Two-sided p = 2 × 0.0195 = **0.039**.

2. Why is the sign test valid for a 5-point pain score when a test using the sizes of the differences may not be?

The sign test uses only the direction of each change (better or worse), which an ordinal score does record. A test that ranks the sizes of differences (Wilcoxon signed-rank) assumes that a change of 2 points is larger than a change of 1 point wherever it occurs on the scale, which is an interval-scale assumption that ordinal pain scores do not guarantee.

Checkpoint 10.2

1. Match each test to its parametric counterpart: Mann-Whitney U, Wilcoxon signed-rank, Kruskal-Wallis, Friedman.

Mann-Whitney U ↔ independent-samples t-test; Wilcoxon signed-rank ↔ paired t-test; Kruskal-Wallis ↔ one-way ANOVA; Friedman ↔ repeated-measures ANOVA.

2. Two groups of 5 are ranked together. Group 1 holds ranks 1, 2, 4, 5 and 7. Compute R_1, U_1 and U_2.

$R_1 = 1 + 2 + 4 + 5 + 7 = 19$. $U_1 = R_1 - n_1(n_1 + 1)/2 = 19 - 15 = \mathbf{4}$. $U_2 = n_1n_2 - U_1 = 25 - 4 = \mathbf{21}$. (Check: group 2 holds ranks 3, 6, 8, 9, 10, sum 36, and 36 − 15 = 21.)

3. A Mann-Whitney test is significant, but the two groups have almost identical medians. How can that happen, and how should you report it?

The Mann-Whitney test asks whether values in one group tend to be larger than in the other. It compares medians only if the two distributions have the same shape. If one group is much more spread out or differently skewed, the test can be significant with similar medians (Hart, 2001). Report it as a difference in distribution ("values in group A tend to be higher"), describe both medians and spreads (e.g. IQRs), and do not claim that the medians differ.

Checkpoint 10.3

1. A 2 × 3 table has 500 observations. How many degrees of freedom does the test of independence have, and why, in terms of free cells?

df = (2 − 1)(3 − 1) = **2**. With the row and column totals fixed, you can fill in only the first two cells of the first row freely. The third cell of that row is fixed by the row total, and every cell of the second row is fixed by the column totals.

2. In a 2 × 2 table, a = 20, b = 30, c = 10, d = 40. Compute the expected count for cell a and the Pearson X^2. Is it significant at 5%?

Row totals 50 and 50; column totals 30 and 70; N = 100. E_a = 50 × 30 ÷ 100 = **15**. $X^2 = N(ad - bc)^2 \div [(a + b)(c + d)(a + c)(b + d)] = 100 \times (800 - 300)^2 \div (50 \times 50 \times 30 \times 70) = 25{,}000{,}000 \div 5{,}250{,}000 = \mathbf{4.76}$. This exceeds 3.84 (the 5% point with 1 df); p = 0.029, so it is significant at 5%. (Check by cells: |O − E| = 5 in every cell; 25/15 + 25/35 + 25/15 + 25/35 = 4.76.)

3. "My 2 × 2 table has an expected count of 2.1, so I applied Yates' correction." What is wrong with this?

Yates' correction is a continuity correction for approximating discrete counts with the continuous χ^2 curve. It does not make the χ^2 approximation valid when expected counts are too small. With an expected count of 2.1, one cell in four (25%) is below 5, which breaks Cochran's convention for a 2 × 2 table. The right analysis is Fisher's exact test (§10.7).

Checkpoint 10.4

1. In a cohort, 30 of 100 smokers and 10 of 100 non-smokers develop a cough. Compute RR, OR, RD and NNH. Why do RR and OR differ?

$R_1 = 0.30$, $R_0 = 0.10$. RR = 0.30/0.10 = **3.0**. OR = (30 × 90)/(70 × 10) = 2,700/700 = **3.86**. RD = 0.30 − 0.10 = **0.20** (20 extra coughs per 100 smokers). NNH = 1/0.20 = **5**. The OR is further from 1 because the outcome is common: $OR = RR \times (1 - R_0)/(1 - R_1) = 3 \times 0.90/0.70 = 3.86$. The two agree only when the outcome is rare in both groups.

2. Why can a case-control study not estimate the risk ratio directly, although it can estimate the odds ratio?

The investigator chooses how many cases and controls to include, so the proportion of cases in each exposure row reflects that choice, not a risk in a population at risk. Sampling a fraction f of non-cases multiplies b and d by f, which changes a/(a + b) but leaves ad/bc unchanged. Also, the odds ratio of exposure

(cases vs controls) equals the odds ratio of disease (exposed vs unexposed), so the case-control OR estimates the same quantity a cohort would.

3. Using Levin's formula, find the PAF when 20% of the population is exposed and RR = 3.

PAF = $p_e(RR - 1) \div [1 + p_e(RR - 1)] = 0.20 \times 2 \div (1 + 0.20 \times 2) = 0.40 \div 1.40 =$ **0.286, or 28.6%**, assuming the association is causal and unconfounded.

Checkpoint 10.5

1. A 2 × 2 table has three expected counts below 5. Which test should you use, and why is Yates' correction not the answer?

Fisher's exact test. It computes the p-value exactly from the hypergeometric distribution, with the margins fixed, so it does not rely on the χ^2 approximation. Yates' correction only adjusts the approximation for the discreteness of counts; it does not repair an approximation that fails because expected counts are too small (the §10.7 pilot shows uncorrected p = 0.036, Yates p = 0.126, exact p = 0.091).

2. In a before–after survey of 150 households, 20 joined CBHI and 8 left; the rest did not change. Compute McNemar's X^2 and state your conclusion at the 5% level.

Discordant pairs: b = 20 (joined), c = 8 (left). $X^2 = (20 - 8)^2 \div (20 + 8) = 144 \div 28 =$ **5.14**, 1 df, p = 0.023. Membership rose significantly after the drive: more households joined than left. (The 122 unchanged households do not enter the test. The exact binomial version, the sign test on the 28 discordant pairs, gives p = 0.036, the same conclusion.)

3. In a matched case-control study, b = 30 and c = 12. Compute the matched OR and its 95% CI.

OR = b/c = 30/12 = **2.5**. SE(ln OR) = $\sqrt{(1/30 + 1/12)} = \sqrt{0.1167} = 0.342$. ln 2.5 = 0.916. CI: $\exp(0.916 \pm 1.96 \times 0.342) = \exp(0.247)$ to $\exp(1.586) =$ **1.28 to 4.88**.

Chapter 11: Correlation

Checkpoint 11.1

1. For 10 women, $S_{xx} = 250$, S_yy = 40 and $S_{xy} = 80$ (haemoglobin in g/dL against daily iron tablets taken over a month). Compute the covariance, r and the t statistic for H_0: ρ = 0. Is r significant at the 5% level?

1. Covariance = $S_{xy} \div (n - 1) = 80 \div 9 =$ **8.89**.
2. $r = S_{xy} \div \sqrt{(S_{xx} \times S_yy)} = 80 \div \sqrt{(250 \times 40)} = 80 \div \sqrt{10{,}000} = 80 \div 100 =$ **0.80**.
3. $t = r\sqrt{(n - 2)} \div \sqrt{(1 - r^2)} = 0.80 \times \sqrt{8} \div \sqrt{0.36} = 2.263 \div 0.60 =$ **3.77** on 8 df.
4. The two-sided 5% critical value of t on 8 df is 2.306. 3.77 > 2.306, so r is significant (p = 0.005). With only 10 women, the association says nothing yet about cause: women who take more tablets may differ in diet or adherence to other advice.

2. A study of 28 anganwadi workers reports r = 0.50 between years of service and a knowledge score. Compute the 95% confidence interval for ρ by Fisher's z.

1. $z = \tfrac{1}{2} \ln[(1 + 0.50) \div (1 - 0.50)] = \tfrac{1}{2} \ln 3 =$ **0.549**.
2. $SE(z) = 1 \div \sqrt{(n - 3)} = 1 \div \sqrt{25} =$ **0.200**.
3. Limits on the z scale: $0.549 \pm 1.96 \times 0.200 = 0.157$ to 0.941.
4. Back-transform with $r = (e^{2z} - 1) \div (e^{2z} + 1)$: **0.16 to 0.74**.

The interval is wide and asymmetric about 0.50: the data fit anything from a weak to a strong association.

3. A researcher correlates ranked examination scores with hours of study. Which coefficient(s) should she use, and why?

A rank coefficient: Spearman's ρ_s or Kendall's τ. Examination rank is ordinal (the gaps between ranks are not equal), so Pearson's r, which uses the actual distances from the mean, is not appropriate. Both rank coefficients use only order. Spearman's ρ_s is Pearson's r on the ranks; Kendall's τ is the proportion of concordant minus discordant pairs. Either is acceptable; report which one, since their values are on different scales.

Checkpoint 11.2

1. In the district example of §11.5 (more health workers, more recorded disease), name one explanation of each of the four kinds (chance, bias, confounding, reverse causation).

- **Chance:** with few districts, or after checking many indicators, a correlation can arise by accident; the p-value and CI address this.
- **Bias (information bias):** more health workers find and record more cases, so recorded disease rises with staffing even if true disease does not.
- **Confounding:** population size (or urbanisation) raises both the number of posts and the number of cases.
- **Reverse causation:** districts with a heavy disease burden are sent more health workers.

2. In Figure 11.4, how many paths join E and W, and how many of them are backdoor paths? A colleague adjusts for ANC visits and analyses only hospital deliveries. Which non-causal paths does she open? What should she adjust for instead?

1. Seven paths: E → W, E → A → W, E → P → W (causal); E ← S → W, E ← S → A → W (**two backdoor paths**); E → A ← S → W and E → H ← C → W (non-causal, each blocked by a collider).
2. Adjusting for A (a collider on E → A ← S → W) opens that path; since she has not adjusted for S, a non-causal association runs from E through A and S to W. It also cuts the causal path E → A → W (overadjustment). Restricting to hospital deliveries conditions on H and opens E → H ← C → W (collider/selection bias).
3. She should adjust for **S alone**, the minimal adjustment set, and study all births, not only hospital births.

3. Wealth is recorded only as "above or below the poverty line". After adjusting for it, the radio–nutrition association shrinks but does not vanish. Is what remains unmeasured or residual confounding? What would reduce it?

It is (at least partly) **residual confounding**: wealth was measured and adjusted for, but a two-category measure leaves large wealth differences within each category. Measuring wealth more finely (an asset index in quintiles, or as a continuous score) and adjusting for that would reduce it. Any confounder that was never recorded (for example, the mother's own interest in health information) would be unmeasured confounding, which better use of the existing data cannot remove.

Checkpoint 11.3

1. Two strata of a cohort give (a, b, c, d) = (10, 20, 5, 40) and (15, 10, 20, 30), with 75 people in each. Compute the stratum ORs, the crude OR and OR_MH. Did the stratifying variable confound the crude OR, and in which direction?

1. Stratum 1: OR = (10 × 40) ÷ (20 × 5) = 400 ÷ 100 = **4.00**. Stratum 2: OR = (15 × 30) ÷ (10 × 20) = 450 ÷ 200 = **2.25**.
2. Crude table: a = 25, b = 30, c = 25, d = 70. Crude OR = (25 × 70) ÷ (30 × 25) = 1,750 ÷ 750 = **2.33**.
3. OR_MH: numerator 10 × 40/75 + 15 × 30/75 = 5.333 + 6.000 = 11.333; denominator 20 × 5/75 + 10 × 20/75 = 1.333 + 2.667 = 4.000. OR_MH = 11.333 ÷ 4.000 = **2.83**.

4. The adjusted OR (2.83) is larger than the crude (2.33): the stratifying variable confounded the crude estimate towards the null (negative confounding). (MH $\chi^2 = 7.55$, $p = 0.006$.) Before pooling, check that 4.00 and 2.25 are compatible with one common OR; with strata of 75 the homogeneity test has little power.

2. In a study of tobacco chewing and oral leukoplakia, the RR is 1.2 among non-drinkers and 4.5 among drinkers; the homogeneity test gives $p = 0.002$. What should the authors report, and why is an MH RR inappropriate?

Drinking is an effect modifier: the RRs differ far more than chance would explain. The authors should report the two stratum-specific RRs (with CIs), and say that the association with tobacco chewing is much stronger among drinkers. An MH RR is a weighted average that assumes one common RR; here it would describe neither group and hide the main finding.

3. In the illustrative Simpson's paradox table of §11.7, apply PHC A's severity-specific failure risks to PHC B's mix of children (20 severe, 200 not severe). What overall failure risk would A have had? Compare it with B's crude risk and explain the paradox in one sentence.

1. Expected failures at A with B's mix: $20 \times 0.12 + 200 \times 0.02 = 2.4 + 4.0 = 6.4$.
2. Risk = $6.4 \div 220$ = **0.029 (2.9%)**, against B's actual $9 \div 220$ = **0.041 (4.1%)**.

With the same case mix, A does better than B; A's crude risk is higher only because two-thirds of its children were severely dehydrated, and severity, not the PHC, drives most failures. (This is direct standardisation, recall §7.4.2, with B's case mix as the standard.)

Chapter 12: Regression – Foundations

Checkpoint 12.1

1. For the clinic register, $S_{xy} = 928.6$, $S_{xx} = 6190.8$, $\bar{x} = 24.35$ months and $\bar{y} = 12.15$ kg. Compute b_1 and b_0. Then find the fitted weight and the residual for a 30-month-old child who weighs 12.0 kg.

$b_1 = S_{xy} \div S_{xx} = 928.6 \div 6190.8 = 0.150$ kg per month. $b_0 = \bar{y} - b_1\bar{x} = 12.15 - 0.150 \times 24.35 = 12.15 - 3.65 = 8.50$ kg. Fitted weight at 30 months: $8.50 + 0.150 \times 30 = 8.50 + 4.50 = 13.00$ kg. Residual = observed − fitted = $12.0 - 13.0 = -1.0$ kg: this child is 1 kg lighter than the model predicts for her age.

2. Children in a school are weighed, and the 20 thinnest are enrolled in a feeding programme. Three months later their mean BMI z-score has risen. Name one reason, other than the programme, why their mean would rise, and one design feature that would remove it.

Regression to the mean: the 20 were selected *because* their first measurement was extreme, and part of that extremeness was measurement error and short-term fluctuation (a recent illness, a measurement slip), which is not repeated at the second weighing. Their mean would move towards the population mean even with no programme. (Normal growth and seasonal change are further reasons.) A randomised control group chosen in the same way (thinnest children, half not fed) regresses by the same amount, so the programme's effect is the difference between groups. Selecting on the mean of two baseline measurements also reduces the problem.

3. Two variables have $r = 0.5$. A person is 2 SD above the mean of x. How many SDs above the mean of y is the predicted value?

In SD units the prediction is r times the predictor: $(\hat{y} - \bar{y})/s_y = r \times (x - \bar{x})/s_x = 0.5 \times 2 = 1.0$ SD above the mean of y.

Checkpoint 12.2

1. A colleague adds a fourth predictor to a model and R^2 rises from 0.41 to 0.43. They conclude the new predictor is clinically important. What is the flaw, and what should they check instead?

Adding any predictor, even random numbers, can only lower SSE, so R^2 can only rise; the rise says nothing about importance. They should check adjusted R^2 (which rises only if the new predictor's |t| exceeds 1, itself a low bar), and, above all, the new predictor's coefficient and 95% CI in clinical units: is the size of the association compatible with the data large enough to matter?

2. Explain, using the error term, why a fitted line almost never passes through every data point, even when the model is a good one.

The model is $Y = \beta_0 + \beta_1 X + \varepsilon$. The error term ε holds everything that affects the outcome besides X: other causes, individual variation and measurement error. The line describes the *average* outcome at each x; individuals scatter around it by their own ε. A good model gets the average right; it was never meant to predict every individual exactly.

3. A regression of haemoglobin (g/dL) on weeks of iron supplementation in 27 women gives $b_1 = 0.12$ g/dL per week, s = 0.9 g/dL and Sxx = 900 weeks^2. Compute SE(b_1), the t statistic and the 95% CI for β_1 ($t_{25} = 2.060$). Is the result compatible with no linear association?

$SE(b_1) = s \div \sqrt{Sxx} = 0.9 \div 30 = 0.03$ g/dL per week. $t = 0.12 \div 0.03 = 4.0$ on $27 - 2 = 25$ df. 95% CI = $0.12 \pm 2.060 \times 0.03 = 0.12 \pm 0.062 = 0.058$ to 0.182 g/dL per week. The interval excludes zero ($p < 0.001$, since $|t| = 4.0$ is far beyond 2.06), so the data are not compatible with a flat line; they are compatible with gains from about 0.06 to 0.18 g/dL per week.

Checkpoint 12.3

1. A residual-vs-fitted plot shows residuals fanning out into a cone as fitted values increase. Which assumption is violated, and what does it threaten: the coefficient, or something else?

Equal variance (homoscedasticity) is violated: heteroscedasticity. It threatens the standard errors, and so the CIs and p-values built from them. The coefficient itself can remain a reasonable estimate; its reported precision is what becomes unreliable.

2. Why does violating independence raise the risk of a false-positive result, rather than just making the estimate less precise?

Correlated observations (children in the same village, repeated visits by one child) carry less information than the same number of independent ones. The model treats each as fresh information, so it computes an SE that is too small. A too-small SE gives too-narrow CIs and too-small p-values, so a null association is declared significant more often than 5% of the time: the Type I error rate rises.

3. The clinic regression (SE of the slope 0.011 kg per month) is refitted on children who were in fact sampled in village clusters, with a design effect of 2.44 for the slope. Roughly what is the correct SE, and what happens to the width of the 95% CI?

DEFF multiplies the variance (recall §6.2.6), so the SE is multiplied by $\sqrt{DEFF} = \sqrt{2.44} = 1.56$. Correct SE $\approx 0.011 \times 1.56 = 0.017$ kg per month. The 95% CI is about 1.56 times as wide: roughly $0.150 \pm 2 \times 0.017 = 0.116$ to 0.184 kg per month instead of 0.129 to 0.171.

Checkpoint 12.4

1. Household wealth and household electricity access are strongly correlated. In a model containing both, each has a wide CI and neither is significant, though R^2 is high. What is the likely explanation, and how would you confirm it?

Multicollinearity. The two predictors share most of their variation, so the model cannot separate their individual contributions: each coefficient's SE is inflated, though together they explain the outcome well (hence the high R^2). Confirm with the variance inflation factor (VIF): large VIFs for both, and a large R^2_j when one is regressed on the other. A joint F-test of the two (the nested-model F-test taught later, in §12.4.5) would be significant even though neither t-test is.

2. A crude regression coefficient for a drug's effect is 0.90, and the coefficient adjusted for age is 0.35. What does this comparison show, and what does it not show?

It shows that age accounts for a large part of the crude association, which is *consistent with* age confounding the drug–outcome relationship (for example, if older patients both received the drug more often and had higher outcome values). It does not by itself prove confounding: the shift could partly reflect chance or model specification, and whether age is a confounder (associated with drug use, affecting the outcome, not on the causal path) must be argued from subject-matter knowledge before fitting. The adjusted 0.35 also remains open to confounding by unmeasured variables.

3. In a model with an interaction, b_1 (supplement) = 0.30 kg and b_3 (supplement × malnourished) = 0.50 kg. What is the supplement's estimated effect in malnourished children? Separately: regressing one predictor on the others gives $R^2_j = 0.75$. What is its VIF, and by what factor is its SE inflated?

Effect in malnourished children = $b_1 + b_3 = 0.30 + 0.50 = 0.80$ kg (0.30 kg applies only to well-nourished children, the reference group). VIF = $1 \div (1 - 0.75) = 1 \div 0.25 = 4$. The variance of that coefficient is 4 times what it would be without the correlation, so the SE is $\sqrt{4} = 2$ times larger.

Checkpoint 12.5

1. A study reports "children with maternal BMI ≥ 25 had 0.4 kg higher birthweight (p = 0.04)", after the authors tried cut-offs of 22, 23, 24, 25 and 26. Give two reasons this analysis is weak, and say how you would reanalyse it.

1. Dichotomising a continuous predictor discards information and power, and treats a mother at BMI 25.1 like one at 35. (ii) The cut-off was chosen as the most significant of five tries, a multiple-comparisons problem that biases the estimate away from zero and inflates the false-positive rate; p = 0.04 after five looks is weak evidence. Reanalyse with maternal BMI continuous, checking the shape with a restricted cubic spline (3–4 knots at fixed centiles) and testing non-linearity by comparing spline and straight-line models; present the fitted curve with its CI band. Categories fixed in advance may be added as a description.

2. Adding three dummy variables for four categories to a trend model lowers SSE from 52.0 to 50.5. The category model has 120 residual df. Compute F for departure from a linear trend and say whether it suggests the trend model is inadequate (the 5% critical value of F on 3 and 120 df is 2.68).

$F = [(52.0 - 50.5) \div 3] \div (50.5 \div 120) = 0.50 \div 0.421 = 1.19$. This is below 2.68 ($p > 0.05$), so there is no evidence of departure from a straight-line trend; the one-coefficient trend model is the better summary. (No evidence of departure is not proof that the trend is exactly linear.)

Checkpoint 12.6

1. One unusual point has an extreme predictor value but lies on the regression line; another has an ordinary predictor value but a very unexpected outcome. Which has high leverage, and which a large residual? Which is more likely to be influential?

The point with the extreme predictor value has high leverage (leverage depends only on x). The point with the ordinary x but unexpected outcome has a large residual. Neither is strongly influential on its own: the high-leverage point lies on the line, so removing it hardly changes the fit, and the large-residual point has little leverage to tilt the line (it mainly shifts the intercept slightly). Influence needs both at once, which Cook's distance measures.

2. Why is an influential point a prompt to investigate rather than a reason to delete?

It may be a genuine observation showing real variation: deleting it discards information and makes the result look more certain than it is. Or it may be a data-entry error or a person outside the study population, who should be corrected or excluded. Only checking the record can tell which. Either way, report the fit with and without the point so readers can see how much the conclusion depends on it.

3. In a regression of 50 adults' systolic BP on age, $\bar{x}$ = 45 years and Sxx = 8,000 years^2. Compute the leverage of a 25-year-old and of a 75-year-old, and compare each with the average leverage.

$h = 1/n + (x - \bar{x})^2/Sxx$. 25-year-old: $1/50 + (-20)^2/8000 = 0.020 + 400/8000 = 0.020 + 0.050 = 0.070$. 75-year-old: $0.020 + 30^2/8000 = 0.020 + 900/8000 = 0.020 + 0.1125 = 0.1325$. The average leverage is $(k + 1)/n = 2/50 = 0.04$, so the 25-year-old has 1.75 times and the 75-year-old 3.3 times the average. The 75-year-old could move the line most if their BP is far from its prediction.

Checkpoint 12.7

1. In a two-predictor model, the adjusted coefficient for maternal education (years) on birthweight is 25 g per year, the coefficient for maternal height is 12 g per cm, and the slope of maternal height on education is 0.5 cm per year. What was the crude coefficient for education?

Crude slope = adjusted slope + (height coefficient) × (slope of height on education) = 25 + 12 × 0.5 = 25 + 6 = 31 g per year. About 6 g of the crude 31 g per year reflected the taller stature of more-educated mothers.

2. A colleague says "age is a confounder in every child-growth study". Using the confounder and precision-covariate definitions, explain when age confounds a wealth–weight association and when it is only a precision covariate.

Age affects weight in every such study, and wealth cannot cause a child's age, so age is never a mediator. But it confounds the wealth–weight association only if it is also associated with wealth in the population studied, for example if wealthier families bring older children to the clinic. Then the crude comparison mixes the wealth difference with an age difference, and adjustment moves the coefficient (as in §12.6: 0.90 → 0.62, with r = 0.16 between age and wealth). If age is unrelated to wealth (say, a sample with the same age distribution in every wealth group), it is only a precision covariate: adjusting leaves the wealth coefficient unchanged but shrinks its SE, because it removes age-driven variation from the residuals. Including it is worthwhile in both cases, but for different reasons.

Chapter 13: Generalised Linear Models

Checkpoint 13.1

1. A colleague fits a logistic regression and finds one predictor has an OR of 45,000 with an enormous standard error. What is the most likely explanation? Is it an error to fix with a different technique?

Almost certainly separation: the predictor (or a combination of predictors) perfectly or almost perfectly predicts the outcome, so the ordinary maximum likelihood estimate does not exist and the coefficient drifts towards infinity. The data are not wrong, and the 45,000 must not be reported. First look at why the predictor separates the outcome so cleanly (a coding error, a predictor that is really a consequence of the outcome, very small numbers). If the pattern is genuine, valid finite estimates can be obtained with Firth's penalised likelihood (Heinze & Schemper, 2002) or exact logistic regression.

2. Explain what "linearity in the logit" means, and why checking it on a plot of raw probability against the predictor would be the wrong approach.

The model assumes that each continuous predictor has a straight-line relation with the log-odds of the outcome: $\text{logit}(p) = b_0 + b_1x$. On the probability scale the correctly specified model is S-shaped, so a curved probability plot is expected and says nothing about whether the assumption holds. Check on the logit scale instead: add a quadratic term (or a spline) and see whether the fit improves, or group the predictor and see whether the log-odds of the groups lie roughly on a line.

3. Of 20 children screened, 5 are anaemic. Write the likelihood for the prevalence p, find the MLE, and give its logit. What probability corresponds to a logit of 1.2?

$L(p) = p^5(1 - p)^{15}$. $\ln L = 5 \ln p + 15 \ln(1 - p)$; its slope $5/p - 15/(1 - p) = 0$ gives $5(1 - p) = 15p$, so $\hat{p} = 5/20 =$ **0.25**. $\text{logit}(0.25) = \ln(0.25/0.75) = \ln(1/3) =$ **−1.099**. Logit 1.2 → p = 1 ÷ (1 + e^(−1.2)) = 1 ÷ (1 + 0.301) = **0.77**.

Checkpoint 13.2

1. A logistic regression reports an OR of 1.08 per year of age. A colleague says this means a 30-year age gap corresponds to 30 × 1.08 = 32.4-fold higher odds. What is wrong, and what is the correct figure?

A per-unit OR multiplies once for each unit; it does not add. The coefficient is additive on the log-odds scale (30 × ln 1.08 = 30 × 0.0770 = 2.309), so the OR for 30 years is e^2.309 = 1.08^{30} = **10.06**, about a 10-fold increase in odds, not 32.4.

2. Two logistic models on the same data report the same OR for a predictor, but one discriminates well between cases and non-cases and the other does not. How is this possible?

Discrimination depends on how well *all* the predictors together separate people who had the event from those who did not, through the spread of the predicted probabilities. A single OR describes one predictor's association. One model may contain other strong predictors that spread the predicted probabilities widely (good discrimination), while the other contains only weak predictors, so everyone gets a similar predicted probability (poor discrimination), even though the shared predictor's OR is the same.

3. In a cohort, the 1-year risk of TB is 1% in adults with neither diabetes nor undernutrition, 1.5% with undernutrition only, 2% with diabetes only and 4% with both. Compute the multiplicative interaction ratio and the RERI, and say what each means.

RR_{10} (undernutrition only) = 1.5/1 = 1.5; RR_{01} (diabetes only) = 2/1 = 2; RR_{11} = 4/1 = 4. Multiplicative: $RR_{11} \div (RR_{10} \times RR_{01}) = 4 \div 3 =$ **1.33**. The joint risk ratio is a third larger than the product of the separate ones:

positive interaction on the multiplicative scale. Additive: RERI = 4 − 1.5 − 2 + 1 = **1.5**. The excess risks would add to 0.5 + 1 = 1.5 percentage points; the observed joint excess is 3 points. Positive additive interaction: if these associations are causal, correcting undernutrition prevents more TB per person reached among diabetics (4% → 2%, 2 cases per 100) than among non-diabetics (1.5% → 1%, 0.5 per 100).

Checkpoint 13.3

1. A multinomial model of cause of under-five death uses "other" as the reference. How would the reported results change if "pneumonia" were the reference, and what would stay the same?

Every set of coefficients would now compare a category with pneumonia (diarrhoea vs pneumonia, malnutrition vs pneumonia, other vs pneumonia) instead of with "other", so all the reported ORs change in value and meaning. The model's fit, its predicted probabilities for each category, and its likelihood are unchanged: the reference is a reporting choice, and any comparison can be recovered from either parameterisation (e.g. diarrhoea vs pneumonia = (diarrhoea vs other) ÷ (pneumonia vs other)). State the reference.

2. Why is ordinal logistic regression a mistake for an unordered outcome such as blood group?

The ordinal model estimates cumulative odds across cut-points in a fixed order (for example A < B < AB < O) and forces one OR per predictor across all cut-points. Blood groups have no order, so the cut-points, the cumulative odds and the single OR are all artefacts of an arbitrary ordering. A different arbitrary order would give different results. A multinomial model is the right one.

3. In a 1:1 matched case-control study of 100 pairs, 30 pairs have only the case exposed and 12 have only the control exposed. Compute the matched OR and its 95% CI. What would an unconditional logistic model with 99 pair dummies report?

OR = b/c = 30/12 = **2.5**. SE(ln OR) = $\sqrt{(1/30 + 1/12)} = \sqrt{0.1167} = 0.342$. 95% CI = exp(ln 2.5 ± 1.96 × 0.342) = exp(0.916 ± 0.670) = **1.28 to 4.88**. Unconditional logistic regression with a dummy for each pair suffers from sparse data; for 1:1 pairs it gives the square of the conditional OR (Pearce, 2016): 2.5^2 = **6.25**, a serious overestimate.

Checkpoint 13.4

1. A Poisson regression of malaria cases on "near standing water", with log(village population) as an offset, gives exp(b) = 2.30. Is this a rate ratio or a count ratio? What would exp(b) be without the offset in these data, and why does it differ?

A rate ratio (incidence rate ratio): with the offset, the model is for log(cases/population). Villages near water have 2.30 times the malaria rate of the others (20.1 vs 8.74 per 1,000). Without the offset, exp(b) is the ratio of mean cases per village, 38.2/23.6 = **1.62**. It is smaller because the villages near water have fewer people (9,500 vs 13,500 in total), so their counts understate their rate.

2. A Poisson model shows clear overdispersion. What does switching to negative binomial regression change, and what does it leave unchanged?

It changes the variance assumption from Var = μ to Var = $\mu + \alpha\mu^2$, so the model is fitted against a different likelihood. The coefficient estimates can shift somewhat (IRR 2.30 → 2.47 in the malaria data), and the SEs and CIs widen to reflect the extra variation (1.83–2.89 → 1.70–3.60), making tests more honest. The meaning of exp(b) is unchanged: still a rate ratio with an offset, a count ratio without.

3. A Poisson model of monthly diarrhoea episodes has Pearson $\chi^2 = 40$ on 10 residual df. Compute the dispersion statistic. If an NB fit gives $\alpha = 0.1$, what variance does it imply for a child whose expected count is 2 episodes? For an expected count of 20?

Dispersion statistic = 40/10 = **4**, well above 1: the counts vary about four times as much as the Poisson model allows (a rough check; confirm with the likelihood-ratio test of $\alpha = 0$). NB variance = $\mu + \alpha\mu^2$. $\mu = 2$: $2 + 0.1 \times 4 =$ **2.4** (close to Poisson). $\mu = 20$: $20 + 0.1 \times 400 =$ **60**, three times the Poisson variance. Overdispersion under the NB model grows with the mean.

Checkpoint 13.5

1. In a cohort, 40% of unexposed people develop the outcome, and a crude OR is 2.0. Convert it to an RR. What would the RR be if the unexposed risk were 2%?

RR = OR ÷ $(1 - p_0 + p_0 \times \text{OR})$. $p_0 = 0.40$: RR = $2.0 \div (1 - 0.40 + 0.80) = 2.0 \div 1.40 =$ **1.43**. $p_0 = 0.02$: RR = $2.0 \div (1 - 0.02 + 0.04) = 2.0 \div 1.02 =$ **1.96**. With a common outcome the OR of 2 corresponds to a much smaller RR; with a rare outcome the two almost coincide.

2. Why does a Poisson model with a log link give valid prevalence ratios for a 0/1 outcome, and why must its SEs be replaced by robust ones?

The coefficients depend only on the model for the mean, $\log(p) = b_0 + b_1x_1 + \ldots$, which is the prevalence-ratio model; the Poisson likelihood is used only as a working device to estimate it, and it gives consistent estimates when the mean model is right (Zou, 2004). But the Poisson model assumes Var(Y) = p, whereas a 0/1 outcome has Var(Y) = $p(1 - p)$, which is smaller. The model-based SEs are therefore too large (by about $1/\sqrt{(1 - p)}$; for the CBHI prevalence of 30%, 0.0816 against the correct 0.0683). The robust (sandwich) SE uses the observed residuals instead of the assumed variance and corrects this.

3. A colleague fits a log-binomial model and gets "algorithm did not converge". Name two alternatives that still give an adjusted prevalence ratio.

1. Modified Poisson regression: a Poisson model with a robust (sandwich) SE (Zou, 2004). (ii) Marginal standardisation from a logistic model: predict every person's probability as if exposed and as if unexposed, average each, and take the ratio. (Supplying starting values to the log-binomial model sometimes helps but, as with the CBHI data, may still fail.)

Chapter 14: Survival Analysis

Checkpoint 14.1

1. In a cohort, S(t) just before month 12 is 0.80. At month 12, 70 members are at risk and 3 drop out. What is S(12)? Is the hazard in that month 3/70 or 3/100 of the original cohort, and why?

S(12) = $0.80 \times (1 - 3/70) = 0.80 \times 67/70 = 0.80 \times 0.957 =$ **0.766**. The month's hazard is estimated as 3/70 = 0.043 per member-month, because only the 70 still enrolled and under observation could drop out that month. Members who had already dropped out or been censored are no longer in the at-risk set, so dividing by the original 100 would understate the rate.

2. In a two-year follow-up of TB patients, the Kaplan–Meier curve ends at 0.64. A colleague reports "median survival 24 months". What should be reported instead?

The curve never falls to 0.5, so the median is **not reached**. Report "median NR" and give survival at stated times, e.g. "64% surviving at 24 months" (with its CI). Twenty-four months is simply the end of follow-up, not a median.

3. Why is a member who moves away at month 5 not simply deleted from a Kaplan–Meier analysis, and what assumption makes it safe to treat them as censored?

They were known to be enrolled and event-free for 5 months, and that is real information: they count in the at-risk set, in every denominator, up to month 5. Deleting them would discard it and bias the sample towards those with complete follow-up. Treating them as censored is valid under **non-informative censoring**: households lost to follow-up are assumed as likely to drop out later as similar households that remained under observation. (If moving away itself ended membership, it would be a competing event, §14.8.)

Checkpoint 14.2

1. A Kaplan–Meier curve for Group A lies above Group B's throughout follow-up, and the log-rank test gives p = 0.03. A colleague concludes: "Group A has a lower hazard at every single time point, guaranteed." What is the overreach?

Two overreaches. First, S(t) is *cumulative*: a curve that stays higher reflects the sum of all hazard up to t. Group A can have a higher hazard than Group B during some periods and still keep a higher S(t), because of an advantage built up earlier; hazards can cross while survival curves do not. Second, the log-rank test is an overall test across the whole follow-up. p = 0.03 is evidence that the survival experiences differ overall, not a guarantee of anything at each time point, and nothing in statistics is "guaranteed".

2. A two-group log-rank analysis gives O = 30, E = 24 in Group A and O = 20, E = 26 in Group B. Compute $\Sigma(O - E)^2/E$, its df and approximate p-value.

$\chi^2 = (30 - 24)^2/24 + (20 - 26)^2/26 = 36/24 + 36/26 = 1.500 + 1.385 =$ **2.88**, df = 2 − 1 = 1. The 5% critical value for 1 df is 3.84, so p > 0.05; exactly, **p ≈ 0.09**. No convincing evidence that survival differs, though Group A had 6 more events than expected.

3. Why can the log-rank test not tell you whether the intervention's association with dropout remains after adjusting for household income?

The log-rank test compares groups defined by one categorical variable. It has no way to include a continuous covariate such as income, and it gives no effect size. Cox regression (§14.6) includes several predictors at once and gives an adjusted hazard ratio with a CI.

Checkpoint 14.3

1. A Cox model reports HR = 0.6 for a treatment, but the two groups' log(−log S) curves clearly diverge over log time, and the Schoenfeld residual test gives p = 0.002 for treatment. What does this mean for the single HR of 0.6?

Proportional hazards is violated: the treatment's hazard ratio changes over follow-up. HR = 0.6 is a weighted average of different effects at different times (for example, strong early and weak later) and should not be read as applying throughout. Report time-specific HRs (split follow-up into periods), let the treatment's effect vary with time (a treatment × time interaction), or, if treatment is only a covariate to be controlled for, stratify on it (§14.6.3 shows how).

2. A Cox coefficient is b = −0.35 with SE 0.12. Give the HR and its 95% CI, and state the result in hazard language.

HR = e^(−0.35) = **0.70**. CI = exp(−0.35 ± 1.96 × 0.12) = exp(−0.35 ± 0.235) = exp(−0.585) to exp(−0.115) = **0.56 to 0.89**. The group (or one-unit increase) has a 30% lower hazard of the event at any time during follow-

up, among those still event-free, with the data compatible with an 11% to 44% lower hazard. It is not "30% lower risk".

3. Explain why proportional hazards concerns the *ratio* of hazards between groups, not the hazard level itself.

Cox's model is $h(t) = h_0(t) \times \exp(bx)$. The baseline hazard $h_0(t)$ may rise or fall freely over time (dropout climbing as a scheme ages); that is not a violation. What must stay constant is the multiplier $\exp(b)$, the ratio between the groups' hazards at each moment, because the model estimates a single b for the whole follow-up.

Checkpoint 14.4

1. A cohort study reports that TB patients who "ever received" a nutrition kit had 20 deaths in 1,000 person-months, against 30 deaths in 1,000 person-months in those who never did. Kits were given at the end of month 2, and 200 of the kit group's person-months fell before the kit. Recompute the rates with kit receipt as a time-varying exposure, and the rate ratio before and after.

Naive: kit 20/1,000 = 2.0 per 100 person-months; no kit 30/1,000 = 3.0 per 100; rate ratio 2.0/3.0 = **0.67**.

The 200 pre-kit person-months are immortal time: a patient had to survive to month 2 to receive the kit, so no deaths could fall in them. They are unexposed time. Corrected: kit 20 deaths in 1,000 − 200 = 800 person-months = **2.5** per 100; no kit 30 deaths in 1,000 + 200 = 1,200 person-months = **2.5** per 100; rate ratio **1.00**. The apparent 33% reduction was entirely immortal-time bias.

2. In the CBHI cohort, you suspect the effect of distance changes over follow-up and you only need to adjust for it. Name a way to do this and what you give up.

Stratify on distance (for example, near vs far, split at the median): a stratified Cox model. Each stratum gets its own baseline hazard, so PH need not hold for distance. You give up an HR for distance, and a categorised distance adjusts less finely than the continuous variable (the intervention HR was 0.61 with a two-level stratum against 0.58 with continuous distance). Alternatively, keep distance in the model with a time-varying coefficient (a distance × time term), which does estimate how its effect changes.

Checkpoint 14.5

1. Five patients: event of interest at week 1, competing event at week 2, event of interest at week 3, censored at week 4, event of interest at week 5. Compute the CIF of the event of interest at week 3, and 1 − KM at week 3 with the competing event censored.

CIF: week 1, 5 at risk, $S(1_-) = 1$: increment $1 \times 1/5 = 0.200$; S becomes 0.800. Week 2, competing event, 4 at risk: $S = 0.800 \times 3/4 = 0.600$. Week 3, 3 at risk: increment $0.600 \times 1/3 = 0.200$. **CIF(3) = 0.400.**

1 − KM (competing event censored): week 1, 1 − 1/5 = 0.8; week 2, no step (censored); week 3, 3 at risk, × 2/3. KM = 0.8 × 2/3 = 0.533, so **1 − KM = 0.467**. It overstates the probability of the event by 0.067, because it treats the patient with the competing event as though they could still have the event of interest.

2. A district study of elderly patients after hip fracture asks (a) whether diabetes raises the rate of re-fracture among those still alive, and (b) what proportion of patients will re-fracture within two years, for service planning. Death is common. Which model suits each question?

1. An aetiological question about the rate among those still at risk: the **cause-specific hazard** model (a Cox model for re-fracture with deaths censored). (b) A question about the actual proportion who re-fracture in a population where many die

first: the **cumulative incidence function** (Aalen–Johansen), and, with covariates, the **Fine–Gray** subdistribution hazard model. 1 − KM would overstate (b) because deaths are common.

Chapter 15: Advanced Topics for the Community Medicine Researcher

Checkpoint 15.1

1. A colleague says, "Model B has lower AIC than model A, so model B is the true model." What is the overstatement?

AIC ranks only the candidates actually compared. A lower AIC means B has the better fit-versus-complexity trade-off of the two, judged by $-2 \ln L + 2k$ on these data. It says nothing about whether B is correct, causally meaningful, or even a good model: both candidates could be poor, and a model not considered could be better. With a small AIC difference (as in §15.1.2, 557.16 against 558.25) the data barely prefer one over the other, and BIC may disagree.

2. Why does adding a purely random predictor never lower a model's maximised likelihood? Give the argument, not just the conclusion.

The larger model contains the smaller one: setting the new coefficient to 0 reproduces the smaller model exactly, with the same likelihood. Maximum likelihood for the larger model searches over all values of that coefficient, including 0, so its maximum can only be equal to or higher than the smaller model's. In a real sample the best value of the new coefficient is almost never exactly 0 (chance alignment with noise), so the likelihood rises slightly. This is why raw fit cannot choose model size and a penalty or a test is needed.

3. A logistic model on 400 children has $\ln L = -210.4$ with 4 parameters. Adding a four-category feeding-practice variable gives $\ln L = -205.9$. Compute the LR statistic, its df, the change in AIC and the change in BIC. Which criteria favour the larger model?

1. $LR = -2 \times (-210.4 - (-205.9)) = -2 \times (-4.5) = \mathbf{9.0}$.
2. A four-category variable adds 3 parameters, so df = **3**. $p = P(\chi^2_3 > 9.0) = \mathbf{0.029}$.
3. ΔAIC = (change in $-2 \ln L$) $+ 2 \times 3 = -9.0 + 6 = \mathbf{-3.0}$: AIC falls, so AIC favours the larger model.
4. $\Delta BIC = -9.0 + 3 \times \ln 400 = -9.0 + 3 \times 5.99 = -9.0 + 17.97 = \mathbf{+8.97}$: BIC rises, so BIC favours the smaller model.
5. The LRT ($p = 0.029$) and AIC favour adding feeding practice; BIC, with its heavier penalty, does not.

Checkpoint 15.2

1. A thesis protocol says, "Variables with $p < 0.2$ in univariable analysis will be entered into a multivariable model, and backward elimination will give the final model of risk factors for low birthweight." Give two problems, one about each step.

Univariable screening: a confounded risk factor can show no crude association and appear only after adjustment, so the screen can wrongly discard important variables (Sun, Shook & Kay, 1996). Backward elimination: the final model depends on the sample and the criterion, is unstable, and the p-values and CIs of the survivors ignore the selection, so they are too optimistic (Heinze et al., 2018). A better protocol states the aim. For the effect of one exposure, choose the adjustment set from a DAG; for a list of predictors, pre-specify candidates and validate the model.

2. For a model predicting which pregnant women will miss their third antenatal visit, is it acceptable to include a variable that is not a cause of missing the visit? Would the same variable be acceptable in a model estimating the effect of ASHA home visits on attendance?

In a predictive model, yes: any variable available at the time of prediction that improves accuracy in new data is acceptable, whether or not it is a cause. In an explanatory model of ASHA visits, not automatically: variables enter because the DAG shows they block backdoor paths. A variable that is a consequence of ASHA visits (a mediator) or a collider must be left out, and one unrelated to the question adds nothing.

3. In the CBHI bootstrap of §15.1.3, household size entered 15.5% of the final AIC-selected models. Out of 200 bootstrap samples, how many is that, and what does it tell you about a single stepwise run?

0.155 × 200 = **31** samples. A variable with no detectable association with membership was selected in about one run in six, so a variable's presence in a single stepwise model is weak evidence that it matters. Reporting inclusion frequencies shows how much the "final model" depends on the particular sample.

Checkpoint 15.3

1. A survey samples 800 households from 40 villages (20 households per village) with an ICC of 0.10. Compute the DEFF, the effective sample size and the factor by which standard errors grow.

1. DEFF = 1 + (20 − 1) × 0.10 = 1 + 1.9 = **2.9**.
2. Effective sample size = 800 ÷ 2.9 ≈ **276** households.
3. DEFT = √2.9 = **1.70**: standard errors are about 1.7 times their SRS values (not 2.9 times; DEFF acts on the variance).

2. One-way ANOVA of a 0/1 outcome across 30 anganwadi centres, 10 children each, gives MSB = 0.40 and MSW = 0.20. Estimate the ICC.

1. σ^2_b = (MSB − MSW) ÷ m = (0.40 − 0.20) ÷ 10 = 0.020.
2. ICC = 0.020 ÷ (0.020 + 0.20) = **0.091**.
3. Check: (0.40 − 0.20) ÷ [0.40 + 9 × 0.20] = 0.20 ÷ 2.20 = 0.091. (The implied DEFF is 1 + 9 × 0.091 = 1.82.)

3. A colleague reports only the inflated variance from a DEFF and stops there. What is missing from the analysis of the exposure–outcome association?

DEFF describes how much precision clustering costs; it is a planning and variance concept. It does not produce an adjusted estimate of an association or a valid test of it. That needs an analysis that handles the clustering: a mixed-effects model (cluster-specific effects), GEE with cluster-robust SEs (population-averaged effects), or a design-based survey analysis with PSUs, strata and weights.

Checkpoint 15.4

1. A programme manager wants to know which villages have unusually low baseline CBHI membership so that she can target them. Would a mixed-effects model or GEE better suit this goal, and why?

A mixed-effects model. It estimates each village's random intercept, shrunk towards the average so that small villages are not over-interpreted, and so answers "how does this village differ from the rest?". GEE treats within-village correlation as a nuisance and estimates a population-averaged effect; it gives no village-specific estimates.

2. Explain why "fixed" and "random" effects are not statements about whether something was chosen at random.

The labels describe how the cluster-specific values are modelled. A fixed effect is estimated as its own separate parameter; a random effect treats the cluster values as draws from a shared distribution, which

lets the model borrow strength across clusters. The same 20 villages could be modelled either way, however they were selected.

3. A cluster trial randomises 8 PHCs. GEE with exchangeable correlation gives a log OR of −0.40 for the intervention, with a robust SE of 0.10; the model-based SE is 0.16. Compute the 95% CI for the OR with each SE. Which SE should worry you, and what would you do?

1. OR = exp(−0.40) = 0.67.
2. Robust SE: exp(−0.40 ± 1.96 × 0.10) = exp(−0.596) to exp(−0.204) = **0.55 to 0.82**.
3. Model-based SE: exp(−0.40 ± 1.96 × 0.16) = exp(−0.714) to exp(−0.086) = **0.49 to 0.92**.

The robust SE. With only 8 clusters the sandwich estimator is imprecise and tends to be too small, so tests based on it reject too often (Mancl & DeRouen, 2001). A robust SE smaller than the model-based one is a warning, not extra precision. Use a small-sample correction (the Mancl–DeRouen bias-corrected estimator; in their simulations the jackknife was no better than the robust SE), use t-based inference with few degrees of freedom, or analyse at the cluster level, and report the more conservative result.

Checkpoint 15.5

1. A block survey samples 100 women from each of two strata: a town of 5,000 women and rural areas with 45,000. Compute each stratum's weight. If 30 town women and 50 rural women are anaemic, compute the unweighted and weighted prevalence.

1. Town: $\pi = 100/5{,}000 = 0.02$, weight = **50**. Rural: $\pi = 100/45{,}000$, weight = **450**.
2. Unweighted prevalence = (30 + 50) ÷ 200 = **40%**.
3. Weighted: $\Sigma wy = 50 \times 30 + 450 \times 50 = 1{,}500 + 22{,}500 = 24{,}000$; $\Sigma w = 50 \times 100 + 450 \times 100 = 50{,}000$. Prevalence = 24,000 ÷ 50,000 = **48%**.
4. The unweighted figure gives the over-sampled town (30% anaemic) half the influence although it holds only 10% of the women.

2. An NFHS analysis of 1,200 children in 60 PSUs across 6 strata reports a naive SE of 1.2 percentage points and a design-based SE of 1.8. What is the DEFF, and how many degrees of freedom does the design have?

DEFF = $(1.8 \div 1.2)^2 = 1.5^2 =$ **2.25**. Design degrees of freedom = PSUs − strata = 60 − 6 = **54** (not 1,199).

3. A student wants the prevalence of anaemia among adolescent girls and deletes all other women from the dataset before declaring the survey design. What should she do instead, and why?

Declare the design on the full dataset and then restrict it with the software's subpopulation (domain) option (in R, `subset(design, age < 20)`). Deleting rows first can drop PSUs or whole strata that contain no adolescents, which changes the design's degrees of freedom and, in some designs, the SE (and can stop the calculation when a stratum is left with one PSU). `subset()` keeps "the original design information about number of clusters, strata" (Lumley, survey package documentation).

Checkpoint 15.6

1. After propensity-score matching, a covariate's p-value comparing the two groups is 0.62. A colleague concludes that balance was achieved. Why is this premature, and what should be reported instead?

A p-value depends on sample size, and matching usually shrinks the sample, so a covariate that is still clearly imbalanced can give a large p-value. Balance is a property of the sample, not a hypothesis about a population. Report the standardised mean difference for every covariate, before and after matching; a common convention treats $|\text{SMD}| < 0.1$ as negligible (Austin, 2009).

2. Before matching, mean age of mothers is 27.0 years (SD 4.0) among members and 25.0 years (SD 5.0) among non-members. Compute the SMD. Is the imbalance negligible by the usual convention?

SMD = $(27.0 - 25.0) \div \sqrt{[(4.0^2 + 5.0^2)/2]} = 2.0 \div \sqrt{20.5} = 2.0 \div 4.53 =$ **0.44**. Far above 0.1: a substantial imbalance.

3. A team matches 1:1 on the propensity score and reports an odds ratio from ordinary (unpaired) logistic regression on the matched sample. What must the team justify, and what are the two positions in the literature?

They must justify ignoring the pairs. Austin's (2011b) simulations found that paired methods (for a binary outcome, McNemar's test and the matched-pair OR b/c, or conditional logistic regression; or robust SEs clustered on the pair) gave type I error and CI coverage closer to nominal. Stuart (2010) records a debate and gives reasons why accounting for the pairs may not be necessary, since matching on the score does not create pairs that are alike on every covariate. The team should state which estimand (ATT among the matched) and which analysis it chose, and report the other as a sensitivity check.

Checkpoint 15.7

1. 15% of income values are missing, mostly among the wealthiest respondents, who are reluctant to disclose income. Will standard MAR-based multiple imputation remove the bias? What should the analyst do?

Not necessarily. If high income itself causes non-response, the data are MNAR: even within groups defined by the observed variables, the missing incomes are higher than the observed ones. Standard MI assumes MAR and would impute values that are too low. The analyst should include every variable that predicts income or its missingness (to make MAR as plausible as possible), then run a sensitivity analysis that imputes incomes shifted upwards by stated amounts and report whether the conclusion changes.

2. Five imputed datasets give estimates 2.0, 2.4, 2.2, 1.8 and 2.1 for a regression coefficient, each with SE 0.30. Apply Rubin's rules to find the pooled estimate and its SE.

1. $\bar{Q} = (2.0 + 2.4 + 2.2 + 1.8 + 2.1) \div 5 = 10.5 \div 5 =$ **2.10**.
2. $\bar{U} = 0.30^2 = 0.09$ (all five SEs equal).
3. Deviations from 2.10: −0.1, 0.3, 0.1, −0.3, 0.0; squares sum to 0.20. $B = 0.20 \div (5 - 1) = 0.05$.
4. $T = 0.09 + (1 + 1/5) \times 0.05 = 0.09 + 0.06 = 0.15$. SE = $\sqrt{0.15} =$ **0.387**, compared with 0.30 in any single completed dataset.

3. Why might the imputation model for missing income include "survey enumerator ID", a variable that will never appear in the analysis model?

The imputation model's job is to make MAR plausible and to predict the missing values well. Non-response often depends on who asked the question, so enumerator ID can explain the missingness; including it moves the mechanism towards MAR given the observed data. It has no role in the research question, so it stays out of the analysis model. The two models serve different purposes (Sterne et al., 2009).

Checkpoint 15.8

1. An ROC curve for a new biomarker lies almost on the diagonal. A second analyst reports an AUC of 0.21 for another marker. What does each result mean, and what should the second analyst check first?

The diagonal is where sensitivity equals the false-positive proportion (1 − specificity) at every cut-off: the test is no better than chance, AUC ≈ 0.5. An AUC of 0.21 is far from chance: the marker separates the groups well, but in the opposite direction to the one specified. The analyst should first check the direc-

tion (for example, whether *low* values indicate disease, as with MUAC). Reversing it gives AUC = 1 − 0.21 = 0.79.

2. A study has 8 diseased and 5 non-diseased people. Of the 40 diseased–non-diseased pairs, the diseased person has the higher biomarker value in 30, the values are tied in 4, and the non-diseased person is higher in 6. Compute the AUC.

AUC = (pairs won + ½ × ties) ÷ all pairs = (30 + ½ × 4) ÷ 40 = 32/40 = **0.80**. (Check: 30 + 4 + 6 = 40 = 8 × 5.)

3. In the MUAC example, compute Youden's J and LR+ at the cut-offs ≤ 12.3 cm and ≤ 12.9 cm. Which would you recommend for community screening, and why does J not settle the question?

≤ 12.3 cm: Se 0.7, Sp 0.9. J = 0.7 + 0.9 − 1 = **0.6**; LR+ = 0.7/(1 − 0.9) = **7.0**. ≤ 12.9 cm: Se 0.9, Sp 0.7. J = 0.9 + 0.7 − 1 = **0.6**; LR+ = 0.9/0.3 = **3.0** (LR− = 0.1/0.7 = 0.14). J is identical, because J counts a missed case and a false alarm as equally bad. In community screening for SAM, a missed child may die while a false positive costs one referral visit, so ≤ 12.9 cm (higher sensitivity) is preferable. The choice rests on the costs of the two errors and on the referral capacity, not on J.

Checkpoint 15.9

1. Two logistic prediction models for 1-year TB treatment default both have a c-statistic of 0.78 in a validation sample. Model A's calibration slope is 1.02 and intercept 0.01; model B's slope is 1.0 but it predicts every patient's risk about twice too high. What does the shared c-statistic tell you, and what does it fail to tell you? Which model would you give to a counsellor who quotes risks to patients?

The c-statistic measures discrimination: in 78% of defaulter–non-defaulter pairs, both models give the defaulter the higher predicted risk. Both rank patients equally well. It says nothing about calibration, whether the predicted probabilities are numerically right. Model B would tell a patient with a true 10% risk that the risk is about 20%. For ranking patients to target extra support, either would do; for quoting risks, use model A (slope ≈ 1 and intercept ≈ 0 mean its probabilities match observed frequencies), or recalibrate B's intercept first.

2. A model's apparent Somers' Dxy is 0.60. Bootstrap validation gives an average optimism of 0.06 in Dxy and a corrected calibration slope of 0.85. Compute the apparent and corrected c-statistics, and explain what a slope of 0.85 says about the predictions.

C = 0.5 + Dxy/2. Apparent C = 0.5 + 0.60/2 = **0.80**. Corrected Dxy = 0.60 − 0.06 = 0.54, so corrected C = 0.5 + 0.54/2 = **0.77**. A calibration slope of 0.85 (below 1) means the predictions are too extreme: in new patients, high predicted risks will be too high and low ones too low, a sign of overfitting. The log-odds predictions should be shrunk towards the mean (for example, multiplied by 0.85) before use.

3. A team reports "Hosmer–Lemeshow p = 0.48, so the model is well calibrated" for 500 households. Why is this conclusion too strong? What would you expect the same test to show if the same model, with the same small miscalibration, were checked in 50,000 households?

A non-significant test is absence of evidence, not evidence of good calibration. The HL test has low power in samples of a few hundred, its result depends on the arbitrary grouping into tenths, and it gives no information about how the model is miscalibrated. A calibration plot with slope and intercept is more informative. In 50,000 households the same small miscalibration would very probably be significant: the test's power rises with n, and in simulations with 50,000 patients a 0.4% deviation made it significant in every model (Kramer & Zimmerman, 2007). Significance there would not mean the model is useless.

Checkpoint 15.10

1. A trial was sized at 91 girls per arm for 80% power with 1:1 allocation. For recruitment reasons the investigators switch to 2:1 (intervention:control) but keep the same total of 182. Compute the relative efficiency and the new power. How many girls would 2:1 need to keep 80% power?

For k:1 allocation the variance of the difference is $(1 + k)^2/(4k)$ times that of 1:1 at the same total. For k = 2: 9/8 = 1.125, so the relative efficiency is 8/9 = **0.889** (11% information lost). At 1:1, $\Delta/SE = z_{\alpha/2} + z_{\beta} = 1.960 + 0.842 = 2.802$. At 2:1 it falls to $2.802/\sqrt{1.125} = 2.641$. Power = $\Phi(2.641 - 1.960) = \Phi(0.681)$ = **0.75**, down from 0.80: a modest loss. To keep 80% power: $180.8 \times 1.125 = 203.4 \rightarrow n_1 = 67.8 \rightarrow 68$ and $n_2 = 135.6 \rightarrow 136$, **204 girls** in all (about 22 more).

2. A calculation gives 200 per group, but 20% loss to follow-up is expected. What should the enrolment target be, and why is 200 × 1.2 slightly wrong?

Enrol n/(1 − loss) = 200/0.80 = **250 per group**, so that 80% of 250 = 200 complete the study. 200 × 1.2 = 240 would leave only 0.80 × 240 = 192, below the required 200. The calculated n assumes complete data on everyone analysed.

3. A cohort study will compare 1-year diarrhoea risk in infants of mothers with and without a latrine at home. Risk without a latrine is 15%; the smallest important difference is a fall to 10%. Compute n per group for 80% power at α = 0.05, and then allow for 10% loss to follow-up.

$p_1 = 0.15$, $p_2 = 0.10$, $\bar{p} = 0.125$, $\Delta = 0.05$. $\sqrt{(2\bar{p}\bar{q})} = \sqrt{(2 \times 0.125 \times 0.875)} = \sqrt{0.21875} = 0.4677$. $\sqrt{(p_1q_1 + p_2q_2)} = \sqrt{(0.1275 + 0.0900)} = \sqrt{0.2175} = 0.4664$. $n = (1.960 \times 0.4677 + 0.8416 \times 0.4664)^2 \div 0.05^2 = (0.9167 + 0.3925)^2 \div 0.0025 = 1.7140 \div 0.0025 = 685.6 \rightarrow$ **686 per group**. With 10% loss: 685.6/0.90 = 761.8 → **762 per group** (1,524 infants). (Software gives the same 685.6.)

Checkpoint 15.11

1. A meta-analysis of ten trials shows $I^2 = 68\%$. A colleague says: "I^2 is over 50%, so we must use a random-effects model." What reasoning does this shortcut, and what should come first?

The model should be chosen before the data are pooled, on clinical and methodological grounds: are the trials similar enough in population, intervention, comparator and outcome definition to share one true effect? If not, or if in doubt, a random-effects model (estimating the mean of a distribution of true effects) fits the question. I^2 can then confirm or challenge that judgement, but it is a descriptive proportion with conventional, overlapping bands (Higgins et al., 2003; Cochrane Handbook), not a switch. $I^2 = 68\%$ should also prompt a search for the sources of heterogeneity (prespecified subgroups) and a prediction interval.

2. A forest plot's pooled diamond is narrower than every study's CI. Explain why, using inverse-variance weights. When might a random-effects diamond be barely narrower than the most precise study?

Each study gets weight $w_i = 1/SE_i^2$, and the pooled estimate $\Sigma w_i y_i/\Sigma w_i$ has $SE = 1/\sqrt{\Sigma w_i}$. Because Σw_i exceeds every single w_i, the pooled SE is smaller than every study's SE, and the pooled CI narrower. Under a random-effects model each weight becomes $1/(SE_i^2 + \tau^2)$. When τ^2 is large compared with the SE_i^2, every weight is limited by τ^2, the pooled SE is about $\sqrt{(\tau^2/k)}$ or more, and the diamond can be little narrower than, or even wider than, the most precise study's CI.

3. Five studies give Q = 12.0. Compute I^2. Another five-study meta-analysis gives Q = 3.0: what is I^2, and why is it reported as that value?

df = k − 1 = 4. I^2 = (12.0 − 4)/12.0 = 8/12 = **66.7%**. For Q = 3.0: (3.0 − 4)/3.0 = −33%, which is truncated to **I^2 = 0%**. Under H_0 (no heterogeneity), Q is expected to be about its df, and chance alone often makes it smaller. A negative excess means no evidence of heterogeneity, and a proportion of variance cannot be negative.

Checkpoint 15.12

1. A meta-analysis of six trials reports "Egger's test p = 0.31, so there is no publication bias." Give two reasons this conclusion is unsafe.

1. With six studies the test has very low power; the Cochrane Handbook advises such tests only when there are at least 10 studies. A non-significant result is not evidence of absence. (ii) The test detects funnel-plot asymmetry, not publication bias as such: bias can exist without asymmetry (for example, if large as well as small null studies are unpublished), and asymmetry can arise without bias (heterogeneity, poorer quality of small studies, chance). A search of trial registries and grey literature says more about missing studies than the test does.

2. In a meta-analysis of 12 trials, the four smallest (SE of ln RR 0.50 each) report RRs of 0.40 to 0.50, and the eight largest (SE 0.10 each) cluster around 0.90. Compute the share of the total fixed-effect weight carried by the four small trials. Would a random-effects analysis increase or decrease their influence, and why does that matter here?

Weights: small trials $1/0.50^2$ = 4 each, 16 in total; large trials $1/0.10^2$ = 100 each, 800 in total. Share = 16/816 = **2.0%**. The fixed-effect estimate is essentially that of the large trials (about 0.90). The small and large trials clearly disagree, so τ^2 will be positive. A random-effects model adds τ^2 to every variance, making the weights more equal and *increasing* the small trials' influence, pulling the pooled RR towards 0.4–0.5. Here the disagreement follows study size (a small-study effect, possibly publication bias), so the random-effects estimate would be the more misleading one. Report both and discuss why the small trials differ.

3. Trim-and-fill imputes four "missing" trials and moves the pooled RR from 0.70 to 0.82. A colleague writes "the true effect is RR 0.82". How would you rewrite the sentence?

"The pooled RR was 0.70 (with its CI). The funnel plot was asymmetric; in a sensitivity analysis assuming the asymmetry reflects unpublished studies, trim-and-fill imputed four trials and gave an RR of 0.82 (with its CI). The effect may therefore be overestimated." Trim-and-fill assumes that the asymmetry is caused by publication bias and that the imputed studies mirror the observed ones; its estimate is a sensitivity analysis (Peters et al., 2007), not the true effect.

Checkpoint 15.13

1. With the prior Beta(20, 30), a survey in a second PHC finds 12 of 40 pregnant women anaemic. Give the posterior distribution and its mean. Is the posterior mean closer to the prior mean or to the sample proportion, and why?

Posterior = Beta(20 + 12, 30 + 28) = **Beta(32, 58)**. Mean = 32/90 = **0.356** (95% credible interval 0.261 to 0.457 from software; by hand, SD = $\sqrt{[32 \times 58 \div (90^2 \times 91)]}$ = 0.050, and 0.356 ± 1.96 × 0.050 = 0.257 to 0.454). Prior mean 0.40; sample proportion 12/40 = 0.30. The posterior mean is a weighted average with weights 50 (the prior's "sample size", a + b) and 40 (the data): (50 × 0.40 + 40 × 0.30)/90 = 0.356. It lies slightly closer to the prior (0.044 away) than to the data (0.056 away), because the prior carries more information (50) than the new sample (40).

2. A report says: "The 95% credible interval for the risk difference was 1% to 9%, so there is a 95% probability that the true risk difference lies in this range." What condition must be added for the sentence to be correct? Could the same sentence be written about a 95% confidence interval?

It must be conditional on the prior and the model: "given the prior distribution and the statistical model used, there is a 95% posterior probability …". With a different prior, or a wrong model, the interval and the probability change. The sentence cannot be written about a frequentist 95% CI: there the 95% refers to the method (95% of intervals built this way, over repeated samples, contain the true value), and the parameter is not given a probability distribution.

3. A clinician's pre-test probability of tuberculous lymphadenitis is 10%. A test with LR+ 6 is positive. Compute the post-test probability, and name the prior, the likelihood and the posterior in this calculation.

Pre-test odds = 0.10/0.90 = 0.111. Post-test odds = 0.111 × 6 = 0.667. Post-test probability = 0.667/1.667 = **0.40**. Prior: the pre-test probability (10%, or odds 0.111). Likelihood: the likelihood ratio, P(positive | disease)/P(positive | no disease) = 6, which carries the evidence of the result. Posterior: the post-test probability, 40%.

Checkpoint 15.14

1. A trial's Table 1 reports baseline weight with "p = 0.04" and the authors write that "randomisation failed". Explain why the p-value is the wrong tool and what the authors should do about a prognostically important imbalance.

If allocation was truly random and concealed, every baseline difference arose by chance, so the null hypothesis of the test is known to be true, and about 1 in 20 such tests will give $p < 0.05$ anyway (more when many variables are tested). A small p says nothing about whether randomisation failed; failure is judged from the process (sequence generation, concealment). What matters is whether an imbalance is large enough to affect the outcome, which depends on how prognostic the variable is, not on its p-value. The remedy is to adjust for prognostic baseline variables specified in the protocol (for example, by ANCOVA), whatever their balance, and to report baseline characteristics descriptively (Senn, 1994; Schulz & Grimes, 2002).

2. In a planned trial of a salt-reduction intervention, the correlation between baseline and 6-month systolic BP is expected to be 0.6. A final-value analysis would need 200 participants per arm. Compute the variance factors for a change-score analysis and for ANCOVA, and the corresponding sample sizes per arm.

Relative to the final-value analysis (factor 1): change score: $2(1 - \rho) = 2 \times 0.4 =$ **0.8** → 200 × 0.8 = **160 per arm**; ANCOVA: $1 - \rho^2 = 1 - 0.36 =$ **0.64** → 200 × 0.64 = **128 per arm**. ANCOVA needs 36% fewer participants than a final-value analysis and 20% fewer than a change-score analysis.

3. A non-inferiority trial of a community-based versus facility-based treatment for a childhood infection uses a margin of −10 percentage points. The ITT difference in cure is −3 points (95% CI −8.5 to +2.5); the per-protocol difference is −6 points (95% CI −11.8 to −0.2). What do you conclude, and why are both analyses reported?

ITT: the lower limit −8.5 lies above −10, so non-inferiority is shown. Per-protocol: the lower limit −11.8 lies below −10, so non-inferiority is not shown; the CI also lies wholly below 0, so community treatment was significantly worse among those who followed the protocol. The conclusion is that non-inferiority is **not established**. In a non-inferiority trial, ITT is not conservative: non-adherence and dropout in both arms make the arms look alike and bias the result towards non-inferiority (ICH, 1998). The per-protocol analysis loses the protection of randomisation, so its bias has no predictable direction; each covers the other's weakness, so both are reported and a convincing claim needs both to agree.

Chapter 16: Measurement Theory – Constructs, Validity, and Reliability

Checkpoint 16.1

1. A new depression screening questionnaire for elderly rural patients is adapted word-for-word from a validated questionnaire for urban working-age adults, without any local review. Which type of validity is most directly at risk, and why?

Content validity: whether the items cover the construct fully and appropriately for *this* population. The way depression shows itself, the idioms used to describe it and the relevant life domains differ by age and culture. Items about work stress may not apply to elderly rural patients, and locally common expressions of distress (for example, bodily complaints) may be missing. An expert and community review, with cultural adaptation (§16.7), is needed before use.

2. A clinician glances at a new anxiety scale and says "these questions look like they measure anxiety." Which more limited type of validity does this comment support?

Face validity only: a surface impression that the instrument appears to measure what it claims. It is weaker than content validity, which needs a systematic expert review of whether the items cover the whole construct.

Checkpoint 16.2

1. A new physical activity scale correlates at r = 0.85 with an existing validated physical activity scale, and at r = −0.30 with a sedentary-behaviour screen (a related but distinct construct). Which strand of validity does each finding support? Why is the second comparison more informative than a correlation with shoe size would be?

The r = 0.85 with another physical-activity measure supports **convergent** validity. The much weaker r = −0.30 with a sedentary-behaviour screen supports **discriminant** validity: the new scale is not simply the reverse of "sitting". Both are strands of construct validity. Sedentary behaviour is a construct a critic could plausibly confuse with physical activity, so a clear gap between 0.85 and a correlation of only −0.30 is informative. Nobody expects physical activity to track shoe size, so a near-zero correlation with it would confirm nothing.

2. A new scale to predict medication adherence correlates well with adherence measured six months later. Which form of validity is this, and how does it differ from correlating with adherence measured on the same day?

Predictive validity: the criterion is measured in the future. Correlation with adherence measured at the same time would be **concurrent validity**. Both are forms of criterion validity, told apart by when the criterion is measured.

3. A tool predicting low birthweight is given at 20 weeks to 250 women. Of 40 babies later born under 2.5 kg, 30 mothers had screened positive; of 210 others, 42 had screened positive. Compute the sensitivity and specificity. Why are these reported instead of r?

1. Sensitivity = 30 ÷ 40 = **0.75**.
2. Specificity = (210 − 42) ÷ 210 = 168 ÷ 210 = **0.80**.

The criterion (low birthweight, yes/no) is binary, and the tool is used to classify. Predictive validity is therefore reported as sensitivity and specificity (and predictive values, or AUC for a continuous score), not a correlation (§16.2.3.2).

Checkpoint 16.3

1. A researcher gives the same questionnaire to the same patients twice, two weeks apart, and finds a high correlation. Which type of reliability is this? Name one threat that would inflate the estimate and one that would deflate it, even for a well-designed instrument.

Test-retest reliability. *Inflating*: respondents remember and repeat their first answers (memory or practice effects), especially with a short interval. *Deflating*: real change in the construct between occasions (for example, a patient's anxiety actually improving), or an intervening event such as health education. A low value does not by itself show the instrument is unreliable. Note also that a correlation ignores a systematic shift between occasions, so ICC(2,1) is preferred (§16.3.2).

2. A four-item scale has item variances 1.2, 1.0, 1.5 and 1.3 and total-score variance 12.0. Compute Cronbach's alpha. A ten-item scale has α = 0.97. What might this very high value signal?

1. $\Sigma s_i^2 = 1.2 + 1.0 + 1.5 + 1.3 = 5.0$.
2. $\alpha = [4 \div 3] \times (1 - 5.0 \div 12.0) = 1.333 \times 0.583 =$ **0.78**, satisfactory for group comparisons by the Bland & Altman (1997) convention.

α = 0.97 on ten items may mean **redundancy**: several items ask the same question in different words, adding length without new information. Terwee et al. (2007) rate α positively only up to 0.95. Also, α rises with the number of items (Spearman–Brown), so a high value partly reflects length. It does not show that the scale is unidimensional or valid.

3. Two nurses classify 200 children as "fever" or "no fever" by touch. Both say fever for 30, neither for 140, nurse A only for 20, nurse B only for 10. Compute p_o, p_e and κ, and give the Landis & Koch label.

1. $p_o = (30 + 140) \div 200 =$ **0.85**.
2. Nurse A says fever for $(30 + 20) \div 200 = 0.25$; nurse B for $(30 + 10) \div 200 = 0.20$.
3. $p_e = 0.25 \times 0.20 + 0.75 \times 0.80 = 0.05 + 0.60 =$ **0.65**.
4. $\kappa = (0.85 - 0.65) \div (1 - 0.65) = 0.20 \div 0.35 =$ **0.57**: "moderate" (0.41–0.60) on the Landis & Koch (1977) scale, a convention. Report the two marginal proportions too: nurse A calls fever more often (bias), and fever is the minority category (prevalence), both of which affect κ.

Checkpoint 16.4

1. A new haemoglobinometer and the laboratory cyanmethaemoglobin method are compared in 40 pregnant women. Mean difference (new − lab) is −0.3 g/dL, SD of differences 0.6 g/dL. Compute the 95% limits of agreement. The study defines anaemia as Hb below 11 g/dL: interpret the limits for a woman whose laboratory value is 11.2 g/dL.

1. Lower limit = $-0.3 - 1.96 \times 0.6 = -0.3 - 1.18 =$ **−1.48 g/dL**.
2. Upper limit = $-0.3 + 1.96 \times 0.6 = -0.3 + 1.18 =$ **0.88 g/dL**.

For about 95% of women, the new device reads between 1.48 g/dL below and 0.88 g/dL above the laboratory. For a woman whose laboratory value is 11.2 g/dL, the device could plausibly read anywhere from about 9.7 to 12.1 g/dL, so she could be labelled anaemic or not. The device under-reads on average (bias −0.3 g/dL) and is too imprecise to replace the laboratory for individual diagnosis near the cut-off, unless the investigators judged a difference of about 1.5 g/dL clinically acceptable beforehand.

2. The same study reports r = 0.96 between the two methods and concludes "excellent agreement". What is wrong with this conclusion?

r measures how closely the points lie along *some* straight line, not whether the readings are equal. It would be 1 even if the device always read exactly 1 g/dL low. It also rises when the women's true haemoglobin values vary widely, which says nothing about agreement. Agreement is assessed by the bias and limits of agreement (Bland & Altman, 1986), as in question 1.

Checkpoint 16.5

1. A team runs PCA on 20 items meant to measure one construct, "quality of life". Three components have eigenvalues above 1 (4.1, 1.3 and 1.1), explaining 20.5%, 6.5% and 5.5% of the variance. What do the percentages mean? What should the team check before concluding that the scale has three sub-dimensions?

With 20 standardised items the total variance is 20. Each percentage is the eigenvalue ÷ 20: the first component carries 4.1 ÷ 20 = 20.5% of the total variance, the next two 1.3 ÷ 20 = 6.5% and 1.1 ÷ 20 = 5.5%.

The eigenvalue > 1 (Kaiser) rule is a convention that can keep too many components. Before concluding there are three sub-dimensions, the team should:

- inspect the scree plot and run parallel analysis. Here the first component dominates, and the next two only just exceed 1;
- look at the loadings. Does each extra component have several items loading highly, with a sensible theme (for example, all the social-relationship items)? Or is it one or two items, or a method artefact such as all the negatively worded items loading together?
- check the result against theory (for example, the WHOQOL domains) and, ideally, confirm the structure by CFA in a new sample.

If the three components prove real, report α and build scores for each sub-scale, not one total.

2. Explain why a highly reliable instrument is not automatically valid. A scale has reliability 0.64 and its criterion has reliability 0.81. What is the largest correlation the scale could show with the criterion?

Reliability is consistency. An instrument can consistently measure the wrong thing, like a scale that always reads 2 kg high. Reliability is necessary but not sufficient for validity.

Largest possible correlation = $\sqrt{(r_{xx} \times r_{yy})} = \sqrt{(0.64 \times 0.81)} = \sqrt{0.5184} =$ **0.72** (§16.4).

3. A Hindi translation of an English depression scale is to be used in Raipur. List the translation steps, and name two measurement properties you would test afresh, with the statistic for each.

Translation (Beaton et al., 2000): (1) two independent forward translations into Hindi; (2) synthesis of the two; (3) back-translation into English by a translator blind to the original; (4) expert committee review to agree a prefinal version; (5) pretest (cognitive interviews) in the target population.

Properties to test afresh, for example:

- **Internal consistency**: Cronbach's α per sub-scale, after checking structure by PCA/EFA (or CFA against the original structure).
- **Test-retest reliability**: ICC(2,1), absolute agreement, over a suitable interval in at least 50 respondents (Terwee et al., 2007).
- **Criterion validity**: sensitivity, specificity and AUC against a psychiatrist's diagnostic interview.

- **Construct validity**: predefined correlations, or known-groups comparisons (for example, attenders at a psychiatry OPD against community controls).

Chapter 17: Threats to Study Validity

Checkpoint 17.1

1. A rigorous RCT conducted only at a single urban tertiary hospital shows that a treatment works, with excellent internal validity. A district health officer wants to apply the finding in rural PHCs. What should she consider first, and what kind of evidence would answer it?

External validity: whether the *effect* is the same in rural PHC patients. That depends on whether the effect is modified by the ways PHC patients differ (severity, baseline nutrition, co-morbidity, adherence, follow-up care). The best evidence is consistency: trials or well-conducted observational studies in similar settings, or subgroup results from the original trial (tested for interaction) for patients like hers. Internal validity was the precondition, and it is met.

2. Explain why internal and external validity can pull against each other in study design.

Choices that strengthen internal validity (narrow eligibility, standardised protocols, intensive monitoring, excluding people likely to drop out) make the participants and conditions less like routine practice. The effect is then estimated cleanly but for a narrower population, so its transfer to others rests more on judgement.

3. Study A (n = 120) reports RR 1.9 (95% CI 0.9 to 4.0). Study B (n = 12,000), using the same exposure questionnaire with the same systematic error, reports RR 1.35 (95% CI 1.25 to 1.46). Which alternative explanation has Study B's size dealt with, and which has it not?

Chance: B's interval is narrow and excludes 1, so random variation is an unlikely explanation for its association. Size does nothing for bias or confounding: the same measurement error is present in B, and the narrow interval is now centred on an estimate that is still biased by it. (Confounding, too, is unchanged by n.)

Checkpoint 17.2

1. A cohort has 100 exposed people (30 develop the outcome) and 100 unexposed (10 develop it), so the true RR is 3.0. Exposure is recorded with Se = 0.90 and Sp = 0.80, equally in people with and without the outcome. Compute the recorded 2 × 2 table and the recorded RR. What kind of misclassification is this, and in which direction did it move the estimate?

1. Among the 40 with the outcome (30 exposed, 10 unexposed): recorded exposed = 0.90 × 30 + 0.20 × 10 = 27 + 2 = **29**; recorded unexposed = 11.
2. Among the 160 without (70 exposed, 90 unexposed): recorded exposed = 0.90 × 70 + 0.20 × 90 = 63 + 18 = **81**; recorded unexposed = 79.
3. Recorded exposed group: 29 + 81 = 110, risk 29 ÷ 110 = 0.264. Recorded unexposed: 11 + 79 = 90, risk 11 ÷ 90 = 0.122.
4. Recorded RR = 0.264 ÷ 0.122 = **2.16**, against a true 3.0.

Non-differential misclassification of a binary exposure (the same Se and Sp whatever the outcome). It moved the RR towards the null, as expected when Se + Sp > 1 (here 1.70).

2. In a case-control study of low birthweight, (a) interviewers who know which mothers had a low-birthweight baby ask them more follow-up questions about tobacco chewing, and (b) mothers of low-birth-

weight babies, looking for a reason, report chewing more completely. Name each bias and give one way to prevent each.

1. Interviewer bias (information bias arising from the data collector): prevent with a scripted questionnaire asked identically of all, training, and interviewers kept unaware of case status where possible. (b) Recall bias (information bias arising from the participant's memory): prevent by using exposure recorded before the outcome (antenatal cards), or by choosing controls with another adverse outcome that would prompt similar reflection. Both are differential misclassification, so the direction cannot be assumed.

3. Women taking an oral contraceptive visit a doctor more often than other women, and more hypertension is recorded among them. Is this observer bias or detection bias? How would you design the study to prevent it?

Detection bias: the exposure brings women under medical attention, so their blood pressure is measured more often and hypertension is found more often, whatever its true frequency. Observer bias would be a measurer reading blood pressure differently because she knows who takes the pill. Prevention: measure blood pressure in every participant on the same schedule with the same method (equal surveillance), ideally by staff blinded to contraceptive use.

Checkpoint 17.3

1. Name the bias in each: (a) a study of coal miners finds their all-cause mortality lower than the state's; (b) a case-control study of tuberculosis and diabetes recruits cases from patients currently on treatment at a DOTS centre; (c) a free cancer screening camp finds that participants have lower cancer mortality than the district; (d) women in a diet study eat more vegetables in the weeks they know they are being observed.

1. Healthy-worker effect: miners were fit enough to be hired and to stay employed; compare exposed with less-exposed miners instead. (b) Neyman (prevalence–incidence) bias: prevalent cases on treatment exclude those who died early; if diabetes raises early TB mortality, diabetic cases are under-represented and the association is understated. Use newly diagnosed (incident) cases. (c) Volunteer (self-selection) bias: people who attend screening camps are often healthier and more health-conscious; a randomised comparison of invited and not-invited populations is needed. (Lead-time and length-time bias also threaten survival comparisons here.) (d) Hawthorne effect.

2. An oral cancer screening programme detects a cancer at age 52 that would have caused symptoms at 55; the patient dies at 58. Compute survival from diagnosis with and without screening and the lead time. What outcome measure should the programme's evaluation use?

1. With screening: 58 − 52 = **6 years**.
2. Without screening: 58 − 55 = **3 years**.
3. Lead time = 55 − 52 = **3 years**, the whole apparent gain; age at death is 58 either way.

Evaluate by oral-cancer mortality (deaths per 100,000 per year) in populations randomised to screening or not, not by survival from diagnosis.

3. Which of Hill's viewpoints is logically necessary for any causal claim, and why would a lack of specificity be a weak argument against tobacco as a cause of lung cancer?

Temporality: a cause must come before its effect; the other eight add weight but none is required. Specificity (one cause, one effect) is weak because many exposures cause several diseases: tobacco causes lung cancer, oral cancer, heart disease and more. Hill himself warned: "We must not, however, over-emphasise the importance of the characteristic."

Checkpoint 17.4

1. Two studies each find no association between a workplace exposure and a health outcome. Study X sampled workers well but did not measure smoking, a known confounder. Study Y measured everything but recruited from a clinic that sees mainly workers who fell ill. Explain why these problems differ, how each could be addressed, and how each could have *hidden* a real effect.

Study X has **unmeasured confounding**. Smoking was never recorded, so it cannot be adjusted for in these data. It can be addressed with new data (measure smoking, or a validation sub-sample), or bounded with a sensitivity analysis (§17.3.1): how strongly would smoking have to be related to exposure and outcome to change the conclusion? It could hide an effect if the confounding is negative: for example, if exposed workers smoke *less* than unexposed ones, and smoking raises the outcome risk, the harm of the exposure is offset.

Study Y has **selection bias** only if entry to the clinic depends on both exposure and outcome. Recruiting only workers who fell ill means everyone has, or is likely to have, the outcome; if exposed workers attend the clinic for exposure-related reasons as well, clinic attendance is a collider (§17.2.1) and the association among attenders can be pushed towards, or across, the null. No analysis of these data restores the missing workers; the fix is redesign (a cohort of all workers, or population-based cases and controls). Separately, a non-representative sample that did not depend on both would limit generalisability without biasing the association.

Both studies could also be affected by non-differential misclassification of exposure, which tends to push a real effect towards the null. "No association" is not "no effect" until these explanations are addressed.

2. An adjusted RR is 1.8 (95% CI 1.2 to 2.7). Compute the E-values for the estimate and the confidence limit. Would a confounder with RR_EU = RR_UD = 2 explain the association away? (For a protective RR of 0.5, what is the E-value?)

1. Estimate: $1.8 + \sqrt{(1.8 \times 0.8)} = 1.8 + \sqrt{1.44} = 1.8 + 1.2 =$ **3.0**.
2. Lower limit (closer to the null): $1.2 + \sqrt{(1.2 \times 0.2)} = 1.2 + \sqrt{0.24} = 1.2 + 0.49 =$ **1.69**.
3. Confounder (2, 2): $B = 4 \div 3 = 1.33$. Corrected RR $\geq 1.8 \div 1.33 = 1.35$, corrected lower limit $\geq 1.2 \div 1.33 = 0.90$. It cannot explain away the point estimate ($2 < 3.0$) but could explain away the lower limit ($2 > 1.69$): the interval could then include 1.
4. Protective RR 0.5: invert to 2.0; E-value $= 2 + \sqrt{(2 \times 1)} = 2 + 1.41 =$ **3.41**.

3. A cohort adjusts for maternal education recorded only as "literate / not literate", and an association remains. A colleague calls what remains "unmeasured confounding". Correct her. Then name the three design-stage methods of controlling confounding and one limitation of each.

Education *was* measured, but coarsely. What remains is **residual confounding**: differences in education within each of the two categories still confound. Recording years of schooling or finer categories would reduce it. Unmeasured confounding refers to a confounder never recorded.

Design-stage methods: (1) **Randomisation**: balances measured and unmeasured confounders on average, but only when the exposure can ethically and practically be assigned. (2) **Restriction**: admit one level of the confounder only; it cannot confound, but numbers fall and the result applies only to that level. (3) **Matching**: make comparison groups alike on the confounder; the analysis must account for the matching (in case-control studies matching alone does not remove confounding), and a matched variable's own effect cannot then be studied.

Checkpoint 17.5

1. Name the reporting guideline for each: (a) a cluster-randomised trial of weekly iron–folic acid in schools; (b) a case-control study of neonatal sepsis; (c) the accuracy of a digital haemoglobin meter against the laboratory; (d) a systematic review of yoga for hypertension; (e) FGDs with ASHAs on home visits; (f) a risk score predicting pre-eclampsia.

1. CONSORT, with its extension for cluster trials. (b) STROBE. (c) STARD. (d) PRISMA. (e) COREQ (or SRQR). (f) TRIPOD+AI, which supersedes TRIPOD 2015 (§17.4).

2. Rewrite for a results section: "Not exclusively breastfed infants had significantly more diarrhoea (χ^2 = 23.37, $p < 0.001$)." Use the numbers of §10.6.4.

"Diarrhoea occurred in 33 of 80 infants (41.3%) who were not exclusively breastfed and 14 of 120 (11.7%) who were, a risk ratio of 3.54 (95% CI 2.02 to 6.17; χ^2 = 23.37, $p < 0.001$)." The effect and its interval come first; the test statistic and p-value follow, if given at all.

3. A district survey finds that only 35% of pregnant women took 100 IFA tablets. The team then interviews 15 women who did not, choosing them to include different ages, parities and blocks, and stops when new interviews add no new reasons. Name the mixed-methods design, the sampling approach and the stopping rule.

Explanatory sequential design (quantitative first, qualitative to explain it). Purposive sampling, with maximum variation across age, parity and block. Stopping at saturation, when new data yield no new codes or themes; 15 in-depth interviews lies within the 9–17 range at which empirical studies reached saturation for fairly homogeneous groups (Hennink & Kaiser, 2022).

R Quick Reference

Every function used in an **In R** box, with the package it comes from and the sections where it is used. Base-R helpers (c(), mean(), table(), and so on) are left out. All examples read the synthetic teaching datasets described in §1.6.

Function	Package	Used in
`abline()`	base R	§16.3.5
`ageadjust.direct()`	epitools	§7.4.2
`AIC()`	base R	§15.1.2
`alpha()`	psych	§16.3.1.1
`anova()`	base R	§12.4.5, §15.1.2, §15.2.2
`aov()`	base R	§9.5.2, §9.5.3
`as.matrix()`	Matrix	§14.8
`as.vector()`	base R	§15.9
`auc()`	pROC	§15.5.1
`BIC()`	base R	§15.1.2
`binom.test()`	base R	§7.2.4, §9.1.1, §10.1.1
`binomial()`	base R	§13.7
`boot()`	boot	§7.2.4
`boot.ci()`	boot	§7.2.4
`calibrate()`	survey	§15.5.2
`chisq.test()`	base R	§10.6.1, §10.6.3, §10.6.5
`ci.auc()`	pROC	§15.5.1
`clogit()`	survival	§13.2.9
`coefci()`	lmtest	§13.7
`coeftest()`	lmtest	§13.7
`cohen.kappa()`	psych	§16.3.3.1
`combn()`	base R	§10.4
`confint.default()`	base R	§13.2.3, §13.2.5, §13.3, §13.5, §13.6
`cooks.distance()`	base R	§12.5.1, §12.6
`coords()`	pROC	§15.5.1

Function	Package	Used in
`cor()`	base R	§12.4.4.1, §15.9, §16.3.2, §16.3.5
`cor.test()`	base R	§11.2, §11.3
`cortest.bartlett()`	psych	§16.5
`cov()`	pROC	§11.2
`cox.zph()`	survival	§14.6.2
`coxph()`	survival	§14.6.2, §14.6.3, §14.7
`crr()`	cmprsk	§14.8
`cuminc()`	cmprsk	§14.8
`cumprod()`	base R	§7.5
`dhyper()`	base R	§10.7
`duplicated()`	base R	§1.6
`eigen()`	base R	§16.5
`epi.2by2()`	epiR	§10.6.4, §11.7, §17.3.1
`escalc()`	metafor	§15.7.3, §15.7.4
`fisher.test()`	base R	§10.7
`fitted()`	base R	§15.3.2
`fixef()`	lme4	§15.2.4, §15.2.4.1
`for()`	base R	§15.3.2
`forest()`	metafor	§15.7.3
`formula()`	base R	§15.1.3
`friedman.test()`	base R	§10.5
`gaussian()`	base R	§13.1
`geeglm()`	geepack	§15.2.5
`glm()`	base R	§13.1, §13.2.3, §13.2.5, §13.3, §13.5, §13.7, §15.1.2, §15.1.3 …
`glm.nb()`	MASS	§13.6
`glmer()`	lme4	§15.2.4.1
`hatvalues()`	base R	§12.5.1
`I()`	base R	§15.9
`ICC()`	psych	§16.3.2, §16.3.5
`if()`	base R	§15.3.2, §17.3.1
`KMO()`	psych	§16.5
`kruskal.test()`	base R	§10.4
`leveneTest()`	car	§9.2.4
`lm()`	base R	§12.2.4, §12.4.4.1, §12.4.5, §12.5.1, §12.6, §13.1, §15.2.2, §15.4.3 …
`lmer()`	lme4	§15.2.4

Function	Package	Used in
`logLik()`	base R	§13.6, §15.1.2
`lrm()`	rms	§15.5.2
`mantelhaen.test()`	base R	§11.7
`mcnemar.test()`	base R	§10.8, §15.3.2
`mice()`	mice	§15.4.3
`na.omit()`	base R	§12.6
`offset()`	base R	§13.5, §13.6
`p.adjust()`	base R	§9.5.3, §10.4
`par()`	base R	§12.5.1
`pbeta()`	base R	§15.8
`plot()`	base R	§12.5.1, §15.5.2, §16.3.5
`poisson.test()`	base R	§7.4.4
`pool()`	mice	§15.4.3
`power.prop.test()`	base R	§15.6
`power.t.test()`	base R	§8.4, §15.6
`predict()`	base R	§12.2.4, §13.3, §13.7, §15.7.3
`principal()`	psych	§16.5
`prop.test()`	base R	§7.2.3, §7.2.4, §8.5, §8.7, §9.1.1, §9.1.2
`prop.trend.test()`	base R	§10.6.5
`pwr.r.test()`	pwr	§15.6
`qbeta()`	base R	§15.8
`quasibinomial()`	base R	§15.2.6
`rank()`	base R	§10.4
`rcs()`	rms	§12.4.5
`regtest()`	metafor	§15.7.4
`replicate()`	base R	§15.1.3, §15.9
`residuals()`	base R	§13.6
`rma()`	metafor	§15.7.3, §15.7.4
`roc()`	pROC	§15.5.1
`rstandard()`	base R	§12.5.1, §12.6
`SE()`	survey	§15.2.2
`setdiff()`	base R	§15.3.2
`shapiro.test()`	base R	§9.3.1
`sign()`	base R	§10.1.1
`step()`	base R	§15.1.3
`strata()`	survival	§13.2.9, §14.6.3

Function	Package	Used in
`suppressPackageStartupMessages()`	base R	§12.4.5
`Surv()`	rms	§14.4, §14.5, §14.6.2, §14.6.3, §14.7
`survdiff()`	survival	§14.5
`survfit()`	survival	§14.4
`svyby()`	survey	§15.2.6, §15.3.2
`svydesign()`	survey	§15.2.2, §15.2.6, §15.3.2
`svyglm()`	survey	§15.2.6
`svymean()`	survey	§15.2.2, §15.2.6
`t.test()`	base R	§7.2.1, §9.2.2, §9.2.3, §16.3.5
`terms()`	base R	§15.1.3
`timepoints()`	cmprsk	§14.8
`transform()`	base R	§13.7
`trimfill()`	metafor	§15.7.4
`TukeyHSD()`	base R	§9.5.3
`unname()`	Matrix	§13.2.5, §17.3.1
`update()`	Matrix	§12.6, §15.1.2, §15.1.3, §15.2.4, §15.2.4.1
`validate()`	rms	§15.5.2
`var.test()`	base R	§9.2.4
`VarCorr()`	lme4	§15.2.4.1
`vcovHC()`	sandwich	§13.7
`vif()`	rms	§12.4.4.1
`which.max()`	base R	§12.6
`which.min()`	base R	§15.3.2
`wilcox.test()`	base R	§10.2, §10.3

References

Every source below was opened while this edition was prepared, unless marked otherwise. Author–year citations in the text point here; web pages are cited by organisation.

1. Armitage P. Tests for linear trends in proportions and frequencies. Biometrics. 1955;11(3):375-86.
2. Arya R, Antonisamy B, Kumar S. Sample size estimation in prevalence studies. Indian J Pediatr. 2012;79(11):1482-8.
3. Austin PC, Lee DS, Fine JP. Introduction to the analysis of survival data in the presence of competing risks. Circulation. 2016;133(6):601-9.
4. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res. 2011;46(3):399-424. doi:10.1080/00273171.2011.568786
5. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28(25):3083-107. doi:10.1002/sim.3697
6. Austin PC. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. Stat Med. 2011;30(11):1292-301. doi:10.1002/sim.4200
7. Banda JP. Nonsampling errors in surveys (Chapter 7, draft Handbook on Designing of Household Sample Surveys). UN Statistics Division, ESA/STAT/AC.93/7; 2003.
8. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. Int J Epidemiol. 2005;34(1):215-20. doi:10.1093/ije/dyh299
9. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine. 2000;25(24):3186-91.
10. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B. 1995;57(1):289-300.
11. Berkson J. Limitations of the application of fourfold table analysis to hospital data. Biometrics. 1946;2(3):47-53. Reprinted Int J Epidemiol. 2014;43(2):511-5.
12. Bhandari M, Montori VM, Swiontkowski MF, Guyatt GH. User's guide to the surgical literature: how to use an article about a diagnostic test. J Bone Joint Surg Am. 2003;85(6):1133-40.
13. Bilukha OO. Old and new cluster designs in emergency field surveys: in search of a one-fits-all solution. Emerg Themes Epidemiol. 2008;5:7. doi:10.1186/1742-7622-5-7
14. Birhanu T, Molenberghs G, Sotto C, Kenward MG. Doubly robust and multiple-imputation-based generalized estimating equations. J Biopharm Stat. 2011;21(2):202-25. doi:10.1080/10543406.2011.550096
15. Bland JM, Altman DG. Cronbach's alpha. BMJ. 1997;314(7080):572.
16. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307-10.
17. Bland JM, Altman DG. The logrank test. BMJ. 2004;328(7447):1073.
18. Bland M. An Introduction to Medical Statistics. 4th ed. Oxford: Oxford University Press; 2015.
19. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527.
20. Boua PR, et al. Undernutrition in children aged 0-59 months by region and over time: secondary analysis of the Burkina Faso 2012-2018 National Nutrition Surveys. BMJ Open. 2023;13:e066509.
21. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol. 2006;3(2):77-101.
22. Breslow NE, Day NE. Statistical Methods in Cancer Research, Volume I: The Analysis of Case-Control Studies. IARC Scientific Publication No. 32. Lyon: IARC; 1980.

23. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol. 2006;163(12):1149-56. doi:10.1093/aje/kwj149

24. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993;46(5):423-9.

25. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012;345:e5661.

26. Catalogue of Bias Collaboration. Admission rate bias (Berkson's bias). Catalogue of Bias, Centre for Evidence-Based Medicine, University of Oxford.

27. Catalogue of Bias Collaboration. Ascertainment bias. Catalogue of Bias.

28. Catalogue of Bias Collaboration. Attrition bias. Catalogue of Bias.

29. Catalogue of Bias Collaboration. Collider bias. Catalogue of Bias.

30. Catalogue of Bias Collaboration. Detection bias. Catalogue of Bias.

31. Catalogue of Bias Collaboration. Hawthorne effect. Catalogue of Bias.

32. Catalogue of Bias Collaboration. Information bias. Catalogue of Bias.

33. Catalogue of Bias Collaboration. Lead time bias. Catalogue of Bias.

34. Catalogue of Bias Collaboration. Length time bias. Catalogue of Bias.

35. Catalogue of Bias Collaboration. Misclassification bias. Catalogue of Bias.

36. Catalogue of Bias Collaboration. Non-response bias. Catalogue of Bias.

37. Catalogue of Bias Collaboration. Observer bias. Catalogue of Bias.

38. Catalogue of Bias Collaboration. Prevalence-incidence (Neyman) bias. Catalogue of Bias.

39. Catalogue of Bias Collaboration. Recall bias. Catalogue of Bias.

40. Catalogue of Bias Collaboration. Volunteer bias. Catalogue of Bias.

41. Centers for Disease Control and Prevention. CDC Learning Connection Quick-Learn: Using an Epi Curve to Determine Mode of Spread. Atlanta: CDC; n.d.

42. Centers for Disease Control and Prevention. Principles of Epidemiology in Public Health Practice, 3rd ed. (SS1978), Lesson 4, Section 4: Tables, graphs and charts (maps, pie charts, bar charts). Atlanta: CDC; 2006, updated 2012.

43. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. Br J Cancer. 2003;89(2):232-8.

44. Cleveland WS, McGill R. Graphical perception: theory, experimentation, and application to the development of graphical methods. J Am Stat Assoc. 1984;79(387):531-54. doi:10.1080/01621459.1984.10478080

45. Cochran WG. Some methods for strengthening the common χ^2 tests. Biometrics. 1954;10(4):417-51. (Not consulted directly; cited as the origin of a convention.)

46. Cohen J. A power primer. Psychol Bull. 1992;112(1):155-9. doi:10.1037/0033-2909.112.1.155.

47. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213-20.

48. Cole TJ, Green PJ. Smoothing reference centile curves: the LMS method and penalized likelihood. Stat Med. 1992;11(10): 1305-19.

49. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.

50. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162(1):55-63.

51. CONSORT Group. CONSORT 2025 Statement (website).

52. Dandona R, George S, Majumder M, Akbar M, Kumar GA. Stillbirth undercount in the sample registration system and national family health survey, India. Bull World Health Organ. 2023;101(3):191-201.

53. Data for India. How births and deaths in India are counted (CRS and SRS explainer). 2025.

54. Davidson M. Known-groups validity. In: Michalos AC, ed. Encyclopedia of Quality of Life and Well-Being Research. Springer, Dordrecht; 2014.

55. de Boer MR, Waterlander WE, Kuijper LDJ, Steenhuis IHM, Twisk JWR. Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate. Int J Behav Nutr Phys Act. 2015;12:4.

56. Delacre M, Lakens D, Leys C. Why psychologists should by default use Welch's t-test instead of Student's t-test. International Review of Social Psychology. 2017;30(1):92-101.

57. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177-88.

58. Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. Epidemiology. 2016;27(3):368-77.

59. Dunn OJ. Multiple comparisons using rank sums. Technometrics. 1964;6(3):241-52.

60. Dunnett CW. New tables for multiple comparisons with a control. Biometrics. 1964;20(3):482-91. (Extends Dunnett CW, JASA 1955;50:1096-121.)

61. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000;56(2):455-63.

62. Efron B. Bootstrap methods: another look at the jackknife. Ann Statist. 1979;7(1):1-26.

63. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629-34.

64. Epicollect5. Free and easy-to-use mobile data-gathering platform. Oxford University Big Data Institute.

65. EQUATOR Network. Enhancing the QUAlity and Transparency Of health Research.

66. Eypasch E, Lefering R, Kum CK, Troidl H. Probability of adverse events that have not yet occurred: a statistical reminder. BMJ. 1995;311(7005):619-20.

67. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543-9.

68. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs—principles and practices. Health Serv Res. 2013;48(6 Pt 2):2134-56.

69. Fewell Z, Davey Smith G, Sterne JAC. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. Am J Epidemiol. 2007;166(6):646-55.

70. Filmer D, Pritchett LH. Estimating wealth effects without expenditure data–or tears: an application to educational enrollments in states of India. Demography. 2001;38(1):115-32.

71. Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika. 1915;10(4):507-21.

72. Fisher RA. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd; 1925. Chapter III.

73. Gavi Zero-Dose Learning Hub. Lot Quality Assurance Sampling (LQAS). Undated web resource; accessed 2026-09-28.

74. GraphPad Prism 11 Statistics Guide. Fisher's Least Significant Difference (LSD). n.d.

75. GraphPad Prism 11 Statistics Guide. Relationship between overall ANOVA and multiple comparisons tests (quoting Maxwell SE, Delaney HD, Designing Experiments and Analyzing Data). n.d.

76. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37-48.

77. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. 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.

78. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924-6.

79. Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. JAMA. 1983;249(13):1743-5.

80. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36.

81. Harrell FE Jr. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2nd ed. Cham: Springer; 2015.

82. Hart A. Mann-Whitney test is not just a test of medians: differences in spread can be important. BMJ. 2001;323(7309):391-3. doi:10.1136/bmj.323.7309.391

83. Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. Int J Epidemiol. 1999;28(2):319-26.

84. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Stat Med. 2002;21(16):2409-19.

85. Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. Biom J. 2018;60(3):431-449. doi:10.1002/bimj.201700067

86. Held L, Micheloud C, Hofer L, von Felten S, Pawel S, Yang M. Clinical Biostatistics (STA404 lecture notes), Chapter 9 Survival analysis. University of Zurich; 2026 (living document).

87. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med. 2022;292:114523.

88. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615-25.

89. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557-60.

90. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A. 2009;172(1):137-59.

91. Hill AB. The environment and disease: association or causation? Proc R Soc Med. 1965;58(5):295-300. Reprinted J R Soc Med. 2015;108(1):32-7.

92. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat. 2001;55(1). doi:10.1198/000313001300339897.

93. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ. 1999;319(7211):670-4.

94. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979;6(2):65-70.

95. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ. 2025;389:e081123.

96. Horn JL. A rationale and test for the number of factors in factor analysis. Psychometrika. 1965;30:179-85.

97. Horwitz RI, Feinstein AR. Alternative analytic methods for case-control studies of estrogens and endometrial cancer. N Engl J Med. 1978;299(20):1089-94.

98. Howe LD, Hargreaves JR, Huttly SRA. Issues in the construction of wealth indices for the measurement of socio-economic position in low-income countries. Emerg Themes Epidemiol. 2008;5:3.

99. Hróbjartsson A, Thomsen AS, Emanuelsson F, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ. 2012;344:e1119.

100. Hyndman RJ, Fan Y. Sample quantiles in statistical packages. American Statistician. 1996;50:361-5 — as documented in R's help page ?quantile (R stats package, opened locally with Rscript).

101. IBM. SAVE subcommand (FACTOR command). IBM SPSS Statistics 23 documentation.

102. ICMR-NCDIR. Report of National Cancer Registry Programme 2020 (2012-2016). Bengaluru: ICMR-NCDIR; 2020.

103. Illowsky B, Dean S. Introductory Statistics. Houston (TX): OpenStax; 2013. §13.4 Test of Two Variances.

104. International Committee of Medical Journal Editors. Recommendations: Preparing a Manuscript for Submission (Statistics).

105. International Conference on Harmonisation. ICH Topic E9: Statistical Principles for Clinical Trials. Note for guidance CPMP/ICH/363/96. London: EMEA; September 1998.

106. International Institute for Population Sciences (IIPS). National Family Health Survey (NFHS-5) 2019-21: survey description. Mumbai: IIPS; accessed 2026-09-28.

107. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? JAMA. 1994;271(9):703-7.

108. Kendall MG. A new measure of rank correlation. Biometrika. 1938;30(1-2):81-93.

109. Kim HY. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. Restor Dent Endod. 2017;42(2):152-5. doi:10.5395/rde.2017.42.2.152

110. Kish L. A procedure for objective respondent selection within the household. J Am Stat Assoc. 1949;44(247):380-7.

111. Knol MJ, VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction. Int J Epidemiol. 2012;41(2):514-20.

112. KoboToolbox. Data collection, management and visualization platform.

113. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-63.

114. Korstjens I, Moser A. Series: Practical guidance to qualitative research. Part 4: Trustworthiness and publishing. Eur J Gen Pract. 2018;24(1):120-4.

115. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. Crit Care Med. 2007;35(9):2052-6.

116. Kubavat AH, Mittal R, Patel PM, Jarsaniya DH, Pawar PR, et al. A clinical trial to assess the immunogenicity and safety of Inactivated Influenza Vaccine (Whole Virion) IP (Pandemic Influenza (H1N1) 2009 Monovalent Vaccine; VaxiFlu-S) in healthy Indian adult population. J Postgrad Med. 2011;57(2):102-8. doi:10.4103/0022-3859.81860

117. Kuo CL, Duan Y, Grady J. Unconditional or conditional logistic regression model for age-matched case-control data? Front Public Health. 2018;6:57.

118. Kwak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. Korean J Anesthesiol. 2017;70(2):144-56. doi:10.4097/kjae.2017.70.2.144

119. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.

120. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol. 2014;43(6):1969-85.

121. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. Am J Epidemiol. 2009;170(2):244-56.

122. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med. 1988;318(26):1728-33. doi:10.1056/NEJM198806303182605

123. Levin ML. The occurrence of lung cancer in man. Acta Unio Int Contra Cancrum. 1953;9(3):531-41.

124. Litvinov SK, Lobanov AV, Peregudov AN. [The cluster method in conducting epidemiological research]. Zh Mikrobiol Epidemiol Immunobiol. 1986;(11):78-84. PMID 3799117.

125. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health. 2002;23:151-69. doi:10.1146/annurev.publhealth.23.100901.140546

126. Lumley T. Analysis of complex survey samples. Journal of Statistical Software. 2004;9(8):1-19. doi:10.18637/jss.v009.i08

127. Lumley T. subset.survey.design: Subset of survey. R package survey documentation.

128. Lumley T. svydesign: Survey sample analysis. R package survey documentation.

129. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics. 2001;57(1):126-34. doi:10.1111/j.0006-341x.2001.00126.x

130. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol. 2010;5(9):1315-6.

131. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959;22(4):719-48.

132. Mathur MB, Ding P, Riddell CA, VanderWeele TJ. Website and R package for computing E-values. Epidemiology. 2018;29(5):e45-e47.

133. Mathur P, Kulothungan V, Leburu S, et al. National noncommunicable disease monitoring survey (NNMS) in India: Estimating risk factor prevalence in adult population. PLoS One. 2021;16(3):e0246712. doi:10.1371/journal.pone.0246712

134. Matthews JN, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research. BMJ. 1990;300(6719):230-5. doi:10.1136/bmj.300.6719.230

135. McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. J Clin Epidemiol. 2014;67(3):267-77.

136. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153-7. doi:10.1007/BF02295996

137. McNutt LA, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. Am J Epidemiol. 2003;157(10):940-3.

138. Meghani A, et al. Optimizing the Health Management Information System in Uttar Pradesh, India: implementation insights and key learnings. Glob Health Sci Pract. 2022.

139. Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. Am J Epidemiol. 1989;129(1):125-37.

140. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res. 2010;19(4):539-49.

141. Mugali RR, et al. Improving immunization in Afghanistan: results from a cross-sectional community-based survey to assess routine immunization coverage. BMC Public Health. 2017;17:290. doi:10.1186/s12889-017-4193-z

142. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med J. 2012;24(3):69-71.

143. Naaz F, Nayak BP, Panigrahi S, Mohakud NK. Psychosocial and economic burden on families of children with cerebral palsy. Cureus. 2025;17(1):e76794. doi:10.7759/cureus.76794

144. Nagele P. Misuse of standard error of the mean (SEM) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. Br J Anaesth. 2003;90(4):514-6. doi:10.1093/bja/aeg087

145. Nemeth R. Respondent selection within the household - a modification of the Kish grid. Conference paper, Ossiach (undated).

146. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med. 1998;17(8):857-72.

147. NIST/SEMATECH e-Handbook of Statistical Methods, §1.3.5.10 Levene Test for Equality of Variances. n.d.

148. NIST/SEMATECH e-Handbook of Statistical Methods, §1.3.5.11 Measures of Skewness and Kurtosis.

149. NIST/SEMATECH e-Handbook of Statistical Methods, §1.3.5.16 Kolmogorov-Smirnov Goodness-of-Fit Test. n.d.

150. NIST/SEMATECH e-Handbook of Statistical Methods, §1.3.5.7 Bartlett's Test. n.d.

151. NIST/SEMATECH e-Handbook of Statistical Methods, §6.3.3.1 'Counts Control Charts'. National Institute of Standards and Technology. n.d.

152. NIST/SEMATECH e-Handbook of Statistical Methods, §7.1.6 What are outliers in the data?

153. NIST/SEMATECH e-Handbook of Statistical Methods, §7.2.4 ‘Does the proportion of defectives meet requirements?’. National Institute of Standards and Technology. n.d.

154. Norman G. Likert scales, levels of measurement and the “laws” of statistics. Adv Health Sci Educ Theory Pract. 2010;15(5): 625-32. doi:10.1007/s10459-010-9222-y

155. O’Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med. 2014;89(9):1245-51.

156. O’Brien RM. A caution regarding rules of thumb for variance inflation factors. Qual Quant. 2007;41(5):673-90. doi:10.1007/s11135-006-9018-6

157. ODK. Collect data anywhere.

158. Office of the Registrar General & Census Commissioner, India. Census of India website.

159. Office of the Registrar General & Census Commissioner, India. Sample Registration System Statistical Report 2020. New Delhi: ORGI.

160. Office of the Registrar General, India. Report on Medical Certification of Cause of Death 2022. New Delhi: ORGI.

161. Office of the Registrar General, India. Vital Statistics of India based on the Civil Registration System 2022. New Delhi: ORGI; 2025.

162. Ozarda Y. Reference intervals: current status, recent developments and future considerations. Biochem Med (Zagreb). 2016;26(1):5-16.

163. Page MJ, Higgins JPT, Sterne JAC. Chapter 13: Assessing risk of bias due to missing results in a synthesis. In: Cochrane Handbook for Systematic Reviews of Interventions, current version. Cochrane.

164. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.

165. Pang M, Kaufman JS, Platt RW. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. Stat Methods Med Res. 2016;25(5):1925-37 (epub 2013).

166. Pearce N. Analysis of matched case-control studies. BMJ. 2016;352:i969.

167. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol. 1995;48(12):1503-10.

168. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373-9.

169. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. Stat Med. 2007;26(25):4544-62.

170. Pewsner D, Battaglia M, Minder C, Marx A, Bucher HC, Egger M. Ruling a diagnosis in or out with “SpPIn” and “SnNOut”: a note of caution. BMJ. 2004;329(7459):209-13.

171. Piaggio G, Elbourne DR, Pocock SJ, Evans SJ, Altman DG. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. JAMA. 2012;308(24):2594-604.

172. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics. 1975;31(1):103-15.

173. Population Education (Population Connection). What are the different types of population pyramids? n.d.

174. Pourhoseingholi MA, Vahedi M, Rahimzadeh M. Sample size calculation in medical studies. Gastroenterol Hepatol Bed Bench. 2013;6(1):14-7.

175. R Core Team. AIC: Akaike’s An Information Criterion. R documentation, package stats.

176. R Core Team. cor.test: Test for Association/Correlation Between Paired Samples. R package stats documentation (R 4.x).

177. R Core Team. plot.lm: Plot Diagnostics for an lm Object. R stats package documentation.

178. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2024 (version 4.3.3 used).

179. R Core Team. varimax/promax — Rotation methods for factor analysis. R package stats documentation.

180. Rayson P, Berridge D, Francis B. Extending the Cochran rule for the comparison of word frequencies between corpora. JADT 2004: 7es Journées internationales d’Analyse statistique des Données Textuelles. 2004.

181. Razali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics. 2011;2(1):21-33.

182. Reed GF, Lynn F, Meade BD. Use of coefficient of variation in assessing variability of quantitative assays. Clin Diagn Lab Immunol. 2002;9(6):1235-9. doi:10.1128/cdli.9.6.1235-1239.2002

183. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ. 2011;342:d549.

184. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med. 2019;38(7):1276-96.

185. Robertson SE, Valadez JJ. Global review of health care surveys using lot quality assurance sampling (LQAS), 1984-2004. Soc Sci Med. 2006;63(6):1648-60. doi:10.1016/j.socscimed.2006.04.011

186. Robinson WS. Ecological correlations and the behavior of individuals. Am Sociol Rev. 1950;15:351-357. doi:10.2307/2087176

187. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

188. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25(1):127-41. doi:10.1002/sim.2331

189. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006;174(4):469-76.

190. Sackett DL. Bias in analytic research. J Chronic Dis. 1979;32(1-2):51-63.

191. Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. Anesth Analg. 2018;126(5): 1763-8.

192. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. Biometrics. 1983;39(2):499-503.

193. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332.

194. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet. 2002;359(9306): 614-8.

195. Scott NW, McPherson GC, Ramsay CR, Campbell MK. The method of minimization for allocation to clinical trials. a review. Control Clin Trials. 2002;23(6):662-74.

196. Senn S. Testing for baseline balance in clinical trials. Stat Med. 1994;13(17):1715-26.

197. Shah D. Healthy worker effect phenomenon. Indian J Occup Environ Med. 2009;13(2):77-9.

198. Shao J. Lecture 24: Sample median and quantiles. STAT 610, University of Wisconsin-Madison. n.d.

199. Shmueli G. To explain or to predict? Statistical Science. 2010;25(3):289-310.

200. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420-8.

201. Singh PK, Tripathi A, Gupta A, Yadav SK. Overview of web-based public health surveillance system in Uttarakhand: Integrated health information platform under integrated disease surveillance programme. Indian J Forensic Community Med. 2023;10(1):1-3. doi:10.18231/j.ijfcm.2023.001

202. Spearman C. The proof and measurement of association between two things. Am J Psychol. 1904;15:72-101. Reprinted Int J Epidemiol. 2010;39(5):1137-50.

203. StataCorp. Stata Base Reference Manual: kappa — Interrater agreement.

204. StataCorp. Stata Base Reference Manual: nbreg — Negative binomial regression. College Station, TX: Stata Press; 2023.

205. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. doi:10.1136/bmj.b2393

206. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54(8):774-81.

207. Streiner DL. Starting at the beginning: an introduction to coefficient alpha and internal consistency. J Pers Assess. 2003;80(1):99-103.

208. Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci. 2010;25(1):1-21. doi:10.1214/09-STS313

209. Sturges HA. The choice of a class interval. J Am Stat Assoc. 1926;21(153):65-66. doi:10.1080/01621459.1926.10502161

210. Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol. 2008;167(4):492-9.

211. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol. 1996;49(8):907-16. doi:10.1016/0895-4356(96)00025-x

212. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 2011;2:53-5.

213. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34-42.

214. The DHS Program. Guide to DHS Statistics: Analyzing DHS data. ICF, n.d.

215. The DHS Program. India: Standard DHS (NFHS-5), 2019-21. Implementing organisation IIPS, Mumbai.

216. The WHOQOL Group. Development of the World Health Organization WHOQOL-BREF quality of life assessment. Psychol Med. 1998;28(3):551-8.

217. Therneau TM. survival: Survival Analysis. R package documentation, cox.zph (citing Grambsch P, Therneau T. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 1994;81:515-26).

218. Therneau TM. survival: Survival Analysis. R package documentation, coxph.

219. Therneau TM. survival: Survival Analysis. R package documentation, tmerge.

220. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care. 2007;19(6):349-57.

221. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Stratification for confounding–part 1: the Mantel-Haenszel formula. Nephron Clin Pract. 2010;116(4):c317-21.

222. Tukey JW. Exploratory Data Analysis. Addison-Wesley; 1977 — as referenced by R's help page ?boxplot.stats (grDevices), opened locally.

223. van Buuren S. Flexible Imputation of Missing Data. 2nd ed. Chapman & Hall/CRC; 2018. Section 2.3.

224. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.

225. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. PLoS Med. 2007;4(10):e297.

226. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. Ann Intern Med. 2017;167(4):268-74.

227. Vickers AJ, Altman DG. Statistics notes: Analysing controlled trials with baseline and follow up measurements. BMJ. 2001;323(7321):1123-4.

228. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol. 2007;165(6):710-8.

229. von Hippel PT. Mean, median, and skew: correcting a textbook rule. Journal of Statistics Education. 2005;13(2).

230. Vyas S, Kumaranayake L. Constructing socio-economic status indices: how to use principal components analysis. Health Policy Plan. 2006;21(6):459-68.

231. Welch HG, Black WC. Overdiagnosis in cancer. J Natl Cancer Inst. 2010;102(9):605-13.

232. Westfall PH. Kurtosis as peakedness, 1905-2014. R.I.P. Am Stat. 2014;68(3):191-5. doi:10.1080/00031305.2014.917055

233. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. Am J Epidemiol. 2013;177(4):292-8. doi:10.1093/aje/kws412

234. Whiting PF, Rutjes AW, Westwood ME, Mallett S; QUADAS-2 Steering Group. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol. 2013;66(10):1093-104.

235. WHO Child Growth Standards. Height-for-age BOYS, 2 to 5 years (z-scores). Geneva: World Health Organization.

236. WHO Child Growth Standards. Weight-for-age BOYS, birth to 5 years (z-scores). Geneva: World Health Organization.

237. WHO Multicentre Growth Reference Study Group. WHO Child Growth Standards based on length/height, weight and age. Acta Paediatr Suppl. 2006;450:76-85. doi:10.1111/j.1651-2227.2006.tb02378.x

238. Wickham H, Stryjewski L. 40 years of boxplots. Working paper; 2011.

239. Woolf B. On estimating the relation between blood group and disease. Ann Hum Genet. 1955;19(4):251-3. doi:10.1111/j.1469-1809.1955.tb01348.x

240. World Bank Microdata Library. India - National Family Survey 2019-2021 (IND_2019_DHS_v01_M), sampling procedure. 2022.

241. World Bank. Age dependency ratio (% of working-age population), SP.POP.DPND. World Development Indicators metadata glossary. n.d.

242. World Health Organization. Data quality review: a toolkit for facility data quality assessment. Module 3: Site assessment of data quality. Version update December 2020. Geneva: WHO; 2020.

243. World Health Organization. Estimating mortality from COVID-19: scientific brief. 4 August 2020.

244. World Health Organization. Global Health Observatory Indicator Metadata Registry: Underweight among children under 5 years of age (IMR 27). n.d.

245. World Health Organization. Global Health Observatory: Stunting in children under 5 (JME) (UNICEF-WHO-WB). n.d.

246. World Health Organization. Global Health Observatory: Wasting prevalence among children under 5 years of age (% weight-for-height <-2 SD) (IMR 302). n.d.

247. World Health Organization. Guideline on haemoglobin cutoffs to define anaemia in individuals and populations. Geneva: WHO; 2024. (Summary of recommendations as published by Guideline Central with WHO.)

248. World Health Organization. Hypertension. Fact sheet, 25 September 2025.

249. World Health Organization. Indicator Metadata Registry: Infant mortality rate (IMR ID 1).

250. World Health Organization. Indicator Metadata Registry: Maternal mortality ratio (IMR ID 26).

251. World Health Organization. Indicator Metadata Registry: Neonatal mortality rate (IMR ID 67).

252. World Health Organization. Indicator Metadata Registry: Perinatal mortality (IMR ID 10325).

253. World Health Organization. Indicator Metadata Registry: Total fertility rate (IMR ID 123).

254. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32-5.

255. Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. J Stat Softw. 2008;27(8). (pscl package vignette 'countreg', opened from the installed package source /usr/lib/R/site-library/pscl/doc/countreg.Rnw)

256. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. JAMA. 1998;280(19):1690-1.

257. Zou G. A modified Poisson regression approach to prospective studies with binary data. Am J Epidemiol. 2004;159(7):702-6.

Feedback and new versions

If something in this book is wrong, unclear, or could be better, please write to **contact@drharshmaheshwari.com**. Put the version and the section in the subject line (for example: *Statistics v3.1, §7.3*). Say what the text says, what you think it should say and, if you can, the source.

Corrections are made in new versions. The version number and the date of the last update are on the copyright page, and the newest version is at **drharshmaheshwari.com**.

Scan for the newest version

drharshmaheshwari.com

contact@drharshmaheshwari.com

Version 3.1 · 28 September 2026

AN MD COMMUNITY MEDICINE COMPANION

Statistics: From First Principles to Regression

Most statistics books hand you formulas to memorise. This one builds every idea from the one before it, from what a variable is to regression, survival analysis and meta-analysis, so that by the time a formula appears it feels like the only reasonable answer.

- Seventeen chapters, one unbroken chain: nothing is used before it is taught.
- Worked examples from Indian community medicine: the under-five clinic, the CBHI survey, the district outbreak.
- Derivations and worked calculations you can follow with a calculator, plus optional R walk-throughs.
- 94 checkpoints, with worked answers at the back.



Dr. Harsh Maheshwari
MD Community Medicine Resident, AIIMS Raipur

contact@drharshmaheshwari.com
drharshmaheshwari.com

Newest version & feedback

Version 3.1 · 28 September 2026 · CC BY-NC-SA 4.0 · free to share, not to sell