Applied Statistics
Using R

Applied Statistics
Using R

A rigorous first course in statistical reasoning
for the life and health sciences

Safaa Dabagh

Department of Mathematics
Loyola Marymount University

Spring 2026 Edition

Applied Statistics Using R

First edition, Spring 2026.

Typeset with pdfLaTeX using the Palatino typeface.
All R code in this book was verified with R version 4.4 or later, using the tidyverse and ggplot2 packages.

Preface

Why this book

Every applied-statistics textbook I have taught from falls into one of two camps. Books in the first camp are brief enough to finish in a semester but too shallow to build real statistical intuition; students leave able to run `t.test()` but unable to say what a p-value means or why they should care. Books in the second camp treat statistics as a mathematical exercise whose biological context is almost an afterthought; students leave able to derive formulas but unable to choose the right one when confronted with a real dataset. I have wanted, for several semesters, a third option that does neither.

This book is that third option. It is written for the students I teach at Loyola Marymount University—motivated, mathematically capable undergraduates, most of them headed for careers in the life and health sciences—who deserve a rigorous treatment but also deserve to see every technique applied in the kind of biological question they care about. The mathematics is developed with care and, where it matters, derivation. The examples come from marine biology, ecology, pharmacology, neuroscience, and public health, chosen deliberately to mirror the research many of you will pursue. Every procedure is demonstrated in R using the tidyverse idioms most working scientists now use.

How to use this book

The chapters build on one another. Chapters 1–4 establish the descriptive-statistics toolkit and the language of estimation. Chapter 5 (Probability) and Chapter 6 (Hypothesis Testing) supply the inferential framework that the rest of the book applies. Chapters 7–10 develop inference for categorical data and the normal distribution. Chapters 11–13 are the heart of the book: inference about means, for one and two samples, and the remedies available when assumptions fail. Chapter 15 handles more-than-two groups with ANOVA. Chapters 16 and 17 close with correlation and simple linear regression.

Every chapter has the same structure: a motivating vignette, learning objectives, six or seven sections of narrative interleaved with Definition, Result, and Example boxes, one or more R Section boxes with complete code, at least one Watch Out box for a common mistake, a chapter summary with formulas and an R-function table, and a graduated set of exercises.

R setup

All R code in this book was verified with R version 4.4 or later, running `library(tidyverse)`. Before running any example, execute `set.seed(205)`: the code will then produce the numbers printed on the page. Datasets referenced in examples are catalogued in Appendix D.

Acknowledgements

This book grew out of four semesters of MATH 205 lecture notes and in-class worked examples, refined one conversation at a time by the students who pushed back on every argument. Their

curiosity and patience shaped the tone as much as any textbook I had on my shelf. I am grateful to them, and to the LMU Mathematics Department for the time to write it.

Safaa Dabagh

Los Angeles, Spring 2026

Contents

List of Figures

Part I

Foundations

1

Introduction to Statistics and Data

"Statistics is the grammar of science."

— Karl Pearson

Statistics is the science of learning from data: collecting it well, summarizing it honestly, and drawing conclusions with a calibrated sense of uncertainty. For biologists, psychologists, public-health researchers, and environmental scientists, it is not an optional tool. The questions you care about—does this drug work, is that population declining, how much variation is there in this trait, is the intervention safe—all translate into statistical questions before they translate into decisions.

This chapter lays the foundation. We distinguish populations from samples, define the kinds of variables you are likely to meet, discuss how data are collected (well, and badly), and place the whole enterprise in the context of the studies that follow throughout the book.

LEARNING OBJECTIVES

- Distinguish a population from a sample and a parameter from a statistic.
- Identify the two major types of studies (observational vs. experimental) and explain which can establish causation.
- Classify variables as categorical (nominal, ordinal) or numerical (discrete, continuous).
- Describe simple random sampling and explain why it is the gold standard for generalizable inference.
- Identify common sources of sampling bias and sketch how each would distort conclusions.

1.1 Population, Sample, Parameter, Statistic

DEFINITION – Population, sample, parameter, statistic

A *population* is the complete collection of entities about which we wish to draw conclusions (all adult female vervet monkeys in Awash National Park, every tablet produced on line 3 today). A *sample* is a subset of the population from which we collect data. A *parameter* is a numerical characteristic of the population (the mean foraging time μ, the defect proportion p). A *statistic* is a numerical characteristic of the sample ($\bar{x}$, $\hat{p}$).

Statistical inference is the process of drawing conclusions about a parameter based on a statistic. The entire rest of this book elaborates the tools for doing this honestly.

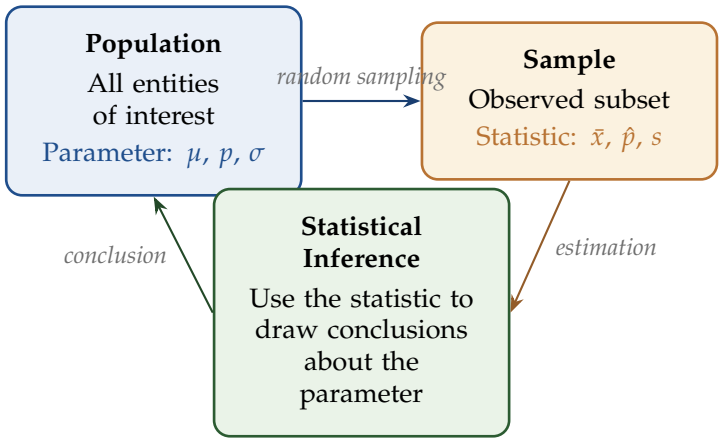


Figure 1.1: The statistical inference loop. Random sampling connects the population to the sample; statistics computed from the sample are used to estimate population parameters; inference closes the loop by reporting what the data say about the population.

1.2 Types of Studies

DEFINITION – Observational study vs. experiment

In an *observational study*, the researcher measures what happens without intervening. In an *experiment*, the researcher imposes a treatment on the subjects and measures the response. The defining feature of an experiment is *random assignment*: subjects are randomly placed into treatment groups so that background variables are balanced on average.

RESULT – Why experiments can establish causation

Random assignment breaks the link between the treatment and all other subject characteristics (including those the researcher has not measured). When a difference in outcomes then appears between groups, it cannot be explained by pre-existing differences. In observational studies, that argument is unavailable—which is why observational research can reveal associations but not causes.

WATCH OUT – Correlation is not causation

An association observed in an observational study can reflect any of: the putative cause, reverse causation, a lurking confounder, a selection effect, or chance. Experimental randomization eliminates the first three; nothing eliminates all four without careful design.

1.3 Types of Variables

DEFINITION – Variable types

A *variable* is a measured characteristic of the sampling units.

- **Categorical (qualitative)**: values are labels.
  - *Nominal*: no natural order (species name, blood type, treatment arm).
  - *Ordinal*: ordered but not numeric (strongly disagree $<$ disagree $<$ neutral $<$ agree; stage I $<$ stage II $<$ stage III cancer).
- **Numerical (quantitative)**: values are numbers.
  - *Discrete*: integer-valued (count of offspring).
  - *Continuous*: can take any value in an interval (body mass, enzyme concentration).

The variable type dictates the display (Chapter 2), the summary statistics (Chapter 3), and the inference procedure (Chapters 7 onward).

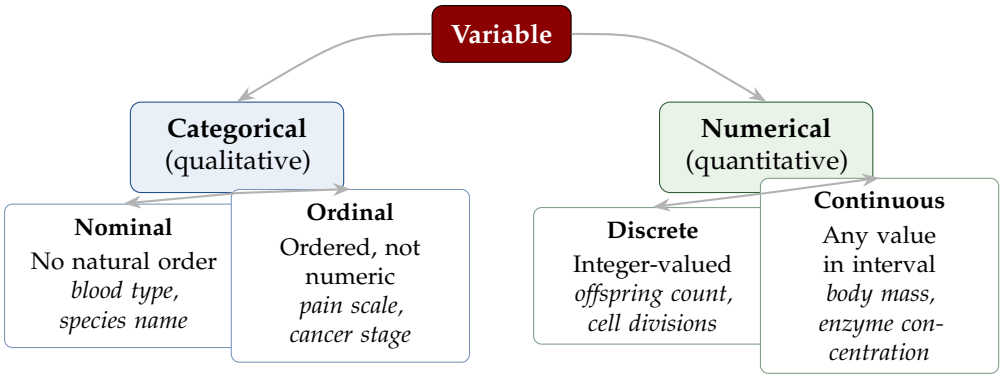


Figure 1.2: Taxonomy of variable types. Choosing the correct type for each variable is the first decision in any statistical analysis; it determines which displays, summaries, and tests are appropriate.

1.4 Random Sampling and Bias

DEFINITION – Simple random sample (SRS)

A *simple random sample* of size n is a sample drawn so that every subset of size n from the population is equally likely to be chosen. SRS is the gold standard because the sampling distribution of statistics under SRS has known, tractable form.

RESULT – Common biases

Four sampling biases appear frequently in applied work:

1. **Selection bias**: the sampling mechanism systematically excludes a portion of the population.
2. **Nonresponse bias**: respondents differ systematically from nonrespondents (online polls; volunteer samples).
3. **Response bias**: the measurement process itself distorts answers (leading questions; socially desirable responses).
4. **Voluntary-response bias**: people with strong opinions are overrepresented.

None can be fixed by increasing n.

1.5 Doing It in R

R SECTION – Loading a dataset and inspecting its variables

```r
library(tidyverse)
set.seed(205)

# The LMU class dataset (see Appendix D)
lmu <- read_csv("lmu_class_data_cleaned.csv")

glimpse(lmu)                    # variable types and first values
summary(lmu)                    # marginal summaries
table(lmu$major)                # frequency table for a categorical
```

Chapter Summary

Concept	Key idea
Population vs. sample	Population: all entities of interest. Sample: the observed subset.
Parameter vs. statistic	Parameter: numerical property of the population (μ, p). Statistic: numerical property of the sample ($\bar{x}$, $\hat{p}$).
Observational study	Measures without intervening; can show association, not causation.
Experiment	Random assignment $\Rightarrow$ causation defensible.
Variable types	Categorical (nominal, ordinal) or Numerical (discrete, continuous).
Simple random sample	Every subset of size n equally likely; gold standard.
Common biases	Selection, nonresponse, response, voluntary-response.

EXERCISES

1. For each variable, classify it as nominal, ordinal, discrete, or continuous: (a) blood type; (b) systolic blood pressure; (c) pain score on a 1–10 scale; (d) number of seeds per pod; (e) species of bird.

2. A newspaper reports that readers of its website approve of a new city policy by a 3:1 margin. What bias is most likely at play?

3. Explain, in two to three sentences, why a randomized experiment can establish causation when an observational study cannot, using an example of your choice.

4. A public-health official wants to estimate the mean systolic blood pressure of adults in a city of 2 million people. Describe a sampling scheme that would produce a simple random sample, and one that would produce a biased sample.

5. Distinguish between a parameter and a statistic using specific symbols (μ, $\bar{x}$, p, $\hat{p}$). Give an example of each in a clinical-trial context.

6. A marine biologist wants to estimate the mean body length of adult hammerhead sharks off the coast of Florida. She tags and measures 40 sharks caught in a commercial fishery. (a) Identify the population and the sample. (b) What parameter is she estimating and with what statistic? (c) What bias might arise from using fishery catch records?

7. A social psychologist randomly assigns 60 college students to two groups: one watches a 10-minute gratitude video, the other watches a neutral video. Both groups then complete a wellbeing questionnaire. (a) Is this an experiment or an observational study? (b) What is the treatment? (c) Can causation be inferred? Explain why or why not.

8. Explain why increasing the sample size from $n = 50$ to $n = 5000$ cannot fix a voluntary-response bias. What is the only way to eliminate sampling bias?

9. A hospital records the following for each patient: (a) diagnosis code (ICD-10); (b) days in hospital; (c) discharge status (home / rehab / deceased); (d) patient satisfaction score

(1–5 stars); (e) total charges (dollars). Classify each variable and state which graphical display you would use first.

10. (Challenge) A researcher finds that cities with more ice-cream shops have higher rates of drowning deaths. Propose at least two alternative explanations for this association that do not involve ice cream causing drowning. Which type of study could rule out your alternatives?

2

Displaying Data

"Above all else show the data."

— Edward R. Tufte

A good graph answers a scientific question at a glance; a bad graph hides the answer or, worse, suggests a wrong one. This chapter teaches you to build the standard displays for categorical and numerical data, to recognize when a particular display is appropriate, and to evaluate graphs critically.

LEARNING OBJECTIVES

- Build frequency and relative-frequency tables and the corresponding bar charts for categorical data.
- Build histograms and density plots for continuous numerical data and interpret their shape.
- Build scatterplots for bivariate numerical data.
- Identify common graphical pitfalls: misleading scales, 3-D effects, truncated axes.
- Produce publication-ready figures in R using ggplot2.

2.1 Displaying Categorical Data

DEFINITION – Frequency table, bar chart

A *frequency table* lists each category alongside the count of observations falling in it. A *relative frequency table* instead reports proportions. A *bar chart* plots categories on one axis and counts (or proportions) on the other with separate bars for each category.

Bar charts use bars whose heights encode counts. Pie charts encode the same information as angular segments; most readers find them harder to decode accurately, and we will prefer bar charts throughout the book.

2.2 Displaying Numerical Data: One Variable

DEFINITION – Histogram

Divide the range of the data into k equal-width intervals (*bins*). For each bin, count the number of observations that fall inside. A *histogram* is the bar chart of these counts. The choice of bin width is a judgment call: too few bins hide features, too many bins add visual noise.

Histograms reveal the *shape* of a distribution: is it symmetric or skewed, unimodal or multimodal, does it have heavy tails or outliers? All of these matter when you choose an inference procedure (Chapters 11–15).

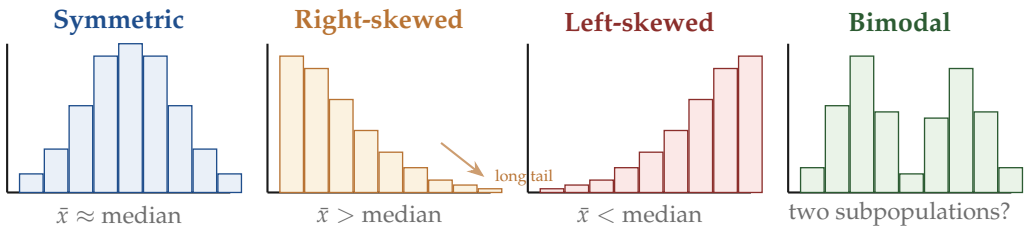


Figure 2.1: Four histogram shapes to recognize. Symmetric distributions are well described by the mean; skewed distributions are better summarized by the median and may need a transformation before parametric inference (Chapter 13). Bimodal distributions often signal a mixture of subpopulations worth investigating separately.

RESULT – Distribution shapes to recognize

- *Symmetric unimodal*: approximately mirror-image around the mode; often normal-looking.
- *Right-skewed*: long tail stretching to the right; common for income, body mass, lab values.
- *Left-skewed*: long tail to the left; less common.
- *Bimodal*: two distinct peaks; often signals a mixture of subpopulations.

WATCH OUT – Bar chart $\neq$ histogram

A bar chart displays counts for categorical data and leaves gaps between bars. A histogram displays counts for numerical data binned into adjacent intervals and has no gaps between bars. They are different plots for different data types; use the right one.

2.3 Displaying Bivariate Numerical Data

DEFINITION – Scatterplot

A *scatterplot* places one variable on each axis and plots a point per observation. It displays the *direction*, *form*, and *strength* of the association between the two variables, as well as any outliers.

Correlation (Chapter 16) and regression (Chapter 17) quantify what a scatterplot shows. Never report a correlation without first looking at the plot.

2.4 Doing It in R

R SECTION – Categorical: bar chart

```
library(tidyverse)

lmu |> ggplot(aes(x = major)) +
  geom_bar(fill = "#8B0000") +
  theme_minimal() +
  labs(x = NULL, y = "Number of students")
```

R SECTION – Numerical: histogram and density

```
lmu |> ggplot(aes(x = height_cm)) +
  geom_histogram(bins = 20, fill = "#EAF0F8",
                 colour = "#1F4E8C") +
  theme_minimal() +
  labs(x = "Height (cm)")

# A smooth density overlay
lmu |> ggplot(aes(x = height_cm)) +
  geom_histogram(aes(y = ..density..), bins = 20,
                 fill = "#EAF0F8", colour = "#1F4E8C") +
  geom_density(colour = "#8B0000", linewidth = 0.8) +
  theme_minimal()
```

R SECTION – Bivariate: scatterplot

```
lmu |> ggplot(aes(x = height_cm, y = weight_kg)) +
  geom_point(colour = "#8B0000", alpha = 0.7) +
  theme_minimal() +
  labs(x = "Height (cm)", y = "Weight (kg)")
```

Chapter Summary

Data type	Display	What it shows
Categorical	Bar chart	Counts / proportions per category
Numerical (1 var)	Histogram	Shape, center, spread, outliers
Numerical (1 var)	Boxplot	Five-number summary, outliers
Numerical (2 var)	Scatterplot	Direction, form, strength

Shape vocabulary: symmetric, right-skewed, left-skewed, bimodal

Bar chart $\neq$ histogram: gaps vs. no gaps; categorical vs. numerical

EXERCISES

1. Explain, in one sentence each, when to use a bar chart, a histogram, a boxplot, and a scatterplot.
2. Describe the shape of a histogram whose sample has (a) $\bar{x} = 50$, median $= 50$; (b) $\bar{x} = 80$, median $= 50$; (c) $\bar{x} = 30$, median $= 50$.
3. Sketch a bimodal distribution and describe a scientific scenario that might produce one.
4. Find a graph in a recent news article that you think is misleading. In one paragraph, explain what you would change and why.
5. Using the `lmu` dataset (Appendix D), produce a histogram of heights faceted by reported sex. Briefly describe the distributions.
6. A study measures cortisol levels (nmol/L) in 80 healthy adults. A histogram shows a strong right skew with a few values near 500 nmol/L. (a) Would a bar chart be appropriate for these data? Why or why not? (b) What transformation might make the histogram more symmetric?
7. Explain what the *bin width* of a histogram controls. What happens visually if you use too few bins? Too many bins? Sketch both extremes.
8. A bar chart shows that 60% of survey respondents prefer Brand A. The chart's y-axis starts at 55% rather than 0%. Explain why this is misleading and sketch a corrected version.
9. Using ggplot2, add a vertical line at the sample mean to a histogram of the `height_cm` variable in the `lmu` dataset. Describe what the line reveals about the distribution's symmetry.
10. A researcher plots seed mass (mg) against germination time (days) for 150 seeds. Describe what each of the following scatterplot features would tell you biologically: (a) a positive, linear, tight cloud; (b) a curved cloud that rises then flattens; (c) two separated clusters.

3

Describing Data

"If I had to boil down all of statistics to a single word, it would be 'variation'."

— paraphrase of G. E. P. Box

A histogram shows the full shape of a distribution; a scatterplot shows the full pattern of an association. Graphs are the natural starting point, but science demands numbers. This chapter develops the numerical summaries that compress a distribution into a handful of values: measures of *center* (where is the middle?) and measures of *spread* (how variable is the data?). Together, these summaries power every hypothesis test and confidence interval in the rest of this book.

LEARNING OBJECTIVES

- Compute and interpret the mean, median, and mode.
- Compute and interpret the variance, standard deviation, range, and interquartile range.
- Build a five-number summary and the corresponding boxplot.
- Identify outliers using the $1.5 \cdot \text{IQR}$ rule.
- Choose the appropriate summary statistics for a given distribution shape.

3.1 Measures of Center

DEFINITION – Mean, median, mode

For a sample $x_1, \ldots, x_n$,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \text{median} = \text{middle value of the sorted data.}$$

The *mode* is the value (or values) that occur most often.

The mean is the arithmetic average and is the summary that plays the starring role in most inferential procedures (because its sampling distribution is well-understood via the CLT). The median is the middle-ranked value and is resistant to outliers: adding one extreme point barely changes it. The mode is rarely used for continuous data but can be informative for discrete or multimodal data.

WATCH OUT – The mean is pulled toward the tail

For symmetric distributions, the mean and median are essentially equal. For right-skewed distributions (income, lab values), the mean exceeds the median because it is pulled by the long right tail. For left-skewed distributions, the mean is less than the median. If a summary statistic disagrees with your sense of the data, revisit the histogram—chances are the distribution is skewed and the mean is misleading.

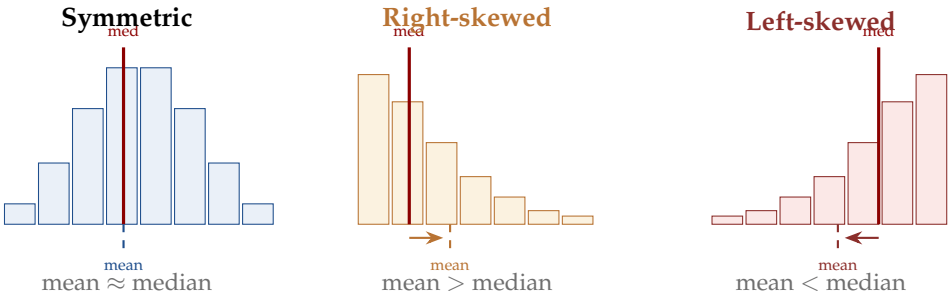


Figure 3.1: How distribution shape affects the relationship between mean and median. In a symmetric distribution they coincide. Skewness pulls the mean toward the longer tail while the median stays near the center of the data.

3.2 Measures of Spread

DEFINITION – Variance, standard deviation

The *sample variance* and *sample standard deviation* are

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad s = \sqrt{s^2}.$$

The divisor $n-1$ (not n) makes s^2 an unbiased estimator of the population variance σ^2—a correction called *Bessel's correction*.

The SD is in the same units as the data; the variance is in squared units and is convenient for theoretical work but less interpretable.

DEFINITION – Range, quartiles, IQR

The *range* is $\max(x) - \min(x)$. The *first quartile* Q_1 is the 25th percentile; the *third quartile* Q_3 is the 75th percentile; the *interquartile range* is $\text{IQR} = Q_3 - Q_1$.

The IQR measures the spread of the middle 50% of the data and is resistant to outliers.

3.3 The Five-Number Summary and Boxplots

DEFINITION – Five-number summary

The *five-number summary* of a sample is

$$\text{min},\ Q_1,\ \text{median},\ Q_3,\ \text{max}.$$

A *boxplot* is a visual rendering of the five-number summary: a rectangle spanning Q_1 to Q_3, a line at the median, and "whiskers" extending to the most extreme values within $1.5 \cdot \text{IQR}$ of the box. Points beyond the whiskers are flagged as potential outliers.

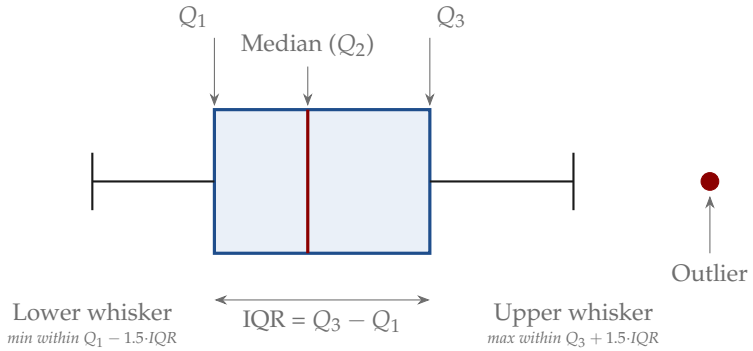


Figure 3.2: Anatomy of a boxplot. The box spans the IQR (Q_1 to Q_3); the red line is the median; whiskers reach to the most extreme non-outlier values; points beyond the whiskers are plotted individually as potential outliers.

RESULT – The $1.5 \cdot \text{IQR}$ rule for outliers

A value x is flagged as a potential outlier if

$$x < Q_1 - 1.5 \cdot \text{IQR} \quad \text{or} \quad x > Q_3 + 1.5 \cdot \text{IQR}.$$

This is a convention, not a law; outliers deserve investigation, not automatic deletion.

WATCH OUT – Outliers are not noise

An outlier is a measurement that falls far from the rest of the data. It might be a data-entry error, a genuine extreme specimen, or evidence of a hidden subpopulation. Deleting outliers without investigation is scientific malpractice. Flag, investigate, and justify any decision to remove, transform, or retain them.

3.4 Choosing Summaries That Fit the Data

For symmetric distributions without outliers, report $\bar{x}$ and s. For skewed distributions or data with outliers, report the median and IQR. A fuller practice—the one we adopt throughout this book—is to report both pairs whenever space allows, along with the sample size and the range.

3.5 Doing It in R

R SECTION – Summary statistics in one call

```r
library(tidyverse)
set.seed(205)

lmu <- read_csv("lmu_class_data_cleaned.csv")

lmu |> summarise(
  n        = n(),
  mean_h   = mean(height_cm),
  sd_h     = sd(height_cm),
  median_h = median(height_cm),
  IQR_h    = IQR(height_cm),
  Q1_h     = quantile(height_cm, 0.25),
  Q3_h     = quantile(height_cm, 0.75)
)
```

R SECTION – Boxplots, one or many groups

```r
# Single variable
lmu |> ggplot(aes(y = height_cm)) +
  geom_boxplot(fill = "#EAF0F8", colour = "#1F4E8C") +
  theme_minimal()

# By group -- side-by-side
lmu |> ggplot(aes(x = sex, y = height_cm, fill = sex)) +
  geom_boxplot(alpha = 0.8) +
  theme_minimal()
```

Chapter Summary

Statistic	Formula	Resistant?
Mean	$\bar{x} = \frac{1}{n}\sum x_i$	No
Median	Middle value (sorted)	Yes
Variance	$s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$	No
SD	$s = \sqrt{s^2}$	No
IQR	$Q_3 - Q_1$	Yes
Outlier rule	$x < Q_1 - 1.5 \cdot \text{IQR}$ or $x > Q_3 + 1.5 \cdot \text{IQR}$	—

When to use: symmetric $\rightarrow \bar{x}$ and s; skewed / outliers $\rightarrow$ median and IQR

EXERCISES

1. For the sample $(3, 6, 7, 9, 12, 13, 20, 22)$, compute mean, median, SD, IQR, and the five-number summary.

2. Explain why the median is "resistant" to an outlier and the mean is not, using the sample in the previous exercise after replacing 22 with 200.

3. Would you expect the mean to exceed, equal, or be smaller than the median for (a) household income; (b) human resting heart rate; (c) time to complete a task? Justify each answer.

4. Construct a sample of size $n = 10$ where the $1.5 \cdot \text{IQR}$ rule flags no outliers but one observation is visually clearly anomalous. What does this tell you about relying on rules of thumb?

5. Using the `lmu` dataset, produce boxplots of height by sex. Report the five-number summary for each group and comment on the comparison.

6. Ecologists measured carapace width (mm) for $n = 12$ blue crabs: $62, 74, 81, 55, 88, 77, 90, 63, 71, 84, 59, 95$. (a) Compute the mean and SD. (b) Apply the $1.5 \cdot \text{IQR}$ rule. (c) Compute the coefficient of variation $\text{CV} = s/\bar{x}$ and explain what it means.

7. Two labs measure enzyme activity (μmol/min) in the same tissue. Lab A reports mean $= 42$, SD $= 8$. Lab B reports mean $= 42$, SD $= 2$. (a) Which lab's measurements are more precise? (b) If you pooled all measurements, would the combined SD be larger or smaller than Lab A's? Explain without computing.

8. A dataset has $Q_1 = 18$, median $= 24$, $Q_3 = 31$. (a) Compute the IQR and the whisker fences. (b) Classify each value as inside-fence or outlier: $5, 12, 45, 52$. (c) Sketch the boxplot.

9. Using R, compute the skewness of the `height_cm` variable in the `lmu` dataset using the formula $g_1 = \frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i - \bar{x}}{s}\right)^3$. Is the distribution left-skewed, symmetric, or right-skewed?

10. (Challenge) Prove algebraically that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Then explain in one sentence why this means the mean is the "balance point" of the distribution.

4

Estimating with Uncertainty

"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."

— Sir Ronald A. Fisher

Chapters 1–3 equipped you to collect, display, and summarize data. Every summary you compute is a *statistic*: a number from a particular sample. Another sample would have produced a slightly different number. Before we can test hypotheses (Chapter 6) or compare means (Chapters 11–12), we need a language for the uncertainty in our statistics. That language has two central ideas—the *sampling distribution* and the *standard error*—and one indispensable summary of uncertainty, the *confidence interval*.

LEARNING OBJECTIVES

- Distinguish a parameter, an estimator, and an estimate.
- Define the sampling distribution of a statistic and explain what its shape, centre, and spread depend on.
- Compute the standard error of a sample mean and a sample proportion.
- Interpret a 95% confidence interval correctly, in both frequentist terms ("capture rate") and scientific terms.
- Use `rnorm()`, `mean()`, and the bootstrap to simulate sampling distributions and verify theoretical results.

4.1 Estimators and Sampling Distributions

DEFINITION – Estimator, estimate

An *estimator* is a procedure—a formula applied to a sample— that produces a number intended to approximate an unknown population parameter. The resulting number is the *estimate*. For example, $\bar{X}$ is an estimator of μ; the realised value $\bar{x}$ computed from a specific sample is an estimate.

The estimator is the recipe; the estimate is the dish you get when you follow it. Both words matter: statistical properties (unbiasedness, variance) belong to estimators, but your paper reports estimates.

DEFINITION – Sampling distribution

The *sampling distribution* of a statistic is the probability distribution of the values the statistic would take across all possible samples of the same size from the same population. It is not observed; it is an idealized object we reason about.

The sampling distribution of $\bar{X}$ is the central object of parametric inference. Chapter 10 (Normal Distribution) and the Central Limit Theorem describe its shape. Two facts carry most of the weight:

$$\mathrm{E}(\bar{X}) = \mu, \qquad \mathrm{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

The first says the sample mean is unbiased. The second says larger samples produce less-variable means—precisely at rate $1/\sqrt{n}$.

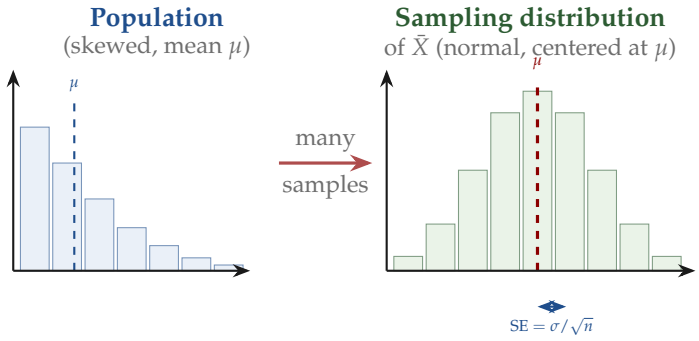


Figure 4.1: From a skewed population (left) to its sampling distribution (right). Even when individual observations are skewed, the distribution of sample means $\bar{X}$ is approximately normal and centered at the true population mean μ. The spread of the sampling distribution shrinks as n grows: $\mathrm{SE}(\bar{X}) = \sigma/\sqrt{n}$.

4.2 The Standard Error

DEFINITION – Standard error

The *standard error* of an estimator is the standard deviation of its sampling distribution. For the sample mean,

$$\mathrm{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}} \quad \text{estimated by} \quad \widehat{\mathrm{SE}}(\bar{X}) = \frac{s}{\sqrt{n}}.$$

For a sample proportion, $\mathrm{SE}(\hat{p}) = \sqrt{p(1-p)/n} \approx \sqrt{\hat{p}(1-\hat{p})/n}$.

The SE answers the question "how much would this estimate wobble if I ran the study again?". It is the scale on which every confidence interval is built and every test statistic is standardized.

4.3 Confidence Intervals

DEFINITION – Confidence interval

A $100(1-\alpha)\%$ confidence interval for a parameter θ is a random interval, computed from the sample, whose endpoints depend on the data and whose construction procedure captures the true θ in $100(1-\alpha)\%$ of its hypothetical repetitions.

For a population mean with known σ, the $100(1-\alpha)\%$ CI is

$$\bar{x} \pm z^*_{\alpha/2} \cdot \sigma / \sqrt{n}.$$

When σ is unknown, z^* is replaced by t^*_{n-1} and σ by s (Chapter 11). The same template—estimate $\pm$ critical value $\times$ standard error—recurs throughout the book.

WATCH OUT – What "95% confident" does not mean

A 95% confidence interval is *not*:

- a statement that μ lies in the interval with probability 0.95;
- a statement about where 95% of the *data* lies;
- a statement about future samples.

It is a statement about the long-run behaviour of the *procedure*: 95% of the intervals built this way, over repeated sampling, would cover the true parameter. The parameter does not move; the interval does.

4.4 Precision, Sample Size, and the $\sqrt{n}$ Penalty

Because $\mathrm{SE}(\bar{X}) = \sigma/\sqrt{n}$, halving the standard error requires *quadrupling* the sample size. Biological studies frequently encounter this $\sqrt{n}$ ceiling: going from $n = 25$ to $n = 100$ halves the SE, but going from $n = 100$ to $n = 200$ reduces it only by a factor of $\sqrt{2}$. Sample-size calculations hinge on this quadratic cost.

4.5 Doing It in R

R SECTION – Simulating a sampling distribution

```
library(tidyverse)
set.seed(205)

# Draw 10,000 samples of size n = 40 from a population with mean 50, sd
    10.
# For each sample, compute the mean. The distribution of these means is
# the (empirical) sampling distribution of x-bar.
means <- replicate(10000, mean(rnorm(40, mean = 50, sd = 10)))

tibble(xbar = means) |>
  ggplot(aes(x = xbar)) +
  geom_histogram(bins = 40, fill = "#EAF3E8", colour = "#2C5B32") +
  theme_minimal() +
  labs(x = "Sample mean", title = "Empirical sampling distribution of
    x-bar")

# Theory predicts mean 50, SD = 10 / sqrt(40) = 1.58
c(mean(means), sd(means))
```

R SECTION – A confidence interval in two lines

```
x <- rnorm(30, mean = 100, sd = 15)
t.test(x, conf.level = 0.95)$conf.int
```

Chapter Summary

Quantity	Formula	Notes
SE of $\bar{X}$ (σ known)	$\sigma/\sqrt{n}$	Exact
SE of $\bar{X}$ (σ unknown)	$s/\sqrt{n}$	Estimated
SE of $\hat{p}$	$\sqrt{\hat{p}(1-\hat{p})/n}$	Estimated
General CI template	estimate $\pm\, z^* \cdot$ SE	$z^* = 1.96$ for 95%
$\sqrt{n}$ penalty	To halve SE, quadruple n	Costly!

95% CI interpretation: the procedure captures μ in 95% of its uses.

NOT: "there is a 95% probability that μ lies in this interval."

EXERCISES

1. Explain, in one sentence each, the difference between (a) parameter and estimator; (b) estimator and estimate; (c) standard deviation and standard error.
2. A sample of $n = 64$ observations has $\bar{x} = 70$, $s = 24$. Compute the standard error of $\bar{X}$ and a 95% confidence interval for μ.
3. In the previous exercise, how large would n need to be to halve the SE (keeping s constant)?
4. Write the correct frequentist interpretation of a 90% confidence interval in one sentence. Then write a plain- English scientific sentence a researcher could paste into a paper.
5. Use `rnorm()` to simulate 10,000 samples of size $n = 25$ from $N(100, 15)$. For each sample, construct a 95% CI using `t.test()`. What fraction of the intervals cover the true $\mu = 100$? Does it match the nominal coverage?
6. A nutritionist estimates mean daily calcium intake in adult women. With $n = 100$ and $s = 310$ mg, the 95% CI is $(812, 935)$ mg. (a) Interpret this interval for a non-statistician. (b) How would the interval change if n were increased to 400 (assuming the same s)?
7. Two researchers study the same population. Researcher A uses $n = 50$ and Researcher B uses $n = 200$. (a) How much narrower is B's 95% CI compared to A's (assuming equal s)? (b) If A reports a CI of $(18.2, 24.6)$, what is the approximate width of B's CI?
8. Explain why the sampling distribution of $\bar{X}$ is narrower than the distribution of individual observations. Use the formula $\text{SE}(\bar{X}) = \sigma/\sqrt{n}$ and a concrete example with $\sigma = 10$, $n = 25$.
9. A wildlife biologist estimates the mean home-range area of wolves. She collects $n = 36$ GPS-tracked animals and finds $\bar{x} = 142$ km^2, $s = 48$ km^2. (a) Construct a 99% CI. (b) The literature reports a mean of 120 km^2. Is that consistent with your interval?
10. (Challenge) Using simulation in R, demonstrate that the width of a 95% CI scales as $1/\sqrt{n}$ by plotting average CI width against n for $n = 10, 25, 50, 100, 200, 500$. What does the shape of the curve tell you about the diminishing returns of larger samples?

5

Probability

"Probability theory is nothing but common sense reduced to calculation."

— Pierre-Simon Laplace

Every inferential procedure in the chapters that follow rests on probability. A p-value is a probability. A confidence level is a probability. The sampling distribution of $\bar{X}$ is described by a probability. Before we can reason about whether an observation is unusual under some hypothesis, we need a language for describing uncertainty. That language is probability, and this chapter gives a working introduction for applied scientists.

LEARNING OBJECTIVES

- Define an event, a sample space, and the probability of an event, and compute probabilities for simple discrete models.
- Apply the addition and multiplication rules for events.
- Compute conditional probabilities and apply Bayes' theorem to update beliefs in light of evidence.
- Check whether two events are independent.
- Define a discrete random variable and compute its expected value and variance.

5.1 Events, Sample Spaces, and Probabilities

DEFINITION – Sample space, event, probability

A *random experiment* is any process whose outcome is not known in advance. The *sample space* S is the set of all possible outcomes. An *event* is a subset of S. A *probability* assigns to each event $A \subseteq S$ a number $P(A) \in [0, 1]$ satisfying

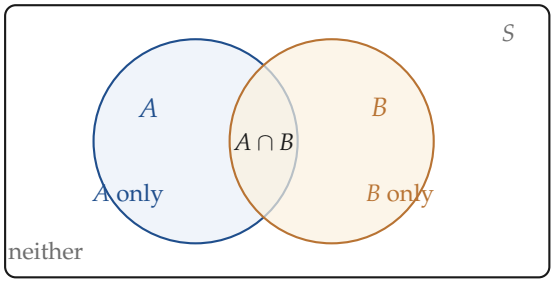
1. $P(S) = 1$,
2. $P(A) \geq 0$ for every event A,

3. $P(A \cup B) = P(A) + P(B)$ whenever A and B are mutually exclusive ($A \cap B = \emptyset$).

RESULT – Basic probability rules

For any events A and B:

$$P(A^c) = 1 - P(A),$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$
$$P(\emptyset) = 0.$$



$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Figure 5.1: A Venn diagram for two events A and B in a sample space S. The union $A \cup B$ covers both circles; the intersection $A \cap B$ is the overlapping region. The general addition rule subtracts the intersection once to avoid double-counting it.

5.2 Conditional Probability and Independence

DEFINITION – Conditional probability

For events A, B with $P(B) > 0$, the *conditional probability of A given B* is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Conditional probability rescales probabilities to the subset of the sample space in which B has occurred. The multiplication rule $P(A \cap B) = P(A \mid B)\, P(B)$ is just this definition rearranged.

DEFINITION – Independence

Events A and B are *independent* if $P(A \cap B) = P(A)P(B)$, or equivalently, $P(A \mid B) = P(A)$.

WATCH OUT – Independence is a mathematical condition, not a feeling

Independence is easy to assume and hard to verify. A rule of thumb: if two outcomes are generated by physically separate processes (different subjects, different trials, different sampling moments) you have a reasonable case for independence. If they are linked by any common cause—genetics, social ties, shared environment—they are almost certainly correlated.

5.3 Bayes' Theorem

RESULT – Bayes' theorem

For events A and B with $P(B) > 0$,

$$P(A \mid B) \;=\; \frac{P(B \mid A)\,P(A)}{P(B)}.$$

In a context with a partition $\{A_1, \dots, A_k\}$ of the sample space, the denominator expands by the law of total probability: $P(B) = \sum_i P(B \mid A_i)\,P(A_i)$.

EXAMPLE – A diagnostic test

A biomarker test for a rare disease has sensitivity $P(+ \mid D) = 0.95$ and specificity $P(- \mid D^c) = 0.98$. The disease prevalence is $P(D) = 0.001$. A randomly selected person tests positive. What is the probability that they actually have the disease?

$$P(D \mid +) = \frac{P(+ \mid D)\,P(D)}{P(+)} = \frac{0.95 \cdot 0.001}{0.95 \cdot 0.001 + 0.02 \cdot 0.999} \approx 0.045.$$

Despite the test's high sensitivity and specificity, a positive result at this prevalence means the person has under a 5% chance of actually being sick. Base-rate neglect—treating $P(D \mid +)$ as if it were close to the sensitivity $P(+ \mid D)$—is among the most consequential statistical errors in medicine.

5.4 Discrete Random Variables

DEFINITION – Discrete random variable, pmf, expectation

A *discrete random variable* X takes countably many values, each with a probability. The *probability mass function* is $p_X(x) = P(X = x)$. The *expected value* and *variance* are

$$\mathrm{E}(X) = \sum_x x\, p_X(x), \qquad \mathrm{Var}(X) = \sum_x \big(x - \mathrm{E}(X)\big)^2 p_X(x).$$

RESULT – Useful facts

For constants a, b and random variables X, Y:

$$\begin{aligned} \mathrm{E}(aX + b) &= a\,\mathrm{E}(X) + b, \\ \mathrm{Var}(aX + b) &= a^2\,\mathrm{Var}(X), \\ \mathrm{E}(X + Y) &= \mathrm{E}(X) + \mathrm{E}(Y) \text{ (always)}, \\ \mathrm{Var}(X + Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) \text{ when } X \perp Y. \end{aligned}$$

5.5 The Binomial Distribution

DEFINITION – Binomial distribution

Let X count the number of "successes" in n independent trials with common success probability p. Then $X \sim \text{Binomial}(n, p)$, with pmf

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Mean: np. Variance: $np(1-p)$.

The binomial distribution is the first probability model you will use in inference (Chapter 7, Analyzing Proportions). It is the foundation for the z-test for a proportion and, via the Poisson limit, for many count-based procedures.

5.6 Doing It in R

R SECTION – Four functions per distribution

R follows a uniform naming convention for every standard distribution: `d`, `p`, `q`, `r`.

```
set.seed(205)

# Binomial(10, 0.3): P(X = 3)
dbinom(3, size = 10, prob = 0.3)

# P(X <= 3)
pbinom(3, size = 10, prob = 0.3)

# The 95th percentile of Binomial(10, 0.3)
qbinom(0.95, size = 10, prob = 0.3)

# Draw 1000 realisations
rbinom(1000, size = 10, prob = 0.3)
```

Chapter Summary

Rule	Formula	When
Complement	$P(A^c) = 1 - P(A)$	Always
Addition (general)	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	Always
Addition (mut. excl.)	$P(A \cup B) = P(A) + P(B)$	$A \cap B = \emptyset$
Multiplication (general)	$P(A \cap B) = P(A \mid B)\,P(B)$	Always
Multiplication (indep.)	$P(A \cap B) = P(A)\,P(B)$	$A \perp B$
Conditional	$P(A \mid B) = P(A \cap B)/P(B)$	$P(B) > 0$
Bayes	$P(A \mid B) = P(B \mid A)P(A)/P(B)$	Updating beliefs
Binomial PMF	$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$	$X \sim \text{Bin}(n, p)$
Binomial mean/var	$\mu = np,\ \sigma^2 = np(1-p)$	

EXERCISES

1. A fair coin is flipped three times. List the sample space and compute the probability of at least two heads.

2. If $P(A) = 0.4$, $P(B) = 0.5$, and $P(A \cap B) = 0.2$, are A and B independent? Compute $P(A \mid B)$ and $P(A \cup B)$.

3. A disease has prevalence 2%. A test has sensitivity 0.90 and specificity 0.95. Compute $P(\text{disease} \mid +)$.

4. Let $X \sim \text{Binomial}(20, 0.25)$. Compute $P(X = 5)$, $P(X \leq 5)$, and $P(X \geq 8)$.

5. Using `rbinom()`, simulate 10,000 realisations of $\text{Binomial}(15, 0.4)$. Estimate the mean and variance from the simulation and compare to the theoretical values.

6. A genetics experiment crosses two heterozygous parents. Under Mendelian inheritance each offspring independently has probability $3/4$ of displaying the dominant phenotype. In a litter of $n = 8$ offspring, what is the probability that exactly 6 show the dominant phenotype? At least 5?

7. Three students each independently attempt a difficult problem. Each has probability 0.3 of solving it. (a) What is the probability that at least one solves it? (b) What is the probability that exactly two solve it?

8. A clinical screening programme tests every patient for a rare condition (prevalence 0.5%). The test has sensitivity 0.99 and specificity 0.97. (a) Compute the positive predictive value $P(D \mid +)$. (b) Why is it so much lower than the sensitivity? (c) What prevalence would be needed for the PPV to exceed 0.50?

9. A random variable X has PMF $P(X = 1) = 0.1$, $P(X = 2) = 0.3$, $P(X = 3) = 0.4$, $P(X = 4) = 0.2$. Compute $\text{E}(X)$, $\text{Var}(X)$, and $P(X \geq 3)$.

10. (Challenge) A jar contains 4 red and 6 blue marbles. You draw 3 without replacement.

Let X = number of red marbles drawn. (a) Write the PMF of X using the hypergeometric distribution. (b) Compute $\mathrm{E}(X)$ directly from the PMF. (c) Verify using the formula $\mathrm{E}(X) = nK/N$ where $N = 10$, $K = 4$, $n = 3$.

6

Hypothesis Testing

"The null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation."

— Sir Ronald A. Fisher

Science advances by proposing hypotheses and putting them to test against data. *Statistical hypothesis testing* is a formal procedure for asking whether the data you observed could plausibly have arisen from a specified null model. This chapter develops the framework—hypotheses, test statistics, p-values, significance levels, and Type I and II errors—in the abstract, then grounds it with the binomial test as a concrete worked example. Every test in later chapters (the t-test, the χ^2 test, the F-test) follows the same template.

LEARNING OBJECTIVES

- Formulate a null and alternative hypothesis from a scientific question, and choose the appropriate alternative (one- vs. two-sided).
- Define p-value, significance level α, Type I error, Type II error, and power; explain the trade-offs among them.
- Carry out a binomial test as a complete example, from hypothesis formulation through decision and scientific conclusion.
- Diagnose common misinterpretations of p-values (e.g. "the probability that H_0 is true").
- Relate hypothesis tests to confidence intervals via the duality between them.

6.1 The Framework

DEFINITION – Null and alternative hypotheses

The *null hypothesis* H_0 is a precise statement about a population parameter that the data will be evaluated against. The *alternative hypothesis* H_a (sometimes H_1) is a complementary statement that is the scientific claim of interest. The null is always of the form "no effect, no difference, no association"; the alternative is the hypothesis we would like to support.

Alternatives come in two flavours. A *two-sided* alternative, $H_a : \theta \neq \theta_0$, asks whether the parameter differs from θ_0 in *either* direction. A *one-sided* alternative, $H_a : \theta > \theta_0$ or $\theta < \theta_0$, asks whether the parameter is larger (or smaller) than θ_0. Choose one-sided only when the study could only have been run to detect a difference in one direction, and never post-hoc.

DEFINITION – p-value

The *p-value* is the probability, computed *assuming H_0 is true*, of observing a test statistic at least as extreme as the one actually observed, where "extreme" is defined by the direction of H_a.

WATCH OUT – What a p-value is *not*

A p-value is *not*:

- the probability that H_0 is true;
- the probability of Type I error in your particular study;
- a measure of the size of an effect.

It is a single probability about the data, under a single stated hypothesis. All three misreadings are widely repeated in the applied literature. None is correct.

DEFINITION – Significance level, Type I and Type II errors

The *significance level* α is the threshold below which the p-value leads us to reject H_0. Two errors are possible:

- **Type I error**: rejecting H_0 when H_0 is true. Probability: α.
- **Type II error**: failing to reject H_0 when H_a is true. Probability: β.

The *power* of a test is $1 - \beta$, the probability of correctly rejecting H_0 when H_a is true.

RESULT – The fundamental trade-off

For a fixed sample size, decreasing α (making rejection harder) increases β (making power worse), and vice-versa. Only by increasing the sample size can you reduce both errors simultaneously. This is why sample-size calculations begin with a target power (often 0.80) for a specified effect size at a chosen α.

6.2 The Test Procedure

The template applies to every test in this book.

1. State H_0 and H_a in terms of population parameters.
2. Choose α *before* looking at the data.
3. Check the assumptions of the test.
4. Compute the test statistic.
5. Compute the p-value under H_0.
6. Decide: if $p < \alpha$, reject H_0; otherwise fail to reject.
7. Translate the decision into a scientific sentence.

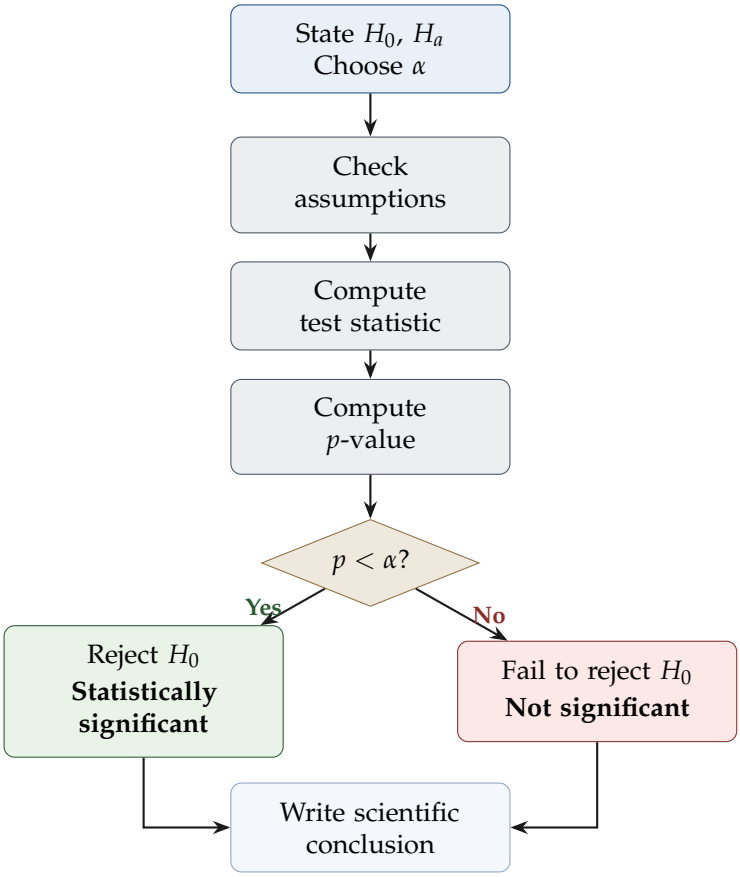


Figure 6.1: The seven-step hypothesis testing procedure used throughout this book. Every test in later chapters (the t-test, χ^2-test, F-test) follows this same template.

WATCH OUT – "Failing to reject" is not "accepting"

The language of hypothesis testing is asymmetric. When $p < \alpha$, we reject H_0. When $p \geq \alpha$, we *fail to reject* H_0—we have not gathered enough evidence to overturn it. We never "accept" H_0, because absence of evidence is not evidence of absence.

6.3 A Worked Example: The Binomial Test

EXAMPLE – Quality control in a pharmaceutical batch

A tablet-production line is said to operate at a defect rate of $p_0 = 0.05$. Inspectors test $n = 60$ tablets from a batch and find 7 defectives, $\hat{p} = 7/60 \approx 0.117$. Is there evidence that this batch exceeds the stated defect rate?

Step 1. $H_0 : p = 0.05$, $H_a : p > 0.05$ (one-sided, because a below-spec batch would not prompt rejection).

Step 2. $\alpha = 0.05$.

Step 3. Under H_0, $X \sim \text{Binomial}(60, 0.05)$.

Step 4 – 5. Observed test statistic $X = 7$; p-value $= P(X \geq 7) = 1 - P(X \leq 6) \approx 0.013$ (from `pbinom`).

Step 6. $p = 0.013 < 0.05$: reject H_0.

Step 7. There is statistically significant evidence that the batch defect rate exceeds 5% (one-sided exact binomial test, $n = 60$, $\hat{p} = 0.117$, $p = 0.013$). Quarantine the batch and investigate the production line.

6.4 Tests and Confidence Intervals

RESULT – Test-interval duality

A two-sided test of $H_0 : \theta = \theta_0$ at level α fails to reject H_0 if and only if θ_0 falls inside the corresponding $100(1-\alpha)\%$ confidence interval for θ.

This duality is why reporting a confidence interval is usually more informative than reporting a test. A CI tells you both the direction of the evidence and the precision with which you have estimated θ; a p-value tells you only whether a single value of θ is consistent with the data.

6.5 Doing It in R

R SECTION – One-line binomial test

```
binom.test(x = 7, n = 60, p = 0.05, alternative = "greater")
```

R output reports the exact p-value (no normal approximation), the observed proportion, and a $100(1-\alpha)\%$ confidence interval for the true proportion (Clopper–Pearson by default).

EXERCISES

Chapter Summary

Term	Definition
H_0 (null)	Precise "no effect" statement about a parameter.
H_a (alternative)	The scientific claim; one- or two-sided.
p-value	$P(\text{data at least this extreme} \mid H_0 \text{ true})$.
Significance level α	Pre-specified rejection threshold (often 0.05).
Type I error	Reject H_0 when H_0 is true; probability $= \alpha$.
Type II error	Fail to reject H_0 when H_a is true; probability $= \beta$.
Power	$1 - \beta$; probability of correctly rejecting H_0.
CI–test duality	μ_0 outside the 95% CI $\Leftrightarrow$ two-sided test rejects at $\alpha = 0.05$.

7-step template: Hypotheses $\rightarrow$ α $\rightarrow$ Assumptions $\rightarrow$ Statistic $\rightarrow$ p-value $\rightarrow$ Decision $\rightarrow$ Conclusion

Conceptual.

1. In a study with $\alpha = 0.05$, a colleague reports $p = 0.12$ and writes "we accept the null hypothesis." Explain in two sentences what is wrong with this phrasing.
2. Describe a situation in biological research where a Type II error is more costly than a Type I error, and vice-versa.
3. Explain the relationship between the significance level, the p-value, and the rejection region of a test.

Computational.

4. A coin is flipped 20 times and lands heads 15 times. Test $H_0 : p = 0.5$ against a two-sided alternative using `binom.test()`. Interpret your result.
5. A clinical claim asserts a cure rate of at least 70%. A trial of $n = 50$ patients produces 31 cures. Is there evidence that the true cure rate is below 70%?
6. Using `qbinom()`, find the smallest k such that $P(X \leq k) > 0.975$ under $X \sim \text{Binomial}(30, 0.2)$. Explain what this k represents.

R exercises.

7. Simulate 10,000 binomial experiments with $n = 60$ and $p_0 = 0.05$. For each experiment compute whether a one-sided binomial test at $\alpha = 0.05$ rejects H_0. What fraction of simulations reject? How does that number compare to α?
8. Repeat the previous exercise but simulate from the *alternative* $p = 0.10$. What is the empirical power of the test? How does it change if you increase n to 200?

9. A journal article reports "$p = 0.049$, statistically significant at the 5% level." A second paper reports "$p = 0.051$, not significant." (a) Are these two studies' evidence really as different as the labels suggest? (b) What additional information would help you decide which result is more practically meaningful?

10. Explain the trade-off between Type I and Type II errors in the context of a drug-safety trial. Which error type is more costly, and how should that influence the choice of α?

11. A wildlife manager tests whether the proportion of injured birds in a migratory population exceeds 8%. She observes 14 injured birds in a sample of $n = 120$. (a) State H_0 and H_a. (b) Use `binom.test()` to find the exact p-value. (c) Construct a 95% CI for the true proportion. (d) Write a two-sentence scientific conclusion.

12. (Challenge) Show algebraically that for a two-sided binomial test with $H_0 : p = p_0$, the p-value equals twice the one-sided p-value when the distribution is symmetric. Under what conditions is the binomial distribution approximately symmetric?

Part II

Categorical Data Analysis

7

Analyzing Proportions

"It is the mark of a truly intelligent person to be moved by statistics."

— George Bernard Shaw

A large fraction of biological data is not continuous. Did the seed germinate (yes/no)? Did the animal survive (yes/no)? Did the patient respond to therapy (yes/no)? Is the observed gene in the active or inactive state? Proportions are everywhere in the life sciences, and they demand their own inferential tools. This chapter develops those tools: the binomial test for a single proportion, the large-sample z-test, the two-sample z-test for comparing two proportions, and the corresponding confidence intervals.

LEARNING OBJECTIVES

- Carry out an exact binomial test for a single proportion and explain when it is preferable to the large-sample z-test.
- Use the one-sample z-test for a proportion and construct the corresponding confidence interval.
- Compare two proportions from independent samples using the two-sample z-test.
- Diagnose when the large-sample approximation fails and fall back on the exact binomial test or Fisher's exact test.

7.1 A Single Proportion: Binomial and z-Test

DEFINITION – One-sample proportion setup

Let X be the number of "successes" in n independent trials with common success probability p. Then $X \sim \text{Binomial}(n, p)$, and the sample proportion is $\hat{p} = X/n$. The sampling distribution of $\hat{p}$ has

$$\mathrm{E}(\hat{p}) = p, \qquad \mathrm{SD}(\hat{p}) = \sqrt{p(1-p)/n}.$$

RESULT – Exact binomial test

To test $H_0 : p = p_0$ against a chosen alternative, compute the exact p-value from the binomial distribution with n trials and success probability p_0. R: `binom.test(x, n, p = p0)`.

RESULT – Large-sample z-test for a proportion

When $np_0 \geq 10$ and $n(1 - p_0) \geq 10$, the statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

is approximately standard normal under H_0. The corresponding Wald-style $100(1 - \alpha)\%$ confidence interval for p is

$$\hat{p} \pm z^*_{\alpha/2}\sqrt{\hat{p}(1 - \hat{p})/n}.$$

Better small-sample intervals (Wilson, Agresti–Coull, Clopper–Pearson) are available in `binom.test()` and the binom package.

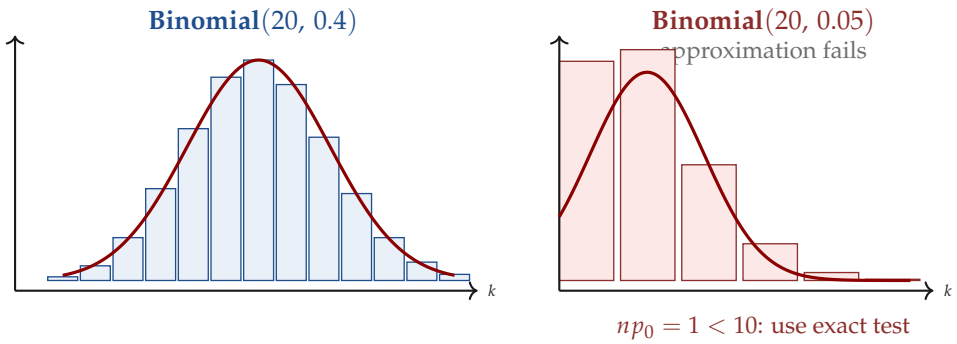


Figure 7.1: The normal approximation to the binomial works well when $np_0 \geq 10$ and $n(1 - p_0) \geq 10$ (left: $p_0 = 0.4$, $n = 20$). When these conditions fail (right: $p_0 = 0.05$, $n = 20$), the binomial is too skewed and the normal overlay is a poor fit; use the exact binomial test or Fisher's exact test instead.

EXAMPLE – Seed germination

A horticulturalist plants $n = 120$ seeds from a new cultivar and observes 87 germinations. The company claims a germination rate of at least 80%. Is there evidence that the true rate is below 80%?
Sample proportion: $\hat{p} = 87/120 = 0.725$.

$$z = \frac{0.725 - 0.80}{\sqrt{0.80 \cdot 0.20/120}} = \frac{-0.075}{0.0365} \approx -2.05.$$

One-sided p-value: $P(Z \leq -2.05) \approx 0.020$. At $\alpha = 0.05$ we reject the claim; there is evidence the true germination rate is below 80%.

7.2 Two Proportions: The z-Test for Two Samples

RESULT – Two-sample z-test for proportions

For two independent samples with counts x_1, x_2 and sizes n_1, n_2, let $\hat{p}_1 = x_1/n_1$ and $\hat{p}_2 = x_2/n_2$. To test $H_0 : p_1 = p_2$, use the pooled proportion $\hat{p}_{\text{pool}} = (x_1 + x_2)/(n_1 + n_2)$ to compute

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_{\text{pool}}(1 - \hat{p}_{\text{pool}})(1/n_1 + 1/n_2)}}.$$

Under H_0, z is approximately standard normal. The $100(1-\alpha)\%$ CI for $p_1 - p_2$ uses the *unpooled* standard error $\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$.

WATCH OUT – When the approximation fails

The z-test needs $n\hat{p}$ and $n(1-\hat{p})$ to be at least ≈ 10 in each group. For rare events (small p), an "adequate" n may be thousands. When the approximation fails, use `fisher.test()` (exact) or `prop.test()` with continuity correction. Always check the counts, not just the proportions.

7.3 Doing It in R

R SECTION – One proportion, three ways

```
library(tidyverse)
set.seed(205)

# Exact binomial test (preferred when counts are small)
binom.test(x = 87, n = 120, p = 0.80, alternative = "less")

# Large-sample z-test via prop.test()
prop.test(x = 87, n = 120, p = 0.80, alternative = "less", correct =
    FALSE)
```

R SECTION – Two proportions

```
# 42 of 150 patients on drug A responded; 29 of 150 on drug B responded
prop.test(x = c(42, 29), n = c(150, 150))

# For small counts, use Fisher's exact
fisher.test(matrix(c(42, 108, 29, 121), nrow = 2))
```

Chapter Summary

Procedure	Test statistic	Use when
Exact binomial	`binom.test()`	Always valid
One-sample z	$z = (\hat{p} - p_0)/\sqrt{p_0(1-p_0)/n}$	$np_0 \geq 10$, $n(1-p_0) \geq 10$
Wald CI	$\hat{p} \pm z^* \sqrt{\hat{p}(1-\hat{p})/n}$	Large n
Two-sample z	$(\hat{p}_1 - \hat{p}_2)/\mathrm{SE}_{\text{pool}}$	Both n large
Fisher's exact	`fisher.test()`	Small expected counts

Two-sample CI uses unpooled SE; test uses pooled SE.

EXERCISES

1. A clinical trial reports 14 of 45 patients responding on the experimental arm. Test $H_0 : p = 0.40$ against a two-sided alternative. Report both the exact binomial and the large-sample z p-values; compare.

2. In an ecology study, 23 of 80 bird nests in a primary forest fledge young, vs. 19 of 110 in a logged forest. Test for a difference in fledging proportions. Report the 95% CI for the difference.

3. Explain why the standard error in the *test* uses the pooled proportion but the standard error in the *confidence interval* uses the unpooled version.

4. A public-health report claims a 3% smoking-cessation rate for a new programme. A replication enrolls $n = 200$ and observes 3 quitters. Is the z-test trustworthy here? Run the appropriate exact procedure instead.

5. A vaccine trial randomises $n_1 = 500$ participants to vaccine and $n_2 = 500$ to placebo. Infected: 12 vaccine, 38 placebo. (a) Compute $\hat{p}_1$, $\hat{p}_2$, and the pooled proportion. (b) Carry out the two-sample z-test. (c) Compute a 95% CI for $p_1 - p_2$. (d) Interpret both the test result and the CI in plain English.

6. Explain, using the conditions $np_0 \geq 10$ and $n(1-p_0) \geq 10$, why the z-test for a proportion fails when p_0 is very small even with a large n. Construct a numerical example that illustrates the failure.

7. Two coral-reef surveys record bleaching in 22/80 quadrats (Site A) and 14/60 quadrats (Site B). (a) Test for a difference using `prop.test()`. (b) Compute the 95% CI for the difference. (c) Compute the relative risk $\mathrm{RR} = \hat{p}_A/\hat{p}_B$ and interpret it ecologically.

8. A researcher tests whether a new therapeutic technique increases the proportion of patients in remission above the historical baseline of 35%. In a pilot of $n = 25$ patients, 12 achieve remission. (a) Check conditions for the z-test. (b) Run the appropriate test. (c) Compute a 90% CI using `binom.test()`.

9. (Challenge) Derive the Wilson confidence interval for a proportion and show algebraically

that it always falls within $[0, 1]$, unlike the Wald interval. For what value of $\hat{p}$ and n do the two intervals differ most dramatically?

8

Fitting Probability Models to Data

"Essentially, all models are wrong, but some are useful."

— George E. P. Box

A researcher has observed counts in several categories: phenotype A, phenotype B, phenotype C. A theoretical model—Mendelian inheritance, say, with a predicted ratio 9:3:3:1—makes specific predictions about the expected counts. How well do the observed counts agree with those predictions? The chi-squared goodness-of-fit (GoF) test gives a principled answer.

LEARNING OBJECTIVES

- State the null hypothesis of a chi-squared GoF test.
- Compute expected counts under a given model and the chi-squared test statistic by hand.
- Identify the correct degrees of freedom when the model parameters are fixed vs. estimated from the data.
- State and diagnose the "all expected ≥ 5" guideline and use `chisq.test()` appropriately.

8.1 The Chi-Squared Test Statistic

DEFINITION – Chi-squared GoF statistic

Let $O_1, \ldots, O_k$ be the observed counts in k mutually exclusive categories, with $\sum O_i = n$. Let $E_1, \ldots, E_k$ be the expected counts under a stated null model. The chi-squared goodness-of-fit statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

Under H_0, for large enough n, χ^2 has approximately a chi-squared distribution with $k - 1 - m$ degrees of freedom, where m is the number of parameters estimated from the data.

Each term $(O_i - E_i)^2 / E_i$ is large when the observed count deviates markedly from its expected value relative to the size of the expected value. Summing over categories aggregates evidence.

RESULT – When the chi-squared approximation is valid

The chi-squared distribution approximates the sampling distribution of the GoF statistic when all expected counts are at least 5 (a conservative rule). If some expected counts are smaller, consider combining adjacent categories or using an exact test.

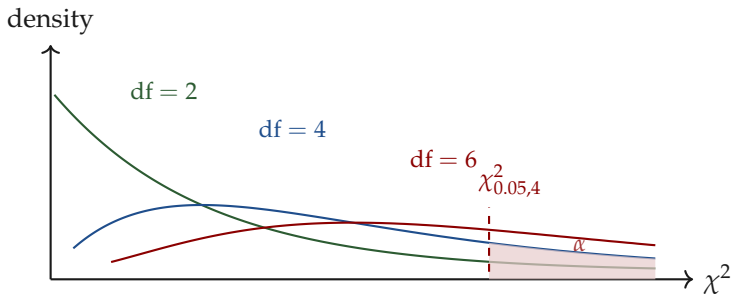


Figure 8.1: Chi-squared distributions for three values of df. As df increases, the distribution shifts right and becomes more symmetric. The shaded region shows the rejection region for df $= 4$ at $\alpha = 0.05$: we reject H_0 when the test statistic exceeds the critical value $\chi^2_{0.05,\,4} \approx 9.49$.

8.2 A Worked Example: Mendelian Inheritance

EXAMPLE – Pea-plant phenotype ratios

Mendel's classic dihybrid cross predicts phenotype ratios $9:3:3:1$ for yellow-round, yellow-wrinkled, green-round, green-wrinkled peas. A student's cross yields $n = 400$ offspring with observed counts $232, 73, 80, 15$.

Expected counts under $9:3:3:1$ with $n = 400$:

$$E_{YR} = 225, \quad E_{YW} = 75, \quad E_{GR} = 75, \quad E_{GW} = 25.$$

Test statistic.

$$\chi^2 = \frac{(232-225)^2}{225} + \frac{(73-75)^2}{75} + \frac{(80-75)^2}{75} + \frac{(15-25)^2}{25} \approx 0.22 + 0.05 + 0.33 + 4.00 = 4.60.$$

Degrees of freedom. No parameters are estimated; $\text{df} = k - 1 = 3$.

P-value. $P(\chi^2_3 \geq 4.60) \approx 0.20$. We fail to reject H_0; the data are consistent with $9:3:3:1$ at $\alpha = 0.05$.

WATCH OUT – Small χ^2 does not "prove" the model

Failing to reject the null only means the observed data are not extreme enough to contradict the model *at the current sample size*. A larger study may reveal discrepancies that a smaller one cannot detect.

8.3 Doing It in R

R SECTION – One line if the expected proportions are supplied

```
observed <- c(232, 73, 80, 15)
expected_p <- c(9, 3, 3, 1) / 16
chisq.test(observed, p = expected_p)
```

Output includes χ^2, df, and the p-value. If the approximation is dubious, pass `simulate.p.value = TRUE` to get a Monte Carlo p-value.

Chapter Summary

Quantity	Formula	Notes
GoF statistic	$\chi^2 = \sum_i (O_i - E_i)^2 / E_i$	k categories
Degrees of freedom	$k - 1 - m$	m = estimated params
Expected count	$E_i = n \cdot p_{0i}$	Under H_0
Validity condition	All $E_i \geq 5$	Otherwise: combine / Monte Carlo
R function	`chisq.test(obs, p = probs)`	

EXERCISES

1. A genetic cross predicts ratios 3 : 1 for two phenotypes. Out of $n = 160$ offspring, 108 are phenotype A and 52 are phenotype B. Test the model.
2. A fair six-sided die is rolled 60 times, giving counts $9, 12, 10, 8, 11, 10$. Carry out a GoF test of fairness and interpret the result.
3. Under what circumstances do you reduce the degrees of freedom below $k - 1$? Give a concrete example.
4. Suppose some expected counts are below 5. Describe two remedies and their trade-offs.
5. A wildlife biologist observes bird species in a forest plot: 50 warblers, 28 sparrows, 12 thrushes, 10 other. A regional survey predicts proportions 45%, 30%, 15%, 10%. Test whether the plot's composition matches the regional model. Compute the contribution of each category to χ^2.
6. A six-sided die is suspected of being loaded. It is rolled $n = 300$ times with results: face 1: 62, 2: 44, 3: 51, 4: 49, 5: 55, 6: 39. (a) State H_0. (b) Compute χ^2 and the p-value. (c) Which face contributes most to the test statistic? (d) What would you conclude?
7. A population genetics study measures genotype frequencies at a biallelic locus. Observed: AA = 180, Aa = 240, aa = 80 ($n = 500$). Hardy–Weinberg equilibrium predicts proportions $p^2 : 2p(1-p) : (1-p)^2$ where $\hat{p} = (2 \times 180 + 240)/(2 \times 500) = 0.6$. (a) Compute expected counts. (b) Note that $\hat{p}$ was estimated from the data: what is the correct df? (c) Test and interpret.
8. Explain why a large χ^2 statistic with many categories does not by itself tell you *which* categories drove the departure from the null model. How would you investigate this further?
9. (Challenge) Write an R simulation that generates 10,000 samples from a Binomial$(60, 0.5)$ distribution, groups the outcomes into 6 equal bins, and applies `chisq.test()` to each. What fraction of tests reject at $\alpha = 0.05$? Compare to the theoretical Type I error rate.

9

Contingency Analysis

"The simple idea of cross-classifying two categorical variables has been one of the most fertile in the history of statistics."

— paraphrase of Leo Goodman

Chapter 8 tested whether counts in a single categorical variable matched a theoretical model. This chapter extends the chi-squared machinery to *two* categorical variables. Given a sample that has been cross-classified by two categorical attributes—treatment vs. outcome, species vs. habitat, exposure vs. disease status— we want to know whether the two attributes are associated or independent.

LEARNING OBJECTIVES

- Build a two-way contingency table from raw data.
- Compute expected cell counts under the null of independence.
- Carry out and interpret a chi-squared test of independence.
- Choose between the chi-squared test and Fisher's exact test based on sample size.
- Quantify the strength of association with the odds ratio or relative risk.

9.1 Two-Way Tables and Independence

Cross-classifying n observations by two categorical variables with r and c levels produces an $r \times c$ contingency table. Let O_{ij} denote the observed count in row i, column j; let R_i, C_j denote the row and column marginals; and let n denote the grand total.

DEFINITION – Expected counts under independence

Under the null hypothesis of independence between the two categorical variables,

$$E_{ij} = \frac{R_i \cdot C_j}{n}.$$

This is the count you would expect in cell (i, j) if the row and column variables were unrelated and only the marginal totals were fixed.

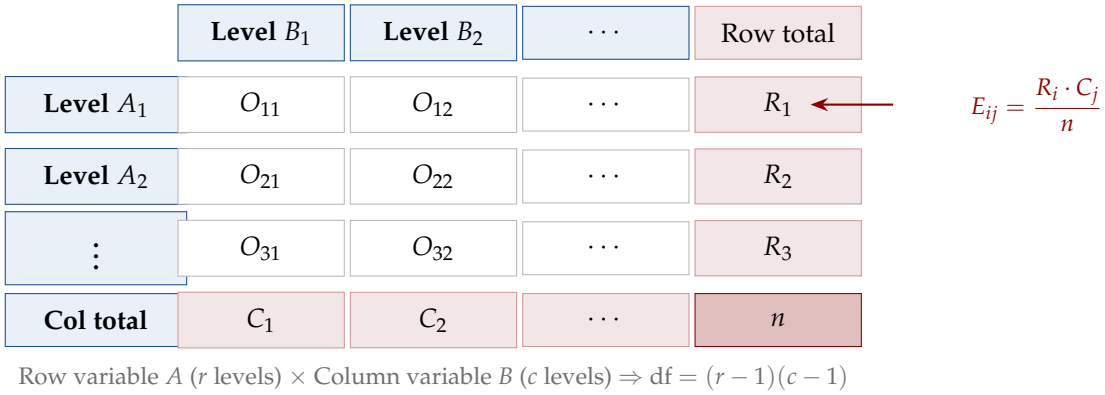


Figure 9.1: Structure of a two-way contingency table. Each cell contains an observed count O_{ij}; marginals R_i and C_j give the row and column totals. Under the null hypothesis of independence, the expected count in cell (i, j) is $E_{ij} = R_i C_j / n$.

RESULT – Chi-squared test of independence

The test statistic

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

has, under H_0, an approximate chi-squared distribution with $(r-1)(c-1)$ degrees of freedom. Large χ^2 is evidence of association.

9.2 A Worked Example

EXAMPLE – Antibiotic resistance by hospital ward

A hospital epidemiologist cross-classifies $n = 240$ bacterial isolates by ward (ICU, surgical, medical) and resistance status (resistant, susceptible). The observed table is

	ICU	Surgical	Medical	Row total
Resistant	42	28	30	100
Susceptible	38	52	50	140
Column total	80	80	80	240

Expected counts under independence for resistant: $E = 100 \cdot 80/240 \approx 33.3$ in each ward; for susceptible, $E \approx 46.7$.
Test statistic.

$$\chi^2 = \frac{(42-33.3)^2}{33.3} + \frac{(28-33.3)^2}{33.3} + \frac{(30-33.3)^2}{33.3} + \frac{(38-46.7)^2}{46.7} + \frac{(52-46.7)^2}{46.7} + \frac{(50-46.7)^2}{46.7} \approx 5.03.$$

df $= (2-1)(3-1) = 2$. $P(\chi^2_2 \geq 5.03) \approx 0.081$.
Conclusion. At $\alpha = 0.05$ we fail to reject, but the data are suggestive of a higher resistance rate in the ICU. A larger study would be warranted.

9.3 Fisher's Exact Test

When some expected counts are small (the rule of thumb: any expected count below 5), the chi-squared approximation becomes unreliable. *Fisher's exact test* computes the exact p-value from the hypergeometric distribution, conditional on the observed margins. It is computationally feasible for any 2×2 table and for small $r \times c$ tables via `fisher.test()` in R.

9.4 Odds Ratios and Relative Risk

DEFINITION – Odds ratio (2×2 table)

For the 2×2 table

	Event	No event
Exposed	a	b
Unexposed	c	d

the odds of the event among the exposed is a/b; among the unexposed, c/d. The *odds ratio* is $\mathrm{OR} = ad/(bc)$.

DEFINITION – Relative risk

The *relative risk* (risk ratio) is

$$\mathrm{RR} = \frac{a/(a+b)}{c/(c+d)} = \frac{\text{risk of event} \mid \text{exposed}}{\text{risk of event} \mid \text{unexposed}}.$$

For rare events, OR and RR are numerically close. For common events they diverge, and RR is usually the more scientifically natural measure of effect size.

9.5 Doing It in R

R SECTION – Chi-squared and Fisher in one page

```r
library(tidyverse)
set.seed(205)

# Build a contingency matrix
resistance <- matrix(c(42, 38, 28, 52, 30, 50),
                     nrow = 2, byrow = FALSE,
                     dimnames = list(status = c("R", "S"),
                                     ward   = c("ICU", "Surg", "Med")))
chisq.test(resistance)                       # chi-squared
fisher.test(resistance)                      # exact (works for small
    tables)

# For a 2x2 with rare cells
small <- matrix(c(2, 8, 1, 14), nrow = 2)
chisq.test(small)                            # approximation warning
fisher.test(small)                           # exact -- use this

# Odds ratio with CI (vcd or epiR)
# vcd::oddsratio(small, log = FALSE)
```

Chapter Summary

Quantity	Formula	Notes
Expected cell count	$E_{ij} = R_i C_j / n$	Under independence
χ^2 independence	$\sum_{i,j}(O_{ij} - E_{ij})^2 / E_{ij}$	df $= (r-1)(c-1)$
Odds ratio (2×2)	OR $= ad/(bc)$	Event vs. no-event
Relative risk	RR $= [a/(a+b)]/[c/(c+d)]$	Exposed vs. unexposed
Fisher's exact	`fisher.test()`	When $E_{ij} < 5$

OR $\approx$ RR only for rare events. For common events, prefer RR.

EXERCISES

1. Construct a 2×2 contingency table for a study where 14/60 exposed and 6/80 unexposed subjects developed the outcome. Compute the odds ratio, relative risk, and a chi-squared p-value.

2. A habitat-preference study cross-classifies $n = 120$ beetles by sex (M/F) and habitat choice (sun/shade). Observed: male-sun 22, male-shade 38, female-sun 33, female-shade 27. Test for association and report the expected cell counts.

3. Explain why the chi-squared approximation can fail even when n is large, if one row or column has very small marginal totals.

4. Explain, in one sentence each, the difference between the odds ratio and the relative risk, and when each is preferable.

5. A nutritional epidemiologist cross-classifies $n = 320$ adults by diet type (plant-based / omnivore) and presence of metabolic syndrome (yes / no). Observed: plant–yes 18, plant–no 142, omni–yes 52, omni–no 108. (a) Compute the χ^2 statistic and p-value. (b) Compute OR and RR. (c) Interpret the OR for a general-audience summary.

6. Two pesticide treatments (A and B) are compared on crop damage (damaged / undamaged). Plot A: 8 damaged, 42 undamaged. Plot B: 19 damaged, 31 undamaged. Some expected counts are borderline. (a) Compute all expected counts. (b) Run both `chisq.test()` and `fisher.test()` in R. (c) Compare the p-values and recommend which to report.

7. Define "statistical independence" in a 2×2 contingency table in two ways: (i) using conditional probabilities, and (ii) using expected cell counts. Show algebraically that the two definitions are equivalent.

8. A 3×3 table cross-classifies education level (low/medium/high) with exercise frequency (rarely/sometimes/regularly). The chi-squared test gives $\chi^2 = 11.4$. (a) Find the p-value. (b) The test is significant but a colleague wants to know *which* cells drive the result. Compute and interpret the standardized residuals $(O_{ij} - E_{ij})/\sqrt{E_{ij}}$ for each cell.

9. (Challenge) Prove that the odds ratio is unchanged when the rows and columns of a 2×2

table are swapped. Then show that the log odds ratio, $\ln(\text{OR}) = \ln(ad) - \ln(bc)$, has an approximate normal distribution under large samples, and derive its standard error.

10

The Normal Distribution

"Everybody believes in the law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact."

— Gabriel Lippmann, attributed

Of all probability distributions, the normal is the most consequential for applied science. It is the limiting shape of the sample mean under almost any distribution (the Central Limit Theorem), the reference curve against which all standard inference tools are built, and the natural model for any measurement that accumulates small, independent sources of variation—body mass, blood pressure, leaf area, reaction time. This chapter develops the normal from the inside out: its parameters, its symmetries, the mechanics of standardization, and— most importantly—the Central Limit Theorem, which explains why the normal refuses to go away.

LEARNING OBJECTIVES

- State the two parameters of the normal distribution and describe their geometric roles.
- Standardize a normal random variable by computing its z-score and use it to find probabilities from a standard normal table or software.
- State the Central Limit Theorem and explain the conditions under which it applies.
- Identify the sampling distribution of $\bar{X}$ and compute its mean and standard error.
- Use the normal distribution in R via `pnorm()`, `qnorm()`, and `rnorm()`.

10.1 The Normal Density

DEFINITION – Normal distribution

A continuous random variable X has a *normal distribution* with mean μ and standard deviation $\sigma > 0$, written $X \sim N(\mu, \sigma)$, if its probability density function is

$$f(x) \;=\; \frac{1}{\sigma\sqrt{2\pi}} \exp\!\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty.$$

The density is symmetric about μ, bell-shaped, and has inflection points at $\mu \pm \sigma$.

Two features of the formula carry almost all the intuition. First, the term $(x-\mu)^2/\sigma^2$ measures, in squared standardized units, how far x is from the centre. Second, the negative exponential compresses this distance into a density that falls off very quickly: the probability of an observation more than three standard deviations from the mean is already below 0.003.

RESULT – The empirical rule

For any $X \sim N(\mu, \sigma)$,

$$\Pr(\mu-\sigma \le X \le \mu+\sigma) \approx 0.683, \quad \Pr(\mu-2\sigma \le X \le \mu+2\sigma) \approx 0.954, \quad \Pr(\mu-3\sigma \le X \le \mu+3\sigma) \approx 0.997.$$

These percentages are not approximations of some deeper truth; they are the area under the normal curve computed exactly. Committing them to memory converts normal probabilities into a one-line estimate.

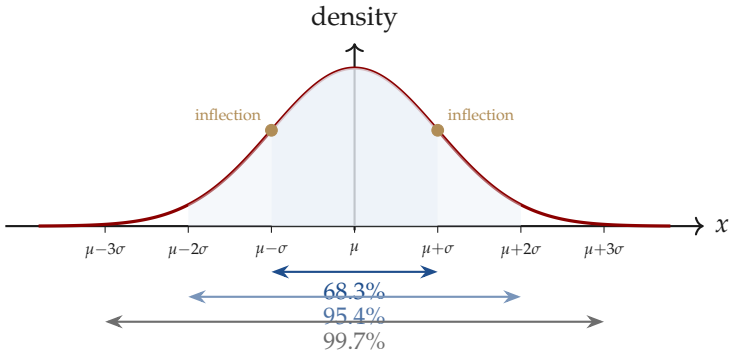


Figure 10.1: The normal distribution and the empirical rule. Shaded regions cover 68.3% ($\pm 1\sigma$) and 95.4% ($\pm 2\sigma$) of the area. The curve has inflection points exactly at $\mu \pm \sigma$, where the curvature changes from concave to convex.

10.2 Standardization and the Z-Score

DEFINITION – Standard normal distribution and z-score

The *standard normal* is $Z \sim N(0,1)$. For any $X \sim N(\mu, \sigma)$, the random variable

$$Z = \frac{X - \mu}{\sigma}$$

has a standard normal distribution. The realised value $z = (x - \mu)/\sigma$ is the *z-score* of x and reports how many standard deviations x sits above or below the mean.

Every normal probability reduces to a standard normal probability. To find $\Pr(X \le x)$ for a general normal, convert x into a z-score and look up the resulting probability in a Z-table or `pnorm()`.

EXAMPLE – Seedling heights

The heights of a species of seedling are normally distributed with $\mu = 6.2$ cm and $\sigma = 0.8$ cm. What fraction of seedlings are shorter than 5.0 cm?

$$z = (5.0 - 6.2)/0.8 = -1.50, \quad \Pr(Z \le -1.50) \approx 0.0668.$$

About 6.7% of seedlings are shorter than 5.0 cm.

10.3 Sampling Distributions and the CLT

All of Chapters 11–15 rest on one fact: the sample mean $\bar{X}$ is less variable than individual observations, and its sampling distribution is approximately normal for moderate-to-large n even when the individual observations are not.

RESULT – Sampling distribution of $\bar{X}$

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a population with mean μ and standard deviation σ. Then

$$\mathrm{E}(\bar{X}) = \mu, \qquad \mathrm{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

If the population is itself normal, $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$ *exactly*. If the population is not normal, the distribution of $\bar{X}$ is *approximately* normal for large enough n.

RESULT – Central Limit Theorem

Let $X_1, X_2, \ldots$ be i.i.d. with mean μ and finite variance σ^2. As $n \to \infty$,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \longrightarrow N(0,1)$$

in distribution. In practice, the normal approximation is good for $n \gtrsim 30$ when the population is moderately skewed, and for much smaller n when the population is symmetric.

WATCH OUT – The CLT is about the mean, not about individual observations

The CLT does *not* say that individual measurements become normal at large n. Individual observations retain whatever shape the population has. What becomes normal is the distribution of the *sample mean* across hypothetical repetitions of the study. Students often collapse these two statements; they are different.

10.4 The Normal in R

R SECTION – Four functions cover it

R exposes the normal distribution through four functions following a uniform naming convention (`d`, `p`, `q`, `r`):

```
set.seed(205)

dnorm(0, mean = 0, sd = 1)        # density at x = 0
pnorm(1.96)                       # P(Z <= 1.96) = cumulative prob
qnorm(0.975)                      # inverse: z such that P(Z<=z) = 0.975
rnorm(10, mean = 100, sd = 15)    # draw 10 values from N(100, 15)
```

R SECTION – Simulating the CLT

```
library(tidyverse)

# Population: heavily right-skewed (exponential with rate 1)
# Draw 10,000 samples of size n = 40; compute the mean of each
means <- replicate(10000, mean(rexp(40, rate = 1)))

tibble(xbar = means) |>
  ggplot(aes(x = xbar)) +
  geom_histogram(bins = 40, fill = "#EAF3E8", colour = "#2C5B32") +
  theme_minimal() +
  labs(x = "Sample mean", title = "Sample means are nearly normal")
```

10.5 Chapter Summary

R function	Purpose
`dnorm(x, mean, sd)`	density at x
`pnorm(q, mean, sd)`	cumulative probability
`qnorm(p, mean, sd)`	quantile (inverse CDF)
`rnorm(n, mean, sd)`	random sample

Chapter Summary

Concept	Formula / fact	Notes
Normal PDF	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$	
z-score	$z = (x - \mu)/\sigma$	Standardises to $N(0,1)$
Empirical rule	68%, 95%, 99.7% within 1σ, 2σ, 3σ	
CLT	$\bar{X}_n \to N(\mu, \sigma/\sqrt{n})$	As $n \to \infty$
SE of $\bar{X}$	$\sigma/\sqrt{n}$	Population σ known
R functions	`pnorm`, `qnorm`, `dnorm`, `rnorm`	

EXERCISES

1. Resting heart rate in a population is normally distributed with $\mu = 72$ bpm and $\sigma = 9$. What fraction of people have a resting rate below 60? Above 90? Between 60 and 90?
2. Convert $x = 165$ from $N(150, 10)$ into its z-score and interpret it in plain English.
3. Verify the empirical rule using `pnorm()`.
4. Draw 100,000 samples of size $n = 30$ from a uniform population on $[0, 1]$ and plot the distribution of their sample means. How closely does it resemble a normal distribution? Match the mean and standard deviation against the theoretical values.
5. Suppose $X \sim N(0, 1)$ and $Y \sim N(0, 1)$ are independent. Is $X + Y$ normal? What are its mean and standard deviation? Verify via simulation.
6. Leaf width (cm) in a plant species is $N(4.2, 0.6)$. (a) What fraction of leaves are wider than 5 cm? (b) What fraction fall between 3.5 cm and 5 cm? (c) Find the 90th percentile width.
7. Blood pressure in a population is $N(120, 15)$ mmHg. Hypertension is defined as SBP $>$ 140 mmHg. (a) What proportion of the population is hypertensive? (b) If a clinic screens 200 patients, about how many would be expected to be hypertensive?
8. Using `rnorm()` in R, simulate 10,000 sample means of size $n = 5$, $n = 30$, and $n = 100$ from an exponential distribution with rate 1. For each n, plot a histogram of the sample means and overlay the theoretical normal density predicted by the CLT. At what n does the CLT approximation look adequate?
9. A normal distribution has mean μ and SD σ. Show that the inflection points of its density function occur exactly at $x = \mu \pm \sigma$. (Hint: set the second derivative of $f(x)$ to zero.)
10. (Challenge) The sum of n independent $N(\mu, \sigma)$ random variables has distribution $N(n\mu, \sigma\sqrt{n})$. Use this fact to derive the sampling distribution of $\bar{X}$, then explain in one sentence why the CLT is remarkable: it extends this result to *non-normal* populations.

Part III

Inference for Means

11

Inference for a Normal Population

"The statistician cannot evade the responsibility for understanding the process he applies or recommends."

— Sir Ronald A. Fisher

A marine biologist weighs 18 green sea turtle (*Chelonia mydas*) hatchlings from a single nest at Juno Beach, Florida. She wants to know whether the mean hatchling mass at her field site differs from the long-term Atlantic-population mean of 24.8 grams reported in the literature. She has a sample mean and a sample standard deviation. She does *not* know the population standard deviation, and she never will.

Chapter 10 assumed σ was known. That assumption was a pedagogical scaffold, not a scientific reality. Almost every inferential problem you will meet in the life sciences begins where Chapter 10 ended: with s instead of σ, and with a sample size that is small enough for the difference to matter. This chapter develops the tools that take over in that setting—the one-sample t-statistic, the t-distribution, the one-sample t-test, and the one-sample t confidence interval. The logic is exactly the logic of Chapter 10. Only one thing changes, and that single change ripples through every formula, every critical value, and every assumption you will check.

LEARNING OBJECTIVES

- Explain why the sampling distribution of $\bar{X}$ changes from Z to t_{n-1} when σ is estimated from the sample.
- Carry out a one-sample t-test from raw data in R and interpret the output in the context of a scientific question.
- Construct and interpret a one-sample t confidence interval and relate it to the corresponding two-sided test.
- State the four assumptions of the one-sample t procedure and diagnose each using R.
- Quantify effect size with Cohen's d and explain why statistical significance is not scientific importance.
- Use a decision guide to select the appropriate inferential procedure for a new dataset.

11.1 From z to t

In Chapter 10 we assumed the population standard deviation σ was known. The sampling distribution of $\bar{X}$ was normal with standard error $\sigma/\sqrt{n}$, and we could standardize the sample mean into a z-statistic whose distribution was the familiar standard normal. That assumption is almost never defensible in practice. You are rarely handed σ on a silver platter; what you have is a sample standard deviation s, which is itself a random variable that changes from sample to sample.

Substituting s for σ feels harmless, but it injects a second source of uncertainty into the standard error. When we divide $\bar{X} - \mu$ by $s/\sqrt{n}$ instead of $\sigma/\sqrt{n}$, we are dividing by a *noisy* denominator. Sometimes s undershoots σ and the standardized statistic is inflated; sometimes s overshoots and the statistic is damped. The standard normal curve is no longer the right reference distribution. We need a curve with heavier tails—heavier precisely because the denominator is random.

DEFINITION – One-sample t-statistic

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with mean μ and unknown variance. Let $\bar{X}$ and s denote the sample mean and the sample standard deviation. The *one-sample t-statistic* for testing $H_0 : \mu = \mu_0$ is

$$t \;=\; \frac{\bar{X} - \mu_0}{s/\sqrt{n}}.$$

When H_0 is true, t follows a t-distribution with $n - 1$ degrees of freedom, written $t \sim t_{n-1}$.

The t-distribution was discovered by William Sealy Gosset, who worked as a brewer at the Guinness factory in Dublin and published under the pseudonym "Student" because Guinness forbade employee publication. Gosset was trying to infer the mean quality of barley from small samples; he realized that the reference distribution for the standardized sample mean depended on the sample size when σ was estimated, and he worked out the exact form. The resulting family of curves is indexed by a single parameter, the *degrees of freedom* $\nu = n - 1$, which counts the pieces of independent information left after using one piece to estimate $\bar{X}$.

RESULT – Properties of the t-distribution

For each positive integer ν, the t_ν distribution is

1. symmetric about zero and bell-shaped;
2. *heavier-tailed* than the standard normal, reflecting the extra uncertainty introduced by estimating σ with s;
3. convergent to the standard normal as $\nu \to \infty$, in the sense that the density of t_ν converges pointwise to the density of Z.

Small ν produces visibly heavier tails. By $\nu = 30$ the difference between t_ν and Z in the middle 95% of either curve is under 2% at every point, which is why older textbooks drew a sharp line at $n = 30$. With modern software, that line is a historical artifact; we simply always use t.

11.1.1 Why heavier tails are exactly right

Figure 11.1 shows the standard normal density alongside t_3 and t_{12}. The t_3 curve is noticeably shorter in the middle and thicker in the tails. The t_{12} curve is already almost indistinguishable from Z near the center but still carries a visible tail premium. The tail inflation is not an ad-hoc correction; it is exactly what the extra variability in s contributes.

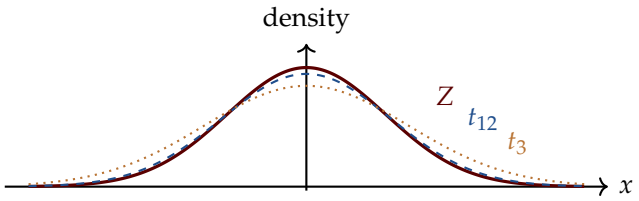


Figure 11.1: The standard normal density and two members of the t-distribution family. Smaller degrees of freedom produce heavier tails and a shorter central peak.

WATCH OUT – The "$n > 30$" myth

Generations of students have been told to use z whenever $n > 30$ and t otherwise. There was a time, before software, when this rule saved table lookups—t_{30} was close enough to Z that a single appendix table could do double duty. It is no longer useful. Always use `t.test()`. The t-distribution converges to Z on its own; there is no cliff at $n = 30$. If you have been reaching for z-procedures out of habit, stop. R does not get tired, and neither does `t.test()`.

11.2 The One-Sample t-Test

With the test statistic in hand, the logic of a hypothesis test follows the template we established in Chapter 6 and refined in Chapter 10. Nothing new is required—only a new reference distribution.

RESULT – The one-sample t-test

To test $H_0 : \mu = \mu_0$ against any of the three alternatives $H_a : \mu \neq \mu_0$, $\mu > \mu_0$, or $\mu < \mu_0$, compute

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \qquad \text{df} = n - 1,$$

and obtain the p-value from the t_{n-1} reference distribution according to the direction of H_a:

Alternative	p-value
$H_a : \mu \neq \mu_0$	$2 \cdot P(T_{n-1} \geq \lvert t_{\text{obs}} \rvert)$
$H_a : \mu > \mu_0$	$P(T_{n-1} \geq t_{\text{obs}})$
$H_a : \mu < \mu_0$	$P(T_{n-1} \leq t_{\text{obs}})$

Reject H_0 when the p-value is smaller than the pre-specified significance level α.

11.2.1 Assumptions

Every inferential procedure rests on assumptions, and the one-sample t-test rests on four. We name them first, then show how to check each in R (Section 11.5).

DEFINITION – Assumptions of the one-sample t-test

1. **Independence.** The observations $X_1, \ldots, X_n$ are independent of one another. This is typically guaranteed by the study design (simple random sampling, random assignment).
2. **Random sampling.** The sample was drawn at random from the population of interest. Without this, $\bar{X}$ is an estimate of a quantity other than the scientific μ you care about.
3. **Normality of the population.** The underlying distribution is approximately normal, *or* the sample is large enough for the Central Limit Theorem to carry the sampling distribution of $\bar{X}$.
4. **No extreme outliers.** Individual points that are far from the body of the data can distort both $\bar{X}$ and s, and therefore the t-statistic.

Item 3 is the one students most often worry about, and it is the one that software cannot fully settle. A Shapiro–Wilk test with $n = 8$ has essentially no power and will rubber-stamp any distribution; the same test with $n = 1000$ will reject any real dataset because real data is never perfectly normal. The right tool is graphical: a Q–Q plot paired with a boxplot. You are looking for catastrophic departures, not for perfection.

11.3 A Fully Worked Example

EXAMPLE – Hatchling mass at Juno Beach

The literature reports that green sea turtle hatchlings from Atlantic nesting beaches have a mean mass of $\mu_0 = 24.8$ g. A biologist studying a single nest at Juno Beach, Florida, weighs $n = 18$ newly emerged hatchlings and records

$$\bar{x} = 26.1 \text{ g}, \qquad s = 2.9 \text{ g}.$$

Is the mean hatchling mass at Juno Beach different from the published population mean?

Step 1 — Hypotheses.
$H_0 : \mu = 24.8, \qquad H_a : \mu \neq 24.8$ (two-sided).

Step 2 — Test statistic.

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{26.1 - 24.8}{2.9/\sqrt{18}} = \frac{1.3}{0.6836} \approx 1.90, \qquad \text{df} = 17.$$

Step 3 — P-value. Under H_0, $t \sim t_{17}$. The two-sided p-value is $2 \cdot P(T_{17} \geq 1.90) \approx 0.074$.

Step 4 — Decision. At $\alpha = 0.05$ we fail to reject H_0. The data do not provide strong enough evidence that Juno Beach hatchlings differ in mean mass from the reported Atlantic population. The result is *suggestive* of a higher local mean—$\bar{x} - \mu_0 = 1.3$ g is about 5% of the reference—and warrants a larger sample.

Step 5 — Scientific translation. The biologist can report: "Our sample of 18 hatchlings had a mean mass 1.3 g higher than the Atlantic reference, but this difference was not statistically significant at the 5% level ($t_{17} = 1.90$, $p = 0.074$). A larger survey is needed before concluding that this nesting site produces heavier hatchlings."

Notice three things about this example. First, the numerical steps are identical to the z-test you learned in Chapter 10; only the reference distribution differs. Second, the four-step structure forces you to keep hypotheses separate from evidence. Third, the scientific translation in Step 5 is not optional. A p-value without a scientific sentence is an unfinished analysis.

11.4 Confidence Intervals for μ

A confidence interval is the natural companion to a test. Where the test asks "is μ_0 plausible?" the interval answers "which values of μ are plausible?". The two perspectives are tightly linked: a value μ_0 lies in the $100(1-\alpha)\%$ two-sided confidence interval for μ if and only if a two-sided t-test of $H_0 : \mu = \mu_0$ fails to reject at level α.

RESULT – One-sample t confidence interval for μ

With the same normality and independence assumptions as the test, a $100(1-\alpha)\%$ confidence interval for the population mean μ is

$$\bar{x} \;\pm\; t^*_{n-1,\,\alpha/2} \cdot \frac{s}{\sqrt{n}},$$

where $t^*_{n-1,\,\alpha/2}$ is the value that cuts off $\alpha/2$ in the upper tail of the t_{n-1} distribution.

For the hatchling data, $t^*_{17,0.025} \approx 2.110$ and $s/\sqrt{n} = 0.6836$, so

$$26.1 \pm 2.110 \cdot 0.6836 \;\approx\; 26.1 \pm 1.44 \;\Rightarrow\; (24.66,\ 27.54).$$

This interval covers the reference value $\mu_0 = 24.8$, consistent with our failure to reject H_0 at $\alpha = 0.05$. It also says something the test does not: *which* values of μ are plausible for Juno Beach hatchlings. Given these 18 measurements, values between roughly 24.7 g and 27.5 g are consistent with the data at the 95% level. A larger sample would narrow this interval.

WATCH OUT – Confidence intervals are about μ, not about $\bar{X}$

A common slip is to say "there is a 95% probability that μ lies in the computed interval." In the frequentist framework, μ is a fixed (unknown) number, not a random variable. The interval is the random object; it either contains μ or it does not. The correct sentence is: "the procedure that produced this interval captures μ in 95% of its uses."

11.5 Assumptions and Diagnostics in R

R SECTION – Loading the data and summarizing

We will work with the hatchling measurements in `juno_hatch.csv`. Every dataset used in this book is catalogued in Appendix D. We load it with tidyverse and print summary statistics.

```
library(tidyverse)
set.seed(205)

hatch <- read_csv("juno_hatch.csv")
glimpse(hatch)
hatch |> summarise(n    = n(),
                   mean = mean(mass_g),
                   sd   = sd(mass_g),
                   min  = min(mass_g),
                   max  = max(mass_g))
```

R SECTION – Checking assumptions graphically

A histogram is not enough for $n < 30$; it is too coarse. Pair a `boxplot` (for outliers and symmetry) with a `qqplot` (for normality).

```
# Boxplot: look for skew and extreme points
hatch |> ggplot(aes(y = mass_g)) +
  geom_boxplot(fill = "#EAF3E8", colour = "#2C5B32") +
  theme_minimal() +
  labs(y = "Hatchling mass (g)", title = "Juno Beach hatchlings")

# Q-Q plot: look for straight-line behaviour
hatch |> ggplot(aes(sample = mass_g)) +
  stat_qq(colour = "#8B0000") +
  stat_qq_line(colour = "#1F4E8C", linewidth = 0.7) +
  theme_minimal() +
  labs(x = "Theoretical quantiles", y = "Sample quantiles",
       title = "Normal Q-Q plot")
```

A Q–Q plot is reading the data against a straight line. Minor wiggles are normal. What you are watching for is a clear curve, a long tail pulling away from the line, or an isolated point far from the pattern. Any of those is a reason to pause before running a t-test.

R SECTION – The test and the confidence interval together

The `t.test()` function delivers both. Call it once; read two answers.

```
t.test(hatch$mass_g, mu = 24.8, conf.level = 0.95)
```

Output (trimmed):

```
        One Sample t-test

data:  hatch$mass_g
t = 1.9014, df = 17, p-value = 0.07434
```

```
alternative hypothesis: true mean is not equal to 24.8
95 percent confidence interval:
 24.66 27.54
sample estimates:
mean of x
    26.10
```

The reported test statistic ($t = 1.90$), degrees of freedom (17), and p-value (0.074) reproduce the hand calculation. The 95% confidence interval, [24.66, 27.54], reproduces the interval we constructed above. A single function call, a complete analysis.

11.6 Effect Size

A p-value tells you whether an effect is detectable; it does not tell you whether the effect is scientifically important. A large enough sample can render any non-zero deviation statistically significant, no matter how trivial. An *effect size* reports the magnitude of the deviation in units that do not depend on the sample size. For one-sample t problems, the standard choice is Cohen's d.

DEFINITION – Cohen's d, one-sample case

For a one-sample problem with sample mean $\bar{x}$, sample standard deviation s, and reference value μ_0,

$$d = \frac{\bar{x} - \mu_0}{s}.$$

Cohen's d measures the difference from the reference in units of standard deviation. Conventional thresholds are: $|d| \approx 0.2$ small, 0.5 medium, 0.8 large.

For the hatchling data,

$$d = \frac{26.1 - 24.8}{2.9} \approx 0.45.$$

This is a medium effect. Read together with the p-value of 0.074, the conclusion is: the biologist's sample shows a medium-sized deviation from the reference in the expected direction, but the sample is too small for the evidence to clear the conventional threshold. That is a very different scientific sentence from "we found nothing."

EXAMPLE – The large-sample trap

Suppose a second site reports a nest with $n = 2500$ hatchlings, mean 26.1 g, and SD 2.9 g. The test statistic explodes to

$$t_{\text{obs}} = \frac{26.1 - 24.8}{2.9/\sqrt{2500}} \approx 22.4, \quad p < 10^{-100}.$$

The p-value is astronomical—but Cohen's d is still 0.45. Nothing about the site changed; only the sample size did. The p-value screams *significant*, while the effect size whispers that the site is about half a standard deviation heavier than average. The effect size is the sentence a biologist wants.

11.7 A Complete R Workflow

We close by demonstrating the whole analysis in a single pipeline. This pattern—load, explore, diagnose, test, interpret—is the one we will use in every chapter going forward.

R SECTION – End-to-end: hatchling analysis

```
library(tidyverse)
library(effectsize)       # for cohens_d()
set.seed(205)

# 1. Load and describe
hatch <- read_csv("juno_hatch.csv")
summary_tbl <- hatch |>
  summarise(n = n(),
            mean_g = mean(mass_g),
            sd_g   = sd(mass_g))
summary_tbl

# 2. Diagnose
hatch |> ggplot(aes(sample = mass_g)) +
  stat_qq(colour = "#8B0000") +
  stat_qq_line(colour = "#1F4E8C") +
  theme_minimal()

# 3. Test + CI
fit <- t.test(hatch$mass_g, mu = 24.8)
fit

# 4. Effect size
cohens_d(hatch$mass_g, mu = 24.8)

# 5. A one-line scientific summary
with(fit,
  sprintf("t(%d) = %.2f, p = %.3f, 95%% CI [%.2f, %.2f]",
          parameter, statistic, p.value,
          conf.int[1], conf.int[2]))
```

WATCH OUT – Reporting raw `t.test` output is not enough

A paper that reads "$p = 0.074$, failed to reject H_0" has reported a procedure, not a finding. The reader wants the effect size, the confidence interval, and a sentence that connects the numbers to the science. The R pipeline above produces all three. Write the sentence.

11.8 Chapter Summary

Key formulas

Quantity	Formula
Standard error of $\bar{X}$ (unknown σ)	$\mathrm{SE}(\bar{X}) = s/\sqrt{n}$
Test statistic	$t_{\text{obs}} = (\bar{x} - \mu_0)/(s/\sqrt{n})$, $\quad \text{df} = n - 1$
Two-sided p-value	$2 \cdot P(T_{n-1} \geq \lvert t_{\text{obs}} \rvert)$
$100(1-\alpha)\%$ CI for μ	$\bar{x} \pm t^*_{n-1,\,\alpha/2} \cdot s/\sqrt{n}$
Cohen's d (one sample)	$d = (\bar{x} - \mu_0)/s$

R functions introduced

Function	Purpose
`t.test(x, mu, conf.level)`	one-sample t-test and CI
`stat_qq()` / `stat_qq_line()`	normal Q–Q plot with reference line
`geom_boxplot()`	boxplot for symmetry and outliers
`effectsize::cohens_d()`	effect size for one-sample t

Decision guide

Before running a one-sample t-test, walk this path:

1. Is the scientific question about a single mean compared to a fixed reference? If not, you probably want Chapter 12 (two samples) or Chapter 15 (several groups).
2. Do you have raw data (not just summary statistics)? If yes, diagnose assumptions with a Q–Q plot and a boxplot.
3. Are the observations independent and randomly sampled? If not, a t-test is not the right tool; consult Chapter 13 for violations.
4. Run `t.test()`. Report t, df, p, and the CI.
5. Compute Cohen's d. Interpret both the p-value *and* the effect size.

Key Formulas — Chapter 11

Quantity	Formula	Notes
t-statistic	$t = (\bar{x} - \mu_0)/(s/\sqrt{n})$	$\text{df} = n - 1$
Two-sided p	$2 \cdot P(T_{n-1} \geq \lvert t_{\text{obs}} \rvert)$	
$100(1-\alpha)\%$ CI	$\bar{x} \pm t^*_{n-1,\alpha/2} \cdot s/\sqrt{n}$	
Cohen's d (1-sample)	$d = (\bar{x} - \mu_0)/s$	$\lvert d \rvert$: 0.2/0.5/0.8 = S/M/L

Always report t, df, p, CI, *and* Cohen's d together.

A p-value without an effect size is an unfinished analysis.

EXERCISES

Conceptual.

1. Explain, in two or three sentences, why the heavier tails of t_{n-1} produce larger critical values than Z at any fixed confidence level.

2. A colleague tells you "my sample size is 45, so I can use a z-test." Write a one-sentence response that is polite and correct.

3. Describe a situation where a p-value below 0.001 does *not* imply scientific importance. Describe a situation where a p-value of 0.12 does *not* imply scientific unimportance.

Computational.

4. A sample of $n = 12$ sockeye salmon from a single spawning run has mean length $\bar{x} = 58.3$ cm and $s = 4.1$ cm. Test whether the run's mean length differs from the published value of 56.0 cm at $\alpha = 0.05$. Report t_{obs}, df, the two-sided p-value, and your scientific conclusion.

5. Construct a 99% confidence interval for μ using the sockeye data in the previous exercise. Does your interval cover 56.0 cm? Explain the relationship between your answer and the decision you reached in the test.

6. In a study of resting heart rate among trained endurance athletes, $n = 9$ participants have $\bar{x} = 49.1$ bpm and $s = 5.8$ bpm. Test whether athletes' mean resting heart rate is below the general adult reference of 72 bpm, using a one-sided alternative.

7. A plant biologist measures leaf area (cm^2) on $n = 24$ greenhouse specimens of a single cultivar and obtains $\bar{x} = 42.7$, $s = 6.9$. The cultivar's published mean is 41.0 cm^2. Conduct a two-sided t-test and compute Cohen's d. Write a one-paragraph scientific summary.

R exercises (use the datasets catalogued in Appendix D).

8. Load `juno_hatch.csv`. Reproduce the test and CI in Section 11.4. Add Cohen's d. Write one sentence that a biologist could paste into a paper.

9. Draw 1,000 samples of size $n = 8$ from a standard normal population, compute the t-statistic against $\mu_0 = 0$ for each sample, and overlay a histogram of your simulated statistics with the theoretical t_7 density. Describe what you see.

10. Repeat the previous exercise with $n = 30$ and with $n = 200$. At what sample size does the difference between your simulated histogram and the standard normal density become hard to see by eye?

11. Consider the `plant_co2.csv` dataset. Treat the `uptake` column for the subset `Type == "Quebec"` and `Treatment == "nonchilled"` as a single sample. Test whether the mean uptake differs from the published species reference of 30.0 μmol/m^2·s. Check assumptions graphically before running the test.

12. (Challenge.) Suppose you observe $\bar{x} = \mu_0$ in your sample. Without plugging numbers into a formula, argue why the two-sided p-value must equal 1 and the confidence interval must be centred on μ_0. Confirm your argument with an R simulation.

12

Comparing Two Means

"Comparisons are the lifeblood of inference: the only way to tell whether an effect is real is to see what happens when it is absent."

— Paraphrase of Sir Austin Bradford Hill

Chapter 11 answered questions about a single population mean, compared to a fixed reference. Most scientific questions are not that simple. A biologist wants to know whether lizards run faster at 30°C than at 20°C. A clinician wants to know whether a new blood-pressure medication lowers mean systolic pressure more than the current standard. An ecologist wants to know whether soil nitrogen differs between two watersheds. In each case the scientific question concerns *two* population means, and the fair answer uses information from *two* samples.

Two very different study designs lead to two very different procedures. When the two samples are drawn from separate, independent groups, we use an *independent two-sample t-test*. When each observation in one sample is naturally paired with an observation in the other— measurements on the same subject before and after treatment, twins, or matched controls—we use a *paired t-test*, which is really the one-sample procedure from Chapter 11 applied to the differences. Choosing the right tool is a design question, not an arithmetic question.

LEARNING OBJECTIVES

- Distinguish between an independent and a paired two-sample design, and justify which procedure is appropriate for a given scientific question.
- Carry out an independent two-sample (Welch) t-test in R, report the test statistic, degrees of freedom, and p-value, and interpret them in context.
- Carry out a paired t-test by taking differences and applying the one-sample procedure of Chapter 11.
- Construct and interpret confidence intervals for a difference of means, $\mu_1 - \mu_2$.
- Diagnose the assumptions of each procedure from raw data using Q–Q plots, boxplots, and side-by-side summaries.
- Report effect size (Cohen's d for two means) alongside the p-value.

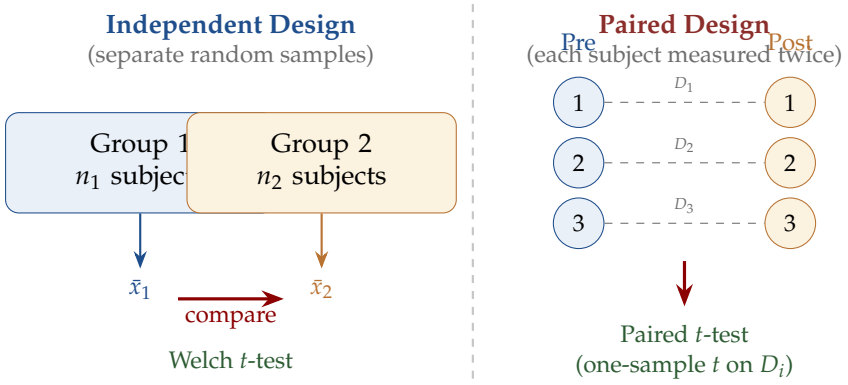


Figure 12.1: Two very different two-sample designs. In the independent design (left), observations in one group have no link to those in the other. In the paired design (right), every subject in one condition is matched to the same subject in the other, producing difference scores $D_i = \text{Pre}_i - \text{Post}_i$ that are analysed with a one-sample t-test.

12.1 From One Mean to Two

In Chapter 11 we had a sample $X_1, \ldots, X_n$ drawn from a single normal population $N(\mu, \sigma)$, and we asked whether μ differed from a fixed reference μ_0. Now we have *two* samples:

$$X_1, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1), \qquad Y_1, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2),$$

and we want to compare μ_1 to μ_2. The natural estimator of $\mu_1 - \mu_2$ is the difference of sample means $\bar{X} - \bar{Y}$, and the question becomes: how variable is that difference?

The sampling distribution of a sum of independent normal random variables is itself normal, with a variance equal to the sum of the variances. Applied to $\bar{X} - \bar{Y}$, this says

$$\text{Var}(\bar{X} - \bar{Y}) \;=\; \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) \;=\; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

If we knew the two population variances, we could standardize by this quantity and use a Z-statistic. We do not. We estimate each σ_i^2 with the sample variance s_i^2, which again forces us off the standard normal curve and onto a t-distribution. The difficulty this time is that the resulting statistic no longer has an exactly-t distribution—and the best fix for that difficulty gives rise to the modern default, Welch's t-test.

12.2 The Independent Two-Sample t-Test

DEFINITION – Welch's two-sample t-statistic

For independent samples with sizes n_1, n_2, sample means $\bar{x}_1, \bar{x}_2$, and sample standard deviations s_1, s_2, the *Welch t-statistic* for testing $H_0 : \mu_1 = \mu_2$ is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}.$$

Under H_0, t has *approximately* a t-distribution whose degrees of freedom are given by the Welch–Satterthwaite formula

$$\text{df} = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}}.$$

The degrees of freedom are typically not an integer; software reports them to two decimals.

The Welch procedure does not assume equal population variances. That matters: the older pooled-variance t-test, which *does* assume $\sigma_1 = \sigma_2$, breaks down badly when the assumption is wrong, especially with unequal sample sizes. Welch behaves almost identically to the pooled test when variances truly are equal and much more reliably when they are not. For this reason, modern statistical practice—and R's `t.test()` default—is Welch.

WATCH OUT – Pooled vs. Welch

Some older textbooks teach the pooled-variance t-test as the "main" two-sample procedure and introduce Welch only as a correction. That order is inverted in practice. Unless you have a strong scientific reason to believe the two populations have identical variances—a belief that is nearly impossible to justify for real data—use Welch. If you see `var.equal = TRUE` in someone else's code, ask why.

RESULT – Independent two-sample t-test

To test $H_0 : \mu_1 = \mu_2$ against the two-sided alternative $H_a : \mu_1 \neq \mu_2$ (or the appropriate one-sided alternative):

1. Verify independence within and between samples.
2. Check approximate normality within each sample (Q–Q plot, boxplot) or rely on the CLT when the sample sizes are large.
3. Compute t_{obs} and df using the Welch formulas.
4. Report the p-value from t_{df} and make a decision at the chosen α.

12.3 A Fully Worked Example (Independent)

EXAMPLE – Sprint speed in side-blotched lizards

A herpetologist measures maximum sprint speed (m/s) on two populations of the side-blotched lizard (*Uta stansburiana*): one from a high-elevation site and one from a low-elevation site. The samples are independent; no animal appears in both groups.

Summary statistics.

Site	n	$\bar{x}$ (m/s)	s (m/s)
High elevation	15	2.34	0.41
Low elevation	18	2.71	0.37

Step 1 — Hypotheses.
$H_0 : \mu_{\text{high}} = \mu_{\text{low}}$, $H_a : \mu_{\text{high}} \neq \mu_{\text{low}}$.

Step 2 — Welch t-statistic.

$$t_{\text{obs}} = \frac{2.34 - 2.71}{\sqrt{\dfrac{0.41^2}{15} + \dfrac{0.37^2}{18}}} = \frac{-0.37}{\sqrt{0.0112 + 0.0076}} = \frac{-0.37}{0.1371} \approx -2.70.$$

Step 3 — Degrees of freedom (Welch).

$$\text{df} = \frac{(0.0112 + 0.0076)^2}{\dfrac{(0.0112)^2}{14} + \dfrac{(0.0076)^2}{17}} \approx 28.6.$$

Step 4 — P-value. With $t_{\text{obs}} \approx -2.70$ and df ≈ 28.6, the two-sided p-value is about 0.012.

Step 5 — Decision + scientific sentence. At $\alpha = 0.05$ we reject H_0. The high-elevation population runs significantly slower on average than the low-elevation population ($t_{28.6} = -2.70$, $p = 0.012$, $\bar{x}_{\text{high}} - \bar{x}_{\text{low}} = -0.37$ m/s). This is consistent with the hypothesis that lower oxygen availability and cooler temperatures at altitude reduce peak locomotor performance in ectotherms.

12.4 The Paired t-Test

Not every two-sample problem is independent. When each observation in one sample is naturally linked to a specific observation in the other, the samples are *paired*, and an independent-samples procedure is wrong—it throws away the pairing, inflates the standard error, and costs power.

DEFINITION – Paired design

A design is *paired* when observations come in linked units (X_i, Y_i) for $i = 1, \ldots, n$, and the pairing is imposed by the study: same subject measured twice, littermates, twins, plots matched on soil and aspect, etc. The natural test statistic is the one-sample t-statistic applied to the sample of differences $D_i = X_i - Y_i$.

RESULT – Paired t-test

Let $D_i = X_i - Y_i$ for $i = 1, \ldots, n$. To test $H_0 : \mu_D = 0$ against a chosen alternative, compute

$$t_{\text{obs}} = \frac{\bar{d} - 0}{s_D / \sqrt{n}}, \qquad \text{df} = n - 1,$$

where $\bar{d}$ and s_D are the sample mean and standard deviation of the differences. The reference distribution is t_{n-1}, exactly as in Chapter 11.

The paired test is the one-sample test in disguise. Nothing is new mathematically; everything is new about the study design. Students sometimes worry that the differences "throw away information." They do not: they consolidate the information that actually matters— the within-subject change—and discard the between-subject variation that the pairing was designed to remove.

EXAMPLE – Blood pressure before and after a beta-blocker

Twelve hypertensive patients have their systolic blood pressure measured at baseline and again four weeks after starting a new beta-blocker. Because each patient contributes two measurements, the design is paired. The differences $D_i = \text{pre}_i - \text{post}_i$ have sample mean $\bar{d} = 11.3$ mm Hg and sample standard deviation $s_D = 6.8$ mm Hg.

Hypotheses. $H_0 : \mu_D = 0$, $H_a : \mu_D > 0$ (the drug lowers pressure, so pre $>$ post).

Test statistic.

$$t_{\text{obs}} = \frac{11.3 - 0}{6.8/\sqrt{12}} \approx 5.76, \qquad \text{df} = 11.$$

P-value. The one-sided p-value is $P(T_{11} \geq 5.76) < 0.001$.

Conclusion. We reject H_0 and conclude that the drug lowers mean systolic blood pressure by a substantial amount. A 95% paired CI for the decrease is $11.3 \pm t^*_{11,0.025} \cdot 6.8/\sqrt{12} \approx (7.0, 15.6)$ mm Hg.

WATCH OUT – Paired vs. independent: a design question

The choice between the paired and the independent t-test is *determined by the study design*, not by the data you end up with. If your study paired observations, you must use the paired

test. Applying an independent-samples test to paired data invalidates the independence assumption (pre and post measurements on the same person are correlated) and typically under-powers the analysis.

12.5 Confidence Intervals for $\mu_1 - \mu_2$

Every test has a companion confidence interval. For the independent (Welch) case,

$$(\bar{x}_1 - \bar{x}_2) \;\pm\; t^{*}_{\mathrm{df},\,\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where the degrees of freedom come from the Welch–Satterthwaite formula. For the paired case, the CI is the one-sample interval for μ_D from Chapter 11 applied to the differences.

Applying this to the lizard example (Welch), the 95% CI for the difference $\mu_{\text{high}} - \mu_{\text{low}}$ is approximately

$$-0.37 \;\pm\; 2.048 \cdot 0.1371 \;\Rightarrow\; (-0.65,\; -0.09)\ \text{m/s}.$$

The interval does not cover zero, agreeing with the test's rejection of H_0. It also quantifies the range of differences consistent with the data: the high-elevation population's mean speed is plausibly between 0.1 and 0.7 m/s slower than the low-elevation population's.

12.6 Assumptions and Diagnostics in R

The assumptions of the Welch test are four:

1. Independence of observations within *and* between samples. Designs that violate this (paired data, clustered samples, repeated measures) require different procedures.
2. Random sampling from each population.
3. Approximate normality within each sample, or large enough n_1 and n_2 for the CLT.
4. Absence of extreme outliers in either sample.

Equal variances is not on the list. Welch does not assume it.

R SECTION – Side-by-side diagnostics

Build your diagnostic pipeline from two small plots: a boxplot to compare spread and spot outliers, and a Q–Q plot per group. Both are one-liners in ggplot2.

```
library(tidyverse)
set.seed(205)

lizards <- read_csv("lizard_speed.csv")    # columns: site, speed_ms

# Boxplot: centre, spread, and any outliers
lizards |> ggplot(aes(x = site, y = speed_ms, fill = site)) +
  geom_boxplot(alpha = 0.8, colour = "#4A5662") +
  scale_fill_manual(values = c("#EAF0F8", "#FBF2E0")) +
  theme_minimal() +
  labs(y = "Sprint speed (m/s)", x = NULL)

# Q-Q plots, one per site
lizards |> ggplot(aes(sample = speed_ms, colour = site)) +
  stat_qq() + stat_qq_line() +
  facet_wrap(~ site) +
  theme_minimal() +
  theme(legend.position = "none")
```

R SECTION – The Welch test and CI, one call

R's `t.test()` defaults to Welch. Supply a formula (`response ~ group`) to compare two groups.

```
t.test(speed_ms ~ site, data = lizards)
```

Output (trimmed):

```
        Welch Two Sample t-test

data:  speed_ms by site
t = -2.70, df = 28.6, p-value = 0.012
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.65 -0.09
sample estimates:
```

```
mean in group high  mean in group low
              2.34               2.71
```

R SECTION – Paired test from long data

If your data are in long format (two rows per subject), reshape them with tidyr before differencing. If they are already wide (pre/post columns on the same row), use `t.test()` directly.

```r
bp <- read_csv("bp_pre_post.csv")   # columns: subject, pre, post

# Wide-form: give t.test() two vectors with paired = TRUE
t.test(bp$pre, bp$post, paired = TRUE, alternative = "greater")

# Or reshape long-form first
bp_long <- bp |>
  pivot_longer(c(pre, post), names_to = "time", values_to = "sbp")
# ...and then fit a mixed model for real repeated-measures problems.
```

12.7 Effect Size

For two-sample problems, the standard effect size is again Cohen's d, using a pooled standard deviation for scale. For independent samples,

$$d \;=\; \frac{\bar{x}_1 - \bar{x}_2}{s_p}, \qquad s_p \;=\; \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}.$$

For paired data, $d = \bar{d}/s_D$ (the one-sample effect size on the differences). Interpret as usual: $|d| \approx 0.2$ small, 0.5 medium, 0.8 large.

For the lizards, $s_p \approx 0.389$, so

$$d \;=\; \frac{2.34 - 2.71}{0.389} \;\approx\; -0.95,$$

a large effect. Read together, "$p = 0.012$, $d \approx -0.95$" is a much more informative sentence than either number alone.

12.8 Chapter Summary

Key formulas

Quantity	Formula
Welch statistic	$t = (\bar{x}_1 - \bar{x}_2)/\sqrt{s_1^2/n_1 + s_2^2/n_2}$
Welch df	Welch–Satterthwaite formula (non-integer)
Paired statistic	$t = \bar{d}/(s_D/\sqrt{n})$, $\text{df} = n - 1$
Welch CI	$(\bar{x}_1 - \bar{x}_2) \pm t^*_{\text{df},\alpha/2}\sqrt{s_1^2/n_1 + s_2^2/n_2}$
Paired CI	$\bar{d} \pm t^*_{n-1,\alpha/2} \cdot s_D/\sqrt{n}$
Cohen's d (indep.)	$(\bar{x}_1 - \bar{x}_2)/s_p$, pooled s_p
Cohen's d (paired)	$\bar{d}/s_D$

R functions introduced

Function	Purpose
`t.test(y ~ group, data)`	Welch two-sample t-test
`t.test(x, y, paired = TRUE)`	paired t-test
`pivot_longer()`	reshape wide paired data to long
`effectsize::cohens_d()`	effect size for two means

Decision guide for two-mean problems

1. Is each observation in one sample linked to a specific observation in the other (same subject, twins, matched pairs)? If yes, use the paired t-test. If no, proceed.
2. Are the two samples independent of each other? If no, you need a different procedure (cluster-robust, mixed model).
3. Check assumptions graphically—boxplots and Q–Q plots.
4. Run `t.test()`. The default (Welch) is the right default. Do not pass `var.equal = TRUE` without a reason.
5. Report t, df, p, the CI for the difference, and Cohen's d.

Key Formulas — Chapter 12

Procedure	Test statistic	df
Welch t	$(\bar{x}_1 - \bar{x}_2)/\sqrt{s_1^2/n_1 + s_2^2/n_2}$	Welch–Satterthwaite
Paired t	$\bar{d}/(s_D/\sqrt{n})$	$n - 1$
Welch CI	$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{s_1^2/n_1 + s_2^2/n_2}$	
Paired CI	$\bar{d} \pm t^*_{n-1} \cdot s_D/\sqrt{n}$	
Cohen's d (indep.)	$(\bar{x}_1 - \bar{x}_2)/s_p$	pooled s_p
Cohen's d (paired)	$\bar{d}/s_D$	

Design question first: paired or independent? Then choose the test.

Default: Welch (no equal-variance assumption). Never `var.equal=TRUE` without reason.

EXERCISES

Conceptual.

1. Explain, in three or four sentences, why the Welch procedure is a more defensible default than the pooled-variance two-sample t-test.
2. A student analyses a pre/post blood-pressure study using `t.test(pre, post)` without `paired = TRUE`. What assumption has been violated, and how would you expect that violation to affect the reported p-value relative to the correct paired analysis?
3. You read in a paper "an independent-samples t-test was used to compare cohorts." What additional information do you need from the methods section before you can evaluate whether the test was appropriate?

Computational.

4. Compute Welch's t-statistic and df for the following summary: $n_1 = 10$, $\bar{x}_1 = 52.4$, $s_1 = 6.1$, $n_2 = 12$, $\bar{x}_2 = 48.9$, $s_2 = 9.8$. Report the two-sided p-value.
5. Six rats are measured for heart rate before and after a standardized exercise protocol. The differences (post − pre) are $+14$, $+22$, $+8$, $+17$, $+11$, $+20$ bpm. Test whether the mean change differs from zero at $\alpha = 0.05$. Report t, df, p, and a 95% CI for the mean change.
6. A pollinator biologist compares flower-visitation rates (visits/hour) for native bees and introduced honeybees at the same site. Native bees: $n = 14$, $\bar{x} = 5.2$, $s = 1.9$. Honeybees: $n = 16$, $\bar{x} = 7.6$, $s = 2.3$. Test for a difference, report the CI, and compute Cohen's d.
7. A biologist measures leaf temperature (°C) at noon on 20 randomly chosen leaves of a plant, then again on the same 20 leaves at 3 pm. The mean difference (noon − 3pm) is -1.8°C with $s_D = 2.7$°C. State whether this is a paired or independent design, run the appropriate test, and interpret.

R exercises (Appendix D).

8. Load `lizard_speed.csv`. Reproduce the test and 95% CI from Section 12.3. Produce the diagnostic plots and comment on the assumptions.

9. Simulate two samples of size $n_1 = 15$ and $n_2 = 15$ from normal populations with equal means but different variances ($\sigma_1 = 1$, $\sigma_2 = 4$). Run a Welch test and a pooled test on the same data. Repeat 1,000 times and report the fraction of false positives for each procedure.

10. Load `bp_pre_post.csv`. Carry out the paired analysis from Section 12.4. Construct a 99% CI instead of the 95% CI and comment on the comparison.

11. (Challenge.) When the sample sizes are equal ($n_1 = n_2$) and the sample variances are equal, show that Welch's t equals the pooled-variance t. What happens when sample sizes are unequal but variances are equal? Support your algebra with a small numerical example in R.

12. An ecologist measures soil nitrogen (mg/kg) in two forest types. Old-growth: $n_1 = 18$, $\bar{x}_1 = 34.6$, $s_1 = 8.2$. Secondary: $n_2 = 22$, $\bar{x}_2 = 28.1$, $s_2 = 11.4$. (a) Carry out a Welch two-sample t-test. (b) Construct a 95% CI for $\mu_1 - \mu_2$. (c) Compute Cohen's d. (d) State why the paired test would be inappropriate here.

13. Twelve patients have their systolic blood pressure recorded before and after 8 weeks of a dietary intervention. Mean change (pre $-$ post) $= 9.2$ mmHg, SD of changes $= 5.7$ mmHg. (a) Test $H_0 : \mu_D = 0$ at $\alpha = 0.05$. (b) Construct a 95% CI for the mean reduction. (c) If someone had incorrectly used an independent two-sample test instead, how would that affect the p-value? Explain conceptually.

14. Using simulation in R, generate two groups of size $n = 15$ from $N(0, 1)$ and $N(0, 4)$ respectively. Run both a Welch test and a pooled test on the same data. Repeat 5,000 times. Report the empirical Type I error rate for each procedure and explain why they differ.

13

Handling Violations of Assumptions

"All models are wrong, but some are useful."

— George E. P. Box

Chapters 11 and 12 developed the one- and two-sample t-tests under an explicit cluster of assumptions: independence, random sampling, approximate normality, no extreme outliers. Real data rarely oblige all four. Clinical measurements are skewed. Ecological counts are bounded below by zero. Psychophysical reaction times have a long right tail. The practical question is not whether your data meet the assumptions—no data meet them exactly—but whether the departures are severe enough to change your scientific conclusion.

Two remedies are available when the departures are serious. The first is a *transformation*: apply a one-to-one function to every observation (log, square root, reciprocal) so that the transformed data sit more comfortably on the normal scale. The second is a *nonparametric test*: abandon the normal-theory machinery altogether and use a procedure that relies only on the ordering of the data. This chapter develops both remedies and—just as importantly—teaches you when *not* to use them.

LEARNING OBJECTIVES

- Diagnose violations of independence, normality, equal variance, and the outlier assumption from raw data.
- Decide whether the Central Limit Theorem rescues the t-test in a given situation, or whether remediation is required.
- Apply log, square-root, and reciprocal transformations and interpret the results on the original scale.
- Carry out the Wilcoxon signed-rank test (paired) and the Mann–Whitney U-test (independent) in R, and report both the W-statistic and the p-value.
- Choose between a transformation, a nonparametric test, and a robust alternative on the basis of the scientific question and the data.

13.1 The Assumptions, Revisited

The t-test has four assumptions, all of which matter, but not equally. *Independence* is the most consequential: when it fails, the nominal p-value and the actual Type I error rate diverge wildly. *Random sampling* is a study-design issue; no statistical procedure can rescue a biased sample. *Normality* and *outliers*, by contrast, are partial concerns: the t-test is reasonably robust to moderate departures, especially at larger sample sizes.

DEFINITION – Robustness

A statistical procedure is *robust* to a particular violation if the actual behaviour of the procedure (its Type I error rate, its power, its confidence coverage) remains close to the nominal behaviour when the violation is present. Robustness is a matter of degree, not a binary property.

RESULT – When the CLT rescues the t-test

For a single sample of size n, the sampling distribution of $\bar{X}$ is approximately normal whenever

- the population is roughly symmetric and $n \gtrsim 15$;
- the population is moderately skewed and $n \gtrsim 30$;
- the population is heavily skewed or bimodal and $n \gtrsim 60$–100;

The widely-repeated "$n > 30$" rule is a reasonable starting point but it is not a guarantee for all shapes.

WATCH OUT – Do not test your way to reassurance

A common mistake is to run a formal normality test (Shapiro–Wilk, Kolmogorov–Smirnov) and decide on that basis whether to use a t-test. Two problems: (1) at small n these tests have almost no power, so the t-test is blessed on data that in fact violates normality badly; (2) at large n any real-world departure from exact normality is detected, so the t-test is condemned precisely when the CLT makes normality irrelevant. The right tool is a Q–Q plot interpreted with judgment.

When a distribution is right-skewed, applying a concave function (log, square root) compresses large values relative to small ones and often renders the transformed distribution approximately normal. Three transformations cover most biological data.

DEFINITION – Common variance-stabilizing transformations

For $x > 0$,

1. **Log transform:** $y = \log(x)$. Appropriate for multiplicative data, counts over a wide dynamic range, and positively-skewed continuous measurements (body mass, enzyme concentrations, gene expression).
2. **Square-root transform:** $y = \sqrt{x}$. Appropriate for count data (individuals per plot) where variance increases with the mean.
3. **Reciprocal transform:** $y = 1/x$. Stabilizes variance when the coefficient of variation

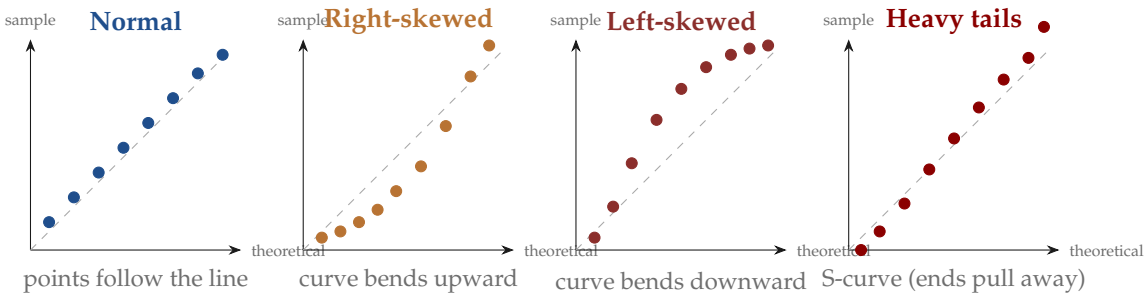


Figure 13.1: Four Q–Q plot patterns to recognize. When data are approximately normal, points fall on the dashed reference line. Right-skew bends the upper tail above the line; left-skew bends it below. Heavy tails produce an S-shaped pattern where both ends deviate from the line in opposite directions.

grows with the mean (rates, waiting times).

EXAMPLE – Log-transforming fish biomass

A fisheries biologist measures catch biomass (kg) across 25 randomly chosen sampling sites in a lake. The raw data range from 0.4 to 312 kg and are strongly right-skewed. A Q–Q plot shows the data curving away from the reference line; a t-test on the raw values would be suspect.

After applying $y = \log_{10}(x)$, the transformed data range from -0.4 to 2.49 and approximate a normal distribution on the log-scale. A one-sample t-test on y against the log of the reference value $\log_{10}(20) \approx 1.30$ is appropriate. If the test yields a 95% CI for μ_y of $(1.02, 1.41)$, the back-transformed interval $(10^{1.02}, 10^{1.41}) \approx (10.5, 25.7)$ is a CI for the *geometric mean* biomass—not the arithmetic mean.

WATCH OUT – Transformations change the parameter you estimate

A t-test on $\log(x)$ does not test a hypothesis about the mean of x. It tests a hypothesis about the mean of $\log(x)$, which on the original scale corresponds to the *geometric* mean, not the arithmetic mean. When you back-transform a confidence interval by exponentiating, label it honestly: "geometric mean" or "median of a log-normal distribution," not "mean."

13.2 Nonparametric Alternatives

When a transformation cannot be found, or when the assumption of normality fails for a reason that transformations cannot fix (multiple modes, heavy ties, ordinal data), switch to a *nonparametric* procedure. These tests replace the observed values with their ranks and base inference on the behaviour of rank sums under the null hypothesis. They assume only that observations are independent and that the distributions under comparison have the same shape.

13.2.1 The Wilcoxon Signed-Rank Test (paired)

RESULT – Wilcoxon signed-rank test

For paired data with differences $D_1, \ldots, D_n$, test $H_0 : \text{median}(D) = 0$ by

1. Discarding zero differences.
2. Ranking $|D_i|$ from smallest to largest.
3. Summing the ranks with positive signs; call this W_+.

Under H_0, W_+ has a known (tabulated) distribution; p-values are obtained from this reference or, for large n, from a normal approximation.

13.2.2 The Mann–Whitney U / Wilcoxon Rank-Sum Test (independent)

RESULT – Mann–Whitney U-test

For two independent samples of sizes n_1 and n_2, test H_0: the two distributions are identical (equivalently, their medians coincide) by

1. Pooling all $n_1 + n_2$ observations and assigning ranks $1, \ldots, n_1 + n_2$.
2. Computing the sum of ranks in sample 1, R_1.
3. Converting to $U = R_1 - n_1(n_1 + 1)/2$.

Under H_0, U has a known distribution; R's `wilcox.test()` returns the p-value.

EXAMPLE – Reaction times across two noise conditions

An experimental psychologist measures reaction time (ms) for 12 participants in a quiet room and 12 participants in a noisy room. The reaction-time distributions are heavily right-skewed. After attempting log and reciprocal transforms without success, the researcher runs a Mann–Whitney test. R reports $W = 41$, $p = 0.008$, and the conclusion is that reaction times tend to be larger in the noisy condition.

WATCH OUT – Rank tests are tests of distributions, not means

A Mann–Whitney test rejects H_0 whenever one distribution is *stochastically larger* than the other. Under the additional assumption that the two distributions have the same shape, this reduces to a statement about medians. If the two distributions differ in shape (one is right-skewed, the other symmetric), a significant Mann–Whitney test is evidence of *some* difference, not necessarily a difference in location.

13.3 Doing It in R

R SECTION – Diagnosing a non-normal sample

```
library(tidyverse)
set.seed(205)

fish <- read_csv("fish_biomass.csv")    # columns: site, biomass_kg

# Raw Q-Q plot: curves away -> skewed
fish |> ggplot(aes(sample = biomass_kg)) +
  stat_qq(colour = "#8B0000") + stat_qq_line() +
  theme_minimal() +
  labs(title = "Raw biomass")

# Log-transformed Q-Q plot
fish |> mutate(log_biomass = log10(biomass_kg)) |>
  ggplot(aes(sample = log_biomass)) +
  stat_qq(colour = "#8B0000") + stat_qq_line() +
  theme_minimal() +
  labs(title = "log10(biomass)")
```

R SECTION – Running a rank-sum test

`wilcox.test()` is the R entry point for Mann–Whitney and Wilcoxon signed-rank tests. Its interface parallels `t.test()`.

```
# Independent samples (Mann-Whitney)
wilcox.test(rt_ms ~ condition, data = rt_data)

# Paired signed-rank test
wilcox.test(bp$pre, bp$post, paired = TRUE, alternative = "greater")

# Confidence intervals for the location shift
wilcox.test(rt_ms ~ condition, data = rt_data, conf.int = TRUE)
```

13.4 Chapter Summary

Remediation options at a glance

Violation	Severity	Tool
Independence	Any	Redesign the study (mixed models, clustered inference)
Random sampling	Any	No statistical fix; discuss limitations
Normality (mild, $n \geq 30$)	Low	Proceed with t-test (CLT)
Normality (strong skew, small n)	High	Transformation or nonparametric test
Outliers	Variable	Investigate first, then robust or nonparametric test
Equal variance (two-sample)	N/A	Welch – no assumption needed

R functions introduced

Function	Purpose
`log()` / `sqrt()` / `1/x`	variance-stabilizing transforms
`wilcox.test(x, y)`	Mann–Whitney (or signed-rank if `paired = TRUE`)
`wilcox.test(x, y, conf.int = TRUE)`	location-shift CI
`shapiro.test()`	formal normality test (use cautiously)

Decision guide

1. Inspect a Q–Q plot and a boxplot per group. Form a visual judgment of how badly normality fails.
2. Ask whether the CLT is likely to rescue you given n and the shape of the data.
3. If not, try a simple transformation. Recheck the Q–Q plot.
4. If no transformation helps, use a nonparametric test.
5. Report what you did and why, in the methods section.

Key Formulas — Chapter 13

Violation	Remedy	When
Independence fails	Redesign / mixed models	Any n
Biased sample	No fix; discuss	Any n
Mild skew	t-test (CLT)	$n \gtrsim 30$
Strong skew, small n	Log/sqrt transform	Check new Q–Q
Heavy outliers	Nonparametric test	When transform fails

Wilcoxon signed-rank (paired): rank $|D_i|$, sum positive ranks W_+

Mann–Whitney U (independent): pool ranks, $U = R_1 - n_1(n_1+1)/2$

Caution: rank tests detect stochastic dominance, not just location shifts.

EXERCISES

Conceptual.

1. Why is the t-test more robust to violations of normality at large n than at small n? Frame your answer in terms of the sampling distribution of $\bar{X}$.
2. Give an example of a biological measurement for which a log transform is likely appropriate, and one for which a square-root transform is likely appropriate. Justify each.
3. Explain why "the Mann–Whitney test is a test of medians" is a true statement only under an additional assumption.

Computational.

4. A sample of 20 plasma concentrations of a hormone has mean $\bar{x} = 420$ ng/mL, median 295 ng/mL, and standard deviation 380 ng/mL. Without plotting, argue which transformation is most plausible and why.
5. Given paired differences $(+4, -2, +7, +13, -1, +5, +9, +3, +8, -4)$, carry out a Wilcoxon signed-rank test by hand (rank the absolute differences, sum positive ranks, consult the table). Report W_+ and your decision at $\alpha = 0.05$ two-sided.
6. You run a Mann–Whitney test on two small samples and obtain $p = 0.03$. A colleague suggests you also run a t-test on the raw data to confirm. Should you? What do you do if the two tests disagree?

R exercises.

7. Load `fish_biomass.csv`. Run a one-sample t-test on the log-transformed data against $\log_{10}(20)$. Back-transform the 95% CI and label it correctly.
8. Load `reaction_time.csv` (columns `condition`, `rt_ms`). Produce boxplots and Q–Q plots per condition. Attempt a log transform. Run both a Welch t-test (on whichever scale is justified) and a Mann–Whitney test, and compare the conclusions.

9. Simulate a two-sample problem in which both distributions have the same median but different variances (e.g. $X \sim N(0,1)$, $Y \sim N(0,4)$, $n_1 = n_2 = 30$). Run `wilcox.test()` on 1,000 replicates. What fraction of the time does the test reject H_0? Explain.

10. An environmental toxicologist measures polycyclic aromatic hydrocarbon (PAH) concentrations (ng/g) in sediment at $n = 18$ sites near an oil refinery. The raw data are strongly right-skewed with values spanning three orders of magnitude. (a) Which transformation would you try first? (b) After log-transformation, $\bar{y} = 2.14$, $s_y = 0.61$. Test $H_0 : \mu_y = \log_{10}(100)$ using a one-sample t-test. (c) Back-transform the 95% CI and label the endpoints correctly (geometric mean? median?).

11. Describe a biological data scenario in which (a) a log-transform succeeds, (b) a square-root transform is more appropriate, and (c) no transformation helps and a nonparametric test is the right choice. Justify each scenario in one sentence.

12. You run both a Welch t-test and a Mann–Whitney test on the same data and get $p = 0.03$ from the t-test and $p = 0.08$ from the Mann–Whitney test. A colleague says: "Since two tests disagree, the result is inconclusive." Explain what the discrepancy most likely indicates about the data, and how you would decide which result to report.

Part IV

Comparing Multiple Groups

15

Comparing Means with ANOVA

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination."

— Sir Ronald A. Fisher

Chapter 12 handled two means. Most interesting biological experiments involve more: three fertilizer regimes, four drug doses, five environmental temperatures, six populations of a species. When the question is whether *any* of the groups differ from one another, the naive strategy—run a t-test on every pair—is both inefficient and mathematically wrong. Fisher's *analysis of variance* (ANOVA) replaces the many pairwise tests with a single global test whose Type I error rate is exactly the one you chose.

One-way ANOVA tests whether the means of $k \geq 3$ populations are all equal against the alternative that at least two differ. It does this by *partitioning* the total variability in the pooled data into a piece that comes from differences *between* group means and a piece that comes from variability *within* groups. If the between-group piece is large relative to the within-group piece, the F-ratio is large and we reject the null. If it is small, the groups' means are plausibly all equal.

LEARNING OBJECTIVES

- Explain why running multiple pairwise t-tests inflates the family-wise Type I error rate, and compute the inflation for k groups at a nominal α.
- Partition the total sum of squares into between-group and within-group components, and write the ANOVA table from summary statistics.
- Carry out a one-way ANOVA in R using `aov()` and interpret the F-statistic, its degrees of freedom, and the p-value.
- State and diagnose the assumptions of one-way ANOVA (independence, approximate normality within groups, homoscedasticity).
- Perform Tukey's HSD post-hoc test and interpret the resulting simultaneous confidence intervals.
- Report effect size for ANOVA using η^2 or ω^2.

15.1 Why Not Just Run Multiple t-Tests?

Suppose you have k groups and decide to compare every pair. With $k = 4$ groups there are $\binom{4}{2} = 6$ pairwise comparisons; with $k = 6$ there are 15. Each test has its own Type I error rate. If the individual tests are conducted at $\alpha = 0.05$ and are *independent*, the probability of at least one false positive across m tests is

$$\Pr(\text{at least one false positive}) \;=\; 1 - (1 - \alpha)^m.$$

For $m = 6$ this is $1 - 0.95^6 \approx 0.265$—more than a quarter of the time, the procedure lies at least once when all null hypotheses are true. For $m = 15$ it is over 0.53.

RESULT – The family-wise error rate

The *family-wise Type I error rate* is the probability of at least one false positive across a family of related tests. Running multiple tests at a fixed per-comparison level α inflates the family-wise rate above α. ANOVA sidesteps the problem by replacing the family of pairwise tests with a single global F-test whose level is exactly α.

15.2 The Core Idea: Partitioning Variance

Suppose we have k groups with n_i observations each. Let Y_{ij} denote the jth observation in group i, $\bar{Y}_i$ the mean of group i, $\bar{Y}$ the overall (grand) mean, and $N = \sum_i n_i$ the total sample size.

DEFINITION – ANOVA sums of squares

$$\begin{aligned} \mathrm{SS}_{\text{total}} &= \sum_{i,j} (Y_{ij} - \bar{Y})^2, \\ \mathrm{SS}_{\text{between}} &= \sum_{i} n_i (\bar{Y}_i - \bar{Y})^2, \\ \mathrm{SS}_{\text{within}} &= \sum_{i,j} (Y_{ij} - \bar{Y}_i)^2. \end{aligned}$$

The partition identity,

$$\mathrm{SS}_{\text{total}} \;=\; \mathrm{SS}_{\text{between}} + \mathrm{SS}_{\text{within}},$$

holds exactly: total variability is the sum of variability among groups and variability within groups.

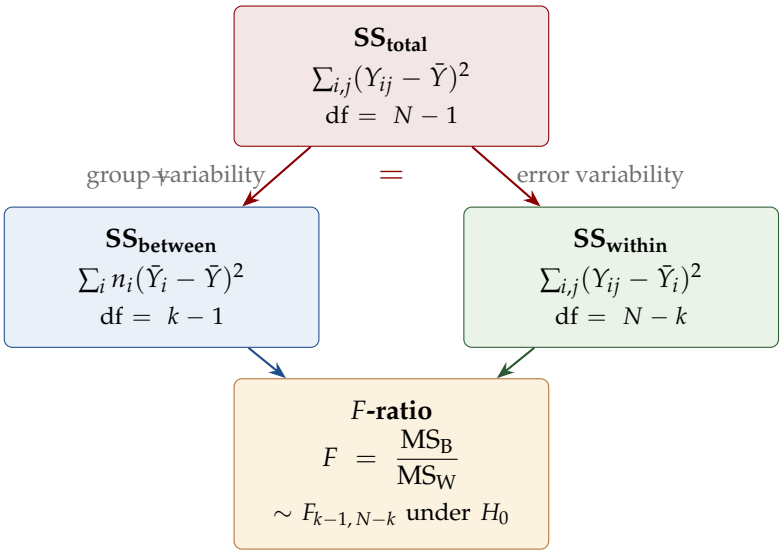


Figure 15.1: The ANOVA variance partition. Total variability ($\mathrm{SS}_{\text{total}}$) splits exactly into between-group variability ($\mathrm{SS}_{\text{between}}$, driven by differences among group means) and within-group variability ($\mathrm{SS}_{\text{within}}$, driven by individual variation inside each group). The F-ratio is the ratio of the corresponding mean squares.

Each sum of squares has its own degrees of freedom—$N - 1$ for total, $k - 1$ for between, $N - k$ for within—and dividing a sum of squares by its df yields a *mean square*. The ratio of the two mean squares is the F-statistic.

RESULT – The one-way F-statistic

Under the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$,

$$F \;=\; \frac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}} \;=\; \frac{\text{SS}_{\text{between}}/(k-1)}{\text{SS}_{\text{within}}/(N-k)} \;\sim\; F_{k-1,\,N-k}.$$

Large F is evidence against H_0; the p-value is the upper tail of the $F_{k-1,N-k}$ distribution.

15.3 A Fully Worked Example

EXAMPLE – Seedling growth under three light regimes

A botanist randomly assigns $n_i = 10$ seedlings of a single species to each of three light regimes: low, medium, and high. After four weeks she measures stem length (cm). The summary statistics are:

Regime	n_i	$\bar{y}_i$	s_i
Low	10	6.1	1.05
Medium	10	8.4	1.18
High	10	7.2	1.10

The grand mean is $\bar{y} = (6.1 + 8.4 + 7.2)/3 = 7.233$.

Between-group SS.

$$\text{SS}_{\text{between}} = 10\big[(6.1 - 7.233)^2 + (8.4 - 7.233)^2 + (7.2 - 7.233)^2\big] = 26.47.$$

Within-group SS. $(n_i - 1)s_i^2$ summed over groups:

$$\text{SS}_{\text{within}} = 9(1.05^2 + 1.18^2 + 1.10^2) = 31.34.$$

Mean squares and F.

$$\text{MS}_{\text{between}} = 26.47/2 = 13.24, \quad \text{MS}_{\text{within}} = 31.34/27 = 1.16, \quad F = 13.24/1.16 \approx 11.41.$$

P-value. With $F_{2,27}$, the upper-tail p-value for $F = 11.41$ is < 0.001.

Conclusion. At $\alpha = 0.05$ we reject H_0. Mean stem length differs among the three light regimes ($F_{2,27} = 11.41$, $p < 0.001$). A post-hoc test (next section) will identify *which* regimes differ.

15.4 Assumptions and Diagnostics

One-way ANOVA assumes three things:

1. Independence of observations within and between groups.
2. Approximate normality within each group, or large enough n_i for the CLT.
3. Homogeneity of variance (homoscedasticity): the k populations share a common variance.

ANOVA is moderately robust to the second assumption, especially for balanced designs (equal n_i). It is less robust to the third; a gross violation of homoscedasticity, particularly with unbalanced designs, can inflate the Type I error rate. When homoscedasticity fails, use Welch's ANOVA (`oneway.test()`), which is to ANOVA what the Welch t-test is to the pooled-variance t-test.

15.5 Post-hoc Comparisons: Tukey's HSD

A significant ANOVA answers "at least two means differ," not "which ones." For the scientific question you usually care about identifying the pairs. Simply running many pairwise t-tests re-introduces the multiple-testing problem we originally solved. The canonical fix is *Tukey's Honestly Significant Difference* (HSD) procedure.

RESULT – Tukey's HSD

For a balanced one-way ANOVA with k groups and common sample size n, Tukey's HSD constructs simultaneous $100(1-\alpha)\%$ confidence intervals for every pairwise mean difference $\mu_i - \mu_j$ using the studentized range distribution. The family-wise coverage probability is exactly $1-\alpha$.

For the seedling example, Tukey's HSD at 95% family-wise confidence reports intervals approximately

$$\text{Medium} - \text{Low} \;:\; (1.0,\, 3.6), \quad \text{High} - \text{Low} \;:\; (-0.2,\, 2.4), \quad \text{Medium} - \text{High} \;:\; (0.0,\, 2.4).$$

The first interval excludes zero; the second includes it; the third sits at the boundary. Interpretation: medium light produces significantly longer stems than low light (with family-wise 95% confidence), but low vs. high and medium vs. high are less clear from this sample.

15.6 Effect Size

For ANOVA, effect size asks: what fraction of total variability is explained by group membership? The natural estimator is

$$\eta^2 \;=\; \frac{\mathrm{SS}_{\text{between}}}{\mathrm{SS}_{\text{total}}}.$$

For the seedling data, $\eta^2 = 26.47/(26.47 + 31.34) \approx 0.46$, meaning 46% of total variability in stem length is explained by light regime. Cohen's conventions for η^2 are 0.01 small, 0.06 medium, 0.14 large; by this scale, the seedling effect is large.

An unbiased alternative is Hays's ω^2, which corrects for sampling variability:

$$\omega^2 = \frac{\mathrm{SS}_{\text{between}} - (k-1)\mathrm{MS}_{\text{within}}}{\mathrm{SS}_{\text{total}} + \mathrm{MS}_{\text{within}}}.$$

15.7 Doing It in R

R SECTION – Fitting the ANOVA and printing the table

```
library(tidyverse)
set.seed(205)

seedlings <- read_csv("seedling_growth.csv")    # columns: regime,
   stem_cm
fit <- aov(stem_cm ~ regime, data = seedlings)
summary(fit)
```

Output:

```
             Df Sum Sq Mean Sq F value Pr(>F)
regime        2  26.47   13.24   11.41 0.000  ***
Residuals    27  31.34    1.16
```

R SECTION – Post-hoc and effect size

```
TukeyHSD(fit)              # simultaneous pairwise CIs
plot(TukeyHSD(fit))        # visual display

library(effectsize)
eta_squared(fit)           # reports eta-squared with a 95% CI
```

R SECTION – Diagnostics

```
# Boxplot by group
seedlings |> ggplot(aes(x = regime, y = stem_cm, fill = regime)) +
  geom_boxplot(alpha = 0.8) +
  theme_minimal()

# Residuals vs fitted (for homoscedasticity) and Q-Q (for normality)
par(mfrow = c(1, 2))
plot(fit, which = 1:2)

# If variances differ, use Welch ANOVA
oneway.test(stem_cm ~ regime, data = seedlings, var.equal = FALSE)
```

15.8 Chapter Summary

The ANOVA table

Source	SS	df	MS	F
Between groups	SS_B	$k-1$	MS_B	$\text{MS}_\text{B}/\text{MS}_\text{W}$
Within groups	SS_W	$N-k$	MS_W	
Total	SS_T	$N-1$		

R functions introduced

Function	Purpose
`aov(y ~ group, data)`	fit one-way ANOVA
`summary(fit)`	print the ANOVA table
`TukeyHSD(fit)`	simultaneous pairwise CIs
`oneway.test(..., var.equal = FALSE)`	Welch ANOVA (unequal variances)
`effectsize::eta_squared(fit)`	effect size

Decision guide

1. Is the scientific question about three or more group means? If two, use Chapter 12.
2. Check assumptions: independence, within-group normality, equal variance.
3. Run `aov()`. Inspect the ANOVA table and F-statistic.
4. If p is small, run Tukey's HSD to identify which pairs differ.
5. Report F, its df, p, the pairwise differences, and η^2.

Key Formulas — Chapter 15

Source	SS	df
Between groups	$\sum_i n_i(\bar{Y}_i - \bar{Y})^2$	$k-1$
Within groups	$\sum_{i,j}(Y_{ij} - \bar{Y}_i)^2$	$N-k$
Total	$\sum_{i,j}(Y_{ij} - \bar{Y})^2$	$N-1$
F-statistic	$\text{MS}_B / \text{MS}_W$	$F_{k-1,N-k}$ under H_0
η^2	$\text{SS}_B / \text{SS}_T$	Effect size
Family-wise error	$1-(1-\alpha)^m$	m comparisons

Post-hoc: Tukey HSD gives simultaneous CIs at exact family-wise level.

Unequal variances: use `oneway.test(var.equal=FALSE)` (Welch ANOVA).

EXERCISES

Conceptual.

1. Compute the family-wise Type I error rate when running 10 independent tests at $\alpha = 0.05$. Why is this a problem for the "just do pairwise t-tests" approach?
2. Explain, in one or two sentences each, why ANOVA's F-test can detect differences that any single pairwise t-test would miss, and why a significant F-test does not tell you *which* groups differ.
3. Under what circumstances would you use Welch's ANOVA instead of classical ANOVA?

Computational.

4. Three fertilizer treatments are applied to $n_i = 8$ plots each. Summary: $\bar{y}_A = 22.1$, $\bar{y}_B = 26.8$, $\bar{y}_C = 24.3$; $s_A = 3.0$, $s_B = 3.4$, $s_C = 2.8$. Compute $\text{SS}_{\text{between}}$, $\text{SS}_{\text{within}}$, mean squares, and the F-statistic. Use an F-table to bound the p-value.
5. For the seedling example in the chapter, show that η^2 equals $(\text{SS}_{\text{between}})/(\text{SS}_{\text{total}})$. What is ω^2?
6. Explain why Tukey's HSD intervals are wider than unadjusted pairwise t-intervals for the same data.

R exercises.

7. Load `seedling_growth.csv`. Run `aov()` and `TukeyHSD()`. Reproduce the ANOVA table in the chapter and the pairwise CIs.
8. Using the CO_2 uptake dataset from Chapter 11, Exercise 11, subset to a single plant type and treat the three uptake regimes as groups. Run ANOVA and post-hoc testing. Comment on the assumptions.

9. Simulate $k = 5$ groups with equal means and $n = 20$ each at $\alpha = 0.05$. Repeat 1,000 times and record the Type I error rate for (a) a single ANOVA and (b) running all 10 pairwise t-tests and reporting any significant difference. Compare the empirical rates.

10. A pharmacologist tests four drug doses (0, 10, 20, 40 mg/kg) on blood glucose reduction in rats ($n_i = 8$ per group). Summary: $\bar{y}_1 = 2.1$, $\bar{y}_2 = 5.8$, $\bar{y}_3 = 9.4$, $\bar{y}_4 = 11.2$ mmol/L; pooled $s = 3.1$. (a) Compute the grand mean and $\text{SS}_{\text{between}}$. (b) Compute $\text{SS}_{\text{within}}$ and the ANOVA table. (c) State the p-value and run Tukey HSD to identify which dose pairs differ significantly.

11. A plant ecologist grows three grass species under identical conditions. Leaf nitrogen content (%) is measured in $n = 15$ plants per species. ANOVA gives $F_{2,42} = 7.34$, $p = 0.002$, $\eta^2 = 0.26$. (a) Interpret η^2. (b) The researcher reports "Tukey HSD shows Species A differs from B ($p < 0.05$) but not from C." What additional information should be reported alongside these comparisons?

12. Explain the relationship between ANOVA and the two-sample Welch t-test when $k = 2$. Show that $F = t^2$ in the equal-variance case (pooled), but that this relationship breaks down for the Welch version. Why?

13. (Challenge) ANOVA assumes homoscedasticity. Simulate three groups with the same means but different standard deviations ($\sigma = 1, 3, 6$, $n = 15$ per group). Run both classical ANOVA and Welch ANOVA (`oneway.test`) on 2,000 replicates. Compare the Type I error rates of the two procedures when the means are equal.

Part V

Relationships Between Variables

16

Correlation Between Variables

"Correlation does not imply causation—but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'."

— Randall Munroe, *xkcd*

Up to this point the response variable has been univariate: we asked whether the mean of one quantity differs between groups, or from a reference. Many biological questions are about the *relationship* between two quantities. Does a bird's wing length increase with its body mass? Does blood pressure rise with sodium intake? Does a plant's leaf area scale with the amount of incident light it receives? The first step in answering such questions is to quantify how tightly the two variables move together. That quantity is the *correlation coefficient*.

LEARNING OBJECTIVES

- Sketch and interpret scatterplots of bivariate numerical data.
- Define the Pearson correlation coefficient r and compute it from raw data or summary statistics.
- Distinguish Pearson (linear), Spearman (rank, monotone), and Kendall's τ (rank, concordance) correlations, and choose appropriately.
- Carry out inference on the population correlation ρ using `cor.test()` and interpret the test statistic, p-value, and confidence interval.
- Explain why correlation is not causation, and name at least three mechanisms that produce correlation without causation.

16.1 Scatterplots First

Before you compute any numbers, *plot* the data. A scatterplot puts one variable on each axis and places one point per observation. Four features of a scatterplot tell you most of what the correlation coefficient summarizes numerically:

1. **Direction.** Does the cloud slope up, slope down, or show no systematic tilt?

2. **Form.** Is the pattern approximately linear, curvilinear, or something more exotic?
3. **Strength.** Are the points tightly packed around a line, or loosely scattered?
4. **Outliers.** Are there isolated points that do not follow the pattern?

Correlation coefficients quantify direction and strength, but only for the particular form they are designed for (usually linear). A nonlinear relationship can be very strong and yet have a Pearson correlation near zero.

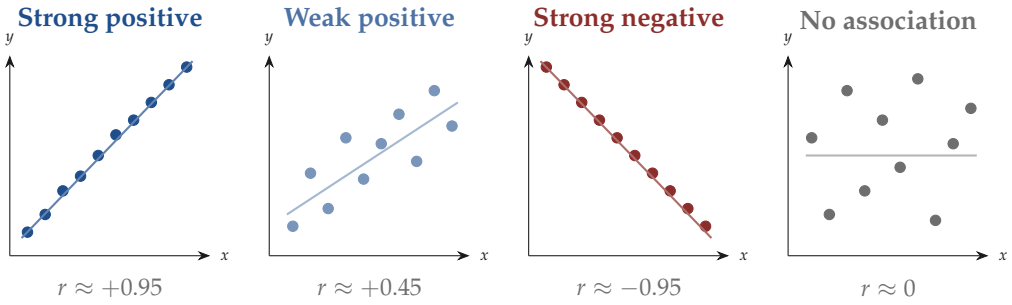


Figure 16.1: Four scatterplot patterns corresponding to different Pearson correlations. Direction is captured by the sign of r; strength by its magnitude. Always plot the data before reporting r — a nonlinear relationship can have $r \approx 0$ even when the two variables are tightly related.

16.2 The Pearson Correlation Coefficient

DEFINITION – Pearson correlation coefficient

For paired observations (x_i, y_i), $i = 1, \ldots, n$, the Pearson correlation coefficient is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

It is a dimensionless number between -1 and $+1$. Values near ± 1 indicate a strong linear association, $r = 0$ indicates no linear association, and the sign records the direction.

An equivalent form of r is the mean product of z-scores:

$$r = \frac{1}{n-1}\sum_{i} z_{x_i} z_{y_i},$$

where z_{x_i} and z_{y_i} are the standardized values. This form reveals why r is dimensionless: it measures average agreement between the standardized versions of the two variables.

WATCH OUT – Linear means linear

r answers one question: how close is the cloud to a straight line? If the true relationship is quadratic, sinusoidal, or otherwise non-linear, r can be misleadingly small even when the two variables carry strong information about each other. Always plot the data before reporting r.

16.3 Inference for ρ

RESULT – Testing $H_0 : \rho = 0$

Under H_0: the population correlation is zero, and X and Y are independent bivariate normal, the test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

follows a t_{n-2} distribution. R's `cor.test()` reports this test statistic, its df, the p-value, and a confidence interval for ρ (via Fisher's z-transform).

EXAMPLE – Body mass and wing length in finches

A researcher measures body mass (g) and wing length (mm) for $n = 28$ Galapagos finches from a single island. The sample correlation is $r = 0.82$. Test $H_0 : \rho = 0$:

$$t = \frac{0.82\sqrt{26}}{\sqrt{1-0.67}} \approx 7.27, \qquad \text{df} = 26, \qquad p < 0.001.$$

There is strong evidence of a positive linear association between body mass and wing length. A 95% CI for ρ is approximately $(0.65, 0.91)$: the population correlation is plausibly between 0.65 and 0.91, consistent with a tight linear relationship.

16.4 Rank-Based Correlations

When the relationship is monotone but nonlinear (doubling one variable multiplies the other by a constant factor, for example), Pearson's r understates the association. Two rank-based alternatives are available.

DEFINITION – Spearman's ρ_s and Kendall's τ

Spearman's ρ_s is the Pearson correlation of the *ranks* of x and y. Kendall's τ counts pairs of observations that are *concordant* (both increase) vs. *discordant* (one increases while the other decreases). Both equal $+1$ for any strictly increasing relationship and -1 for any strictly decreasing one, regardless of linearity.

Spearman and Kendall are also more robust to outliers because ranks are insensitive to the magnitude of extreme values.

16.5 Correlation is Not Causation

A strong correlation between X and Y is compatible with at least five scientific explanations:

1. X causes Y.
2. Y causes X.
3. A third variable Z causes both (*confounding*).
4. X and Y are both outcomes of a selection process (*selection bias*).
5. The correlation is a sampling artifact (low power against the null of zero correlation).

Quantifying the correlation tells you *that* the variables are associated. Identifying *which* of these mechanisms is at work is a scientific exercise informed by study design, domain knowledge, and often an experiment.

16.6 Doing It in R

R SECTION – Scatterplot + fit line

```
library(tidyverse)
set.seed(205)

finches <- read_csv("finch_measurements.csv")     # columns: mass_g,
    wing_mm

finches |> ggplot(aes(x = mass_g, y = wing_mm)) +
  geom_point(colour = "#8B0000", alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE, colour = "#1F4E8C") +
  theme_minimal() +
  labs(x = "Body mass (g)", y = "Wing length (mm)")
```

R SECTION – Correlation coefficients and tests

```
cor(finches$mass_g, finches$wing_mm)                    # Pearson r
cor(finches$mass_g, finches$wing_mm, method = "spearman")
cor(finches$mass_g, finches$wing_mm, method = "kendall")

cor.test(finches$mass_g, finches$wing_mm)               # t-test for rho
    = 0
```

Key Formulas — Chapter 16

Quantity	Formula	Notes
Pearson r	$\sum(x_i - \bar{x})(y_i - \bar{y}) / [(n-1)s_x s_y]$	Linear; $-1 \le r \le 1$
Test stat for $\rho = 0$	$t = r\sqrt{n-2}/\sqrt{1-r^2}$	t_{n-2} under H_0
Spearman r_s	Pearson r of ranks	Monotone relationships
Kendall τ	(concordant $-$ discordant) / $\binom{n}{2}$	Robust to ties

Always plot first. $r = 0$ does not imply independence (nonlinear association possible).

Causation: correlation supports none of: $X \to Y$, $Y \to X$, $Z \to$ both, selection.

EXERCISES

1. Sketch four scatterplots that all have approximately the same Pearson r but look qualitatively different (this is the classic Anscombe's Quartet). Explain what each plot teaches you beyond the correlation coefficient.

2. For the following pairs of variables, predict the sign of the Pearson correlation and whether you would expect it to be strong, moderate, or weak: (a) height and weight of college students; (b) study time and exam score; (c) daily temperature and the number of hot chocolates sold.

3. A researcher reports "$r = 0.6$, $p = 0.03$." Write the scientific interpretation in one sentence, then a one-sentence caveat about causation.

4. Compute Pearson's r by hand for the pairs $(1,2)$, $(2,5)$, $(3,6)$, $(4,8)$, $(5,10)$.

5. Using `cor.test()`, find the 95% CI for ρ given $r = 0.6$, $n = 25$. How would the CI change if $n = 100$?

6. A marine biologist measures fin length (mm) and swimming speed (m/s) in $n = 40$ dolphins. She reports $r = 0.71$. (a) Test $H_0 : \rho = 0$ and report the t-statistic, df, and p-value. (b) Compute a 95% CI for ρ using Fisher's z-transform (available via `cor.test()`). (c) She claims fin length *causes* faster swimming. Evaluate this claim.

7. Describe a dataset where Pearson $r \approx 0$ but Spearman r_s is large and positive. Explain what this tells you about the relationship between the two variables.

8. An exercise physiologist finds $r = 0.82$ between $\dot{V}O_2$max and weekly training mileage in $n = 25$ athletes. (a) Is this statistically significant at $\alpha = 0.01$? (b) Interpret r^2. (c) Would you recommend Pearson or Spearman here, given that training mileage data are typically right-skewed?

9. (Challenge) Simulate $n = 30$ pairs (X, Y) where $Y = X^2$ and $X \sim N(0,1)$. Compute Pearson r, Spearman r_s, and Kendall τ. Why does Pearson r severely underestimate the relationship? What does each coefficient actually measure in this case?

17

Simple Linear Regression

"Regression to the mean is a great leveller. It is why the very best and the very worst of any cohort tend to look more ordinary when measured again."

— Francis Galton, paraphrased

Correlation tells you whether two variables move together. Regression goes one step further: it fits a line—or, more generally, a function—to the data so that the line can be used to *predict* one variable from the other. This is the workhorse of applied statistics. Regression underlies standard curves in biochemistry, dose-response modelling in pharmacology, growth curves in ecology, and the entire enterprise of observational epidemiology.

This chapter develops the simplest form, *simple linear regression* (SLR), in which a single numerical predictor X is used to predict a single numerical response Y. We derive the least-squares estimates of slope and intercept, interpret them, place a standard error and confidence interval on the slope, and build a prediction interval for a new observation. Multiple regression—more than one predictor—is the subject of a more advanced course.

LEARNING OBJECTIVES

- Write the simple linear regression model and identify its parameters.
- Derive the least-squares estimates of slope and intercept, and compute them from summary statistics.
- Interpret the slope and intercept in the context of a scientific problem.
- Compute R^2 and explain what it measures.
- Carry out inference on the slope (test and CI) using `lm()` output.
- Produce and interpret residual plots as diagnostic tools.
- Construct confidence and prediction intervals for a new y at a given x.

17.1 The Model

DEFINITION – Simple linear regression model

For paired observations (x_i, y_i), the simple linear regression model is

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2) \text{ i.i.d.}$$

The parameter β_0 is the *intercept* (the expected value of Y when $x = 0$), β_1 is the *slope* (the expected change in Y for a one-unit increase in x), and σ is the residual standard deviation of Y around the line.

The error terms ε_i carry four assumptions, collectively known as LINE: **L**inearity, **I**ndependence, **N**ormality (of residuals), and **E**qual variance of residuals (homoscedasticity). We diagnose them with residual plots.

17.2 Least-Squares Estimation

The *least-squares* estimates $\hat{\beta}_0, \hat{\beta}_1$ minimize the sum of squared vertical distances between the observed y_i and the fitted $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

RESULT – Closed-form solutions

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = r \cdot \frac{s_y}{s_x}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

The fitted line always passes through $(\bar{x}, \bar{y})$.

Notice the compact form of $\hat{\beta}_1$: it is the Pearson correlation r scaled by the ratio of sample standard deviations. Correlation and regression slope are the same information in two different units: r is unit-free, $\hat{\beta}_1$ is in units of Y per unit of X.

17.3 Interpreting Slope and Intercept

EXAMPLE – Wing length vs. body mass in finches

Continuing the finch example from Chapter 16, suppose regression of wing length (mm) on body mass (g) yields

$$\hat{y} = 42.3 + 0.79x.$$

Slope. For each additional gram of body mass, we estimate an additional 0.79 mm of wing length. The units check out: mm per gram.

Intercept. A hypothetical finch with body mass $x = 0$ would have predicted wing length 42.3 mm. Because no finch has body mass zero, this intercept is not biologically meaningful on its own—it is an anchor for the line, not a prediction you would trust. Most real-world intercepts are like this.

Prediction. For a finch of mass 15 g, predicted wing length is $42.3 + 0.79 \cdot 15 = 54.15$ mm.

17.4 Residuals, R^2, and Variance Explained

The *residual* of observation i is $e_i = y_i - \hat{y}_i$. The sum of squared residuals, $\text{SSE} = \sum e_i^2$, measures the total "unexplained" variability. Partitioning variance as in ANOVA,

$$\underbrace{\sum (y_i - \bar{y})^2}_{\text{SST}} \;=\; \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{\text{SSR}} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{\text{SSE}},$$

and defining

$$R^2 \;=\; \frac{\text{SSR}}{\text{SST}} \;=\; 1 - \frac{\text{SSE}}{\text{SST}},$$

we get a number between 0 and 1 that reports the fraction of total Y-variability explained by the line. For SLR, $R^2 = r^2$.

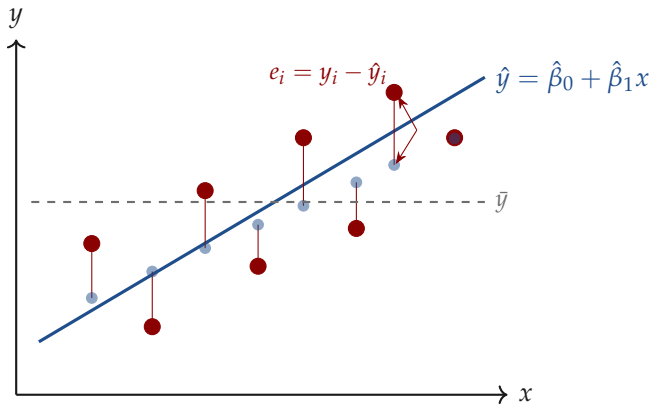


Figure 17.1: A fitted regression line with residuals (red vertical segments). Each residual $e_i = y_i - \hat{y}_i$ measures the vertical distance between the observed point (solid circle) and the fitted value on the line (open circle). Least squares minimizes the sum of squared residuals $\sum e_i^2$.

WATCH OUT – A high R^2 does not guarantee a useful model

R^2 measures the strength of fit, not the correctness of the model. A non-linear relationship can yield a moderate R^2 while the straight line is a poor description. R^2 also says nothing about causation. Always pair it with a residual plot.

17.5 Inference on the Slope

The sampling distribution of $\hat{\beta}_1$ is normal with mean β_1 and standard error

$$\mathrm{SE}(\hat{\beta}_1) \;=\; \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}}, \qquad \hat{\sigma}^2 = \mathrm{SSE}/(n-2).$$

The test statistic $t = \hat{\beta}_1 / \mathrm{SE}(\hat{\beta}_1)$ has a t_{n-2} distribution under $H_0 : \beta_1 = 0$, and the $100(1-\alpha)\%$ CI for β_1 is

$$\hat{\beta}_1 \;\pm\; t^*_{n-2,\alpha/2} \cdot \mathrm{SE}(\hat{\beta}_1).$$

RESULT – Confidence vs. prediction intervals

A *confidence interval* for the mean response at $x = x^*$ estimates $\mathrm{E}(Y \mid X = x^*)$. A *prediction interval* at the same x^* estimates where a single new observation will fall. The prediction interval is always wider because it includes residual variability, not just the uncertainty in the fitted line.

17.6 Doing It in R

R SECTION – Fit, summarize, diagnose

```
library(tidyverse)
set.seed(205)

finches <- read_csv("finch_measurements.csv")
fit <- lm(wing_mm ~ mass_g, data = finches)
summary(fit)            # slope, intercept, std errors, t-stats, p, R^2
confint(fit)            # 95% CIs for intercept and slope
```

Typical `summary()` output:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   42.30       3.21   13.18  < 0.001 ***
mass_g         0.79       0.11    7.19  < 0.001 ***

Multiple R-squared:  0.67,   Adjusted R-squared:  0.66
F-statistic: 51.7 on 1 and 26 DF,  p-value: < 0.001
```

R SECTION – Diagnostic plots

```
par(mfrow = c(2, 2))
plot(fit)   # residuals vs fitted, Q-Q, scale-location, leverage
```

The four standard plots check the four LINE assumptions.

R SECTION – Predictions with uncertainty

```
new_data <- tibble(mass_g = c(12, 15, 18))

predict(fit, new_data, interval = "confidence")   # mean response CI
predict(fit, new_data, interval = "prediction")   # new observation PI
```

17.7 Chapter Summary

Key formulas

Slope	$\hat{\beta}_1 = r \cdot s_y / s_x$
Intercept	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
Residual	$e_i = y_i - \hat{y}_i$
Residual variance	$\hat{\sigma}^2 = \text{SSE}/(n-2)$
SE of slope	$\hat{\sigma} / \sqrt{\sum (x_i - \bar{x})^2}$
R^2	$1 - \text{SSE}/\text{SST}$

R functions introduced

`lm(y ~ x, data)`	fit simple linear regression
`summary(fit)`	coefficient table, R^2, F-test
`confint(fit)`	95% CIs for coefficients
`predict(fit, newdata, interval)`	CI or PI at specified x
`plot(fit)`	four diagnostic plots

Key Formulas — Chapter 17

Quantity	Formula	Notes
Slope	$\hat{\beta}_1 = r \cdot s_y / s_x$	Also: $\sum(x_i - \bar{x})(y_i - \bar{y}) / \sum(x_i - \bar{x})^2$
Intercept	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$	Line passes through $(\bar{x}, \bar{y})$
Residual	$e_i = y_i - \hat{y}_i$	
Residual var	$\hat{\sigma}^2 = \text{SSE}/(n-2)$	$\text{df} = n - 2$
SE of slope	$\hat{\sigma} / \sqrt{\sum(x_i - \bar{x})^2}$	
R^2	$1 - \text{SSE}/\text{SST} = r^2$ (SLR)	

LINE assumptions: Linearity, Independence, Normality of residuals, Equal variance.

CI vs PI: CI estimates mean Y at x^*; PI is wider (includes residual variability).

EXERCISES

Conceptual.

1. State the four LINE assumptions and explain which residual plot you would use to check each.
2. Why is the fitted intercept often biologically meaningless even when the model fit is excellent?
3. Explain, in one or two sentences each, the difference between a confidence interval and a prediction interval at a fixed value of x.

Computational.

4. Given $\bar{x} = 10$, $\bar{y} = 25$, $s_x = 2$, $s_y = 5$, $r = 0.8$, compute $\hat{\beta}_0$, $\hat{\beta}_1$, and predict $\hat{y}$ at $x = 12$.
5. A regression output reports $\hat{\beta}_1 = 2.4$, $\text{SE}(\hat{\beta}_1) = 0.6$, $n = 20$. Test $H_0 : \beta_1 = 0$ at $\alpha = 0.05$ and give a 95% CI for β_1.
6. For the same regression, suppose $R^2 = 0.52$. Interpret R^2 in one sentence. What is the corresponding Pearson correlation r?

R exercises.

7. Load `finch_measurements.csv`. Fit the regression of wing length on body mass. Report $\hat{\beta}_0$, $\hat{\beta}_1$, the 95% CI for β_1, and R^2. Produce a scatterplot with the fitted line.
8. Using the same data, compute a 95% confidence interval for the mean wing length of finches with $x = 15$ g and a 95% prediction interval for a single new finch of the same mass. Explain why the prediction interval is wider.
9. Simulate a dataset of size $n = 50$ from the model $Y_i = 2 + 3x_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, 1)$ and $x_i \sim \text{Uniform}(0, 10)$. Fit a regression to recover the parameters. Repeat 1,000 times and report the empirical distribution of $\hat{\beta}_1$. Is it centred on 3?

10. (Challenge.) Show algebraically that $R^2 = r^2$ for simple linear regression. Verify the identity using a simulated dataset.

11. A biochemist fits a standard curve relating absorbance (y) to protein concentration (x, μg/mL) using $n = 8$ calibration points. `lm()` output: intercept $= 0.043$ (SE $= 0.011$), slope $= 0.0198$ (SE $= 0.0008$), $R^2 = 0.997$. (a) Interpret the slope in context. (b) Predict absorbance at 50 μg/mL. (c) Compute a 95% CI for the slope. (d) Is the intercept biologically meaningful?

12. Produce the four standard diagnostic plots (`plot(fit)`) for a regression of body weight on femur length in $n = 45$ red deer. Describe what each plot checks and what a "good" plot looks like for each LINE assumption.

13. A study reports that shoe size predicts height with $R^2 = 0.64$ ($n = 200$, $p < 0.001$). A journalist writes: "Shoe size explains 64% of height." (a) Is this statistically accurate? (b) Does it imply that shoe size *causes* height? (c) What alternative predictor might account for the correlation?

14. (Challenge) The prediction interval at x^* is wider than the confidence interval at the same x^*. (a) Write out both interval formulas explicitly. (b) Show algebraically that the PI is always wider. (c) At what value of x^* are both intervals narrowest, and why?

A

R Reference Guide

This appendix collects every R function used in the book, organized by chapter. All examples assume tidyverse is loaded (`library(tidyverse)`) and `set.seed(205)` for reproducibility.

Chapters 1–4 (Foundations)

Function	Purpose
`read_csv("file.csv")`	read a CSV file into a tibble
`glimpse(df)`	transposed data summary
`summary(df)`	marginal summaries of every column
`table(x)`	frequency table for a categorical variable
`n(),mean(),median()`	row count, arithmetic mean, median
`sd(),var()`	sample SD, sample variance
`range(),IQR(),quantile()`	range, interquartile range, percentiles
`min(),max()`	min and max
`ggplot(df, aes(...))`	begin a ggplot
`geom_histogram()`	histogram
`geom_boxplot()`	boxplot (single or grouped)
`geom_bar()`	bar chart for categorical
`geom_point()`	scatterplot
`geom_density()`	kernel density estimate
`labs(...),theme_minimal()`	axis/title labels, theme
`facet_wrap( var)`	small-multiples by a grouping variable
`replicate(n, expr)`	repeat an expression and collect results

Chapters 5 and 6 (Probability, Hypothesis Testing)

`dbinom(x, size, prob)`	pmf of the binomial distribution
`pbinom(q, size, prob)`	CDF
`qbinom(p, size, prob)`	quantile (inverse CDF)
`rbinom(n, size, prob)`	random draws
`binom.test(x, n, p, alternative)`	exact binomial test
`set.seed(205)`	fix the random-number seed

Chapter 7 (Proportions)

`prop.test(x, n, p)`	one- or two-sample z-test for proportions
`fisher.test(matrix)`	exact test for a 2×2 (or small $r \times c$) table

Chapters 8 and 9 (Chi-squared)

`chisq.test(x, p)`	goodness-of-fit chi-squared test
`chisq.test(matrix)`	chi-squared test of independence
`chisq.test(..., simulate.p.value = TRUE)`	Monte Carlo p-value

Chapter 10 (Normal Distribution)

`dnorm(x, mean, sd)`	density
`pnorm(q, mean, sd)`	cumulative probability
`qnorm(p, mean, sd)`	quantile (inverse CDF)
`rnorm(n, mean, sd)`	random draws

Chapters 11 and 12 (t-tests)

`t.test(x, mu)`	one-sample t-test with CI
`t.test(y ~ group, data)`	Welch two-sample t-test
`t.test(x, y, paired = TRUE)`	paired t-test
`stat_qq(),stat_qq_line()`	normal Q–Q plot with reference line
`effectsize::cohens_d()`	Cohen's d effect size

Chapter 13 (Violations)

`log(x),sqrt(x),1/x`	variance-stabilizing transformations
`wilcox.test(x, y)`	Mann–Whitney U (or signed-rank if paired)
`shapiro.test(x)`	Shapiro–Wilk normality test (use cautiously)

Chapter 15 (ANOVA)

`aov(y ~ group, data)`	fit a one-way ANOVA
`summary(fit)`	ANOVA table
`TukeyHSD(fit)`	simultaneous pairwise comparisons
`oneway.test(..., var.equal = FALSE)`	Welch's ANOVA
`effectsize::eta_squared(fit)`	effect size for ANOVA

Chapters 16 and 17 (Correlation, Regression)

`cor(x, y, method)`	Pearson / Spearman / Kendall correlation
`cor.test(x, y)`	inference for ρ
`lm(y ~ x, data)`	fit simple linear regression
`summary(fit)`	coefficient table, R^2, F-statistic
`confint(fit)`	95% CIs for coefficients
`predict(fit, newdata, interval)`	CI (mean) or PI (new obs) at x
`plot(fit)`	four standard regression diagnostic plots

B

Statistical Formula Sheet

A compact reference for every formula used in the book, grouped by topic.

Descriptive Statistics (Ch. 3)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{sample mean}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{sample variance}$$

$$s = \sqrt{s^2} \qquad \text{sample SD}$$

$$z_i = \frac{x_i - \bar{x}}{s} \qquad \text{sample } z\text{-score}$$

$$\text{IQR} = Q_3 - Q_1 \qquad \text{interquartile range}$$

Probability (Ch. 5)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \text{conditional}$$

$$P(A \cap B) = P(A)\,P(B) \qquad \text{if independent}$$

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)} \qquad \text{Bayes' theorem}$$

$$\mathrm{E}(X) = \sum_{x} x\, p_X(x) \qquad \text{expectation (discrete)}$$

$$\mathrm{Var}(X) = \mathrm{E}\big[(X - \mathrm{E}(X))^2\big]$$

Binomial: $X \sim \text{Binomial}(n, p)$

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{E}(X) = np, \ \text{Var}(X) = np(1-p).$$

Sampling Distributions (Ch. 4, 10)

$$\text{E}(\bar{X}) = \mu, \quad \text{SE}(\bar{X}) = \sigma/\sqrt{n} \qquad \text{population } \sigma \text{ known}$$

$$\widehat{\text{SE}}(\bar{X}) = s/\sqrt{n} \qquad \sigma \text{ estimated}$$

$$\widehat{\text{SE}}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n} \qquad \text{one proportion}$$

Central Limit Theorem: $\dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \to N(0,1)$ in distribution.

Confidence Intervals (General Template)

$$\text{estimate} \ \pm \ (\text{critical value}) \cdot (\text{standard error}).$$

Parameter	Estimator	CI
μ (known σ)	$\bar{x}$	$\bar{x} \pm z^* \cdot \sigma/\sqrt{n}$
μ (unknown σ)	$\bar{x}$	$\bar{x} \pm t^*_{n-1} \cdot s/\sqrt{n}$
p (Wald)	$\hat{p}$	$\hat{p} \pm z^* \sqrt{\hat{p}(1-\hat{p})/n}$
$\mu_1 - \mu_2$ (Welch)	$\bar{x}_1 - \bar{x}_2$	$\pm t^*_{\text{df}} \sqrt{s_1^2/n_1 + s_2^2/n_2}$
μ_D (paired)	$\bar{d}$	$\bar{d} \pm t^*_{n-1} \cdot s_D/\sqrt{n}$
β_1 (regression)	$\hat{\beta}_1$	$\hat{\beta}_1 \pm t^*_{n-2} \cdot \text{SE}(\hat{\beta}_1)$

Test Statistics (General Template)

$$\text{test statistic} \ = \ \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error}}.$$

Test	Statistic	Reference dist.
One-sample z (Ch. 7)	$(\hat{p} - p_0)/\sqrt{p_0(1-p_0)/n}$	$N(0,1)$
Two-sample z (Ch. 7)	$(\hat{p}_1 - \hat{p}_2)/\text{SE}_{\text{pooled}}$	$N(0,1)$
One-sample t (Ch. 11)	$(\bar{x} - \mu_0)/(s/\sqrt{n})$	t_{n-1}
Welch t (Ch. 12)	$(\bar{x}_1 - \bar{x}_2)/\sqrt{s_1^2/n_1 + s_2^2/n_2}$	t_{df} (Welch)
Paired t (Ch. 12)	$\bar{d}/(s_D/\sqrt{n})$	t_{n-1}
Chi-squared GoF (Ch. 8)	$\sum (O_i - E_i)^2/E_i$	χ^2_{k-1-m}
Chi-squared independence (Ch. 9)	$\sum (O_{ij} - E_{ij})^2/E_{ij}$	$\chi^2_{(r-1)(c-1)}$
ANOVA (Ch. 15)	MS_B/MS_W	$F_{k-1,\,N-k}$
Correlation (Ch. 16)	$r\sqrt{n-2}/\sqrt{1-r^2}$	t_{n-2}
Slope (Ch. 17)	$\hat{\beta}_1/\text{SE}(\hat{\beta}_1)$	t_{n-2}

ANOVA (Ch. 15)

$$
\begin{aligned}
\text{SS}_{\text{total}} &= \textstyle\sum_{i,j}(Y_{ij}-\bar{Y})^2 \\
\text{SS}_{\text{between}} &= \textstyle\sum_{i} n_i(\bar{Y}_i-\bar{Y})^2 \\
\text{SS}_{\text{within}} &= \textstyle\sum_{i,j}(Y_{ij}-\bar{Y}_i)^2 \\
\text{MS}_B &= \text{SS}_B/(k-1), \quad \text{MS}_W = \text{SS}_W/(N-k), \quad F = \text{MS}_B/\text{MS}_W \\
\eta^2 &= \text{SS}_B/\text{SS}_{\text{total}}
\end{aligned}
$$

Regression (Ch. 17)

$$
\begin{aligned}
\hat{\beta}_1 &= r\cdot s_y/s_x, \qquad \hat{\beta}_0 = \bar{y}-\hat{\beta}_1\bar{x} \\
e_i &= y_i-\hat{y}_i \\
\hat{\sigma}^2 &= \text{SSE}/(n-2) \\
\text{SE}(\hat{\beta}_1) &= \hat{\sigma}/\sqrt{\textstyle\sum(x_i-\bar{x})^2} \\
R^2 &= 1-\text{SSE}/\text{SST} = r^2 \text{ (SLR)}
\end{aligned}
$$

Effect Sizes

$$
\begin{aligned}
d_{\text{one-sample}} &= (\bar{x}-\mu_0)/s \\
d_{\text{two-sample}} &= (\bar{x}_1-\bar{x}_2)/s_p, \quad s_p = \sqrt{\tfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \\
d_{\text{paired}} &= \bar{d}/s_D \\
\eta^2 &= \text{SS}_{\text{between}}/\text{SS}_{\text{total}} \\
r, R^2 &\text{ in regression / correlation}
\end{aligned}
$$

Conventional thresholds (Cohen): $|d| \approx 0.2$ small, 0.5 medium, 0.8 large; $\eta^2 \approx 0.01/0.06/0.14$.

C

Statistical Tables

All tables in this appendix were generated in R using `pnorm`, `qt`, `qchisq`, and `qf`; the source script is provided with the book's instructor package. In practice you should always use software for exact values—but printed tables remain useful for quick reference, in-class illustration, and exam settings.

Standard Normal Distribution (Z)

Entries give $P(Z \le z)$ for $z \sim N(0,1)$. Row label is z to one decimal, column label is the second decimal.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005
-3.3	0.0005	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007
-3.2	0.0007	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009
-3.1	0.0010	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013
-3.0	0.0013	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018
-2.9	0.0019	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025
-2.8	0.0026	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034
-2.7	0.0035	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045
-2.6	0.0047	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060
-2.5	0.0062	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080
-2.4	0.0082	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104
-2.3	0.0107	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136
-2.2	0.0139	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174
-2.1	0.0179	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222
-2.0	0.0228	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281
-1.9	0.0287	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351
-1.8	0.0359	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436
-1.7	0.0446	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537
-1.6	0.0548	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655
-1.5	0.0668	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793
-1.4	0.0808	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951
-1.3	0.0968	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131
-1.2	0.1151	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335
-1.1	0.1357	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562
-1.0	0.1587	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814
-0.9	0.1841	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090
-0.8	0.2119	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389
-0.7	0.2420	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709
-0.6	0.2743	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050
-0.5	0.3085	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409
-0.4	0.3446	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783
-0.3	0.3821	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168
-0.2	0.4207	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562
-0.1	0.4602	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Student's t Distribution

Entries $t^*_{\mathrm{df},\alpha}$ satisfy $P(T_{\mathrm{df}} \geq t^*) = \alpha$.

df	$\alpha = 0.10$	0.05	0.025	0.01	0.005	0.001
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
60	1.296	1.671	2.000	2.390	2.660	3.232
120	1.289	1.658	1.980	2.358	2.617	3.160
∞	1.282	1.645	1.960	2.326	2.576	3.090

Chi-Squared Distribution

Entries $\chi^2_{\mathrm{df},\alpha}$ satisfy $P(\chi^2_{\mathrm{df}} \geq \chi^2_*) = \alpha$.

df	$\alpha = 0.995$	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	79.082	83.298	88.379	91.952
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169

F Distribution, $\alpha = 0.05$ (upper tail)

Entries $F^*_{\mathrm{df}_1,\mathrm{df}_2,0.05}$ satisfy $P(F_{\mathrm{df}_1,\mathrm{df}_2} \geq F^*) = 0.05$.

df_2	1	2	3	4	5	6	8	10	12	15	20	24	30	60
	Upper-tail $\alpha = 0.050$													
1	161.45	199.50	215.71	224.58	230.16	233.99	238.88	241.88	243.91	245.95	248.01	249.05	250.10	252.20
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.40	19.41	19.43	19.45	19.45	19.46	19.48
3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.79	8.74	8.70	8.66	8.64	8.62	8.57
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.96	5.91	5.86	5.80	5.77	5.75	5.69
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.74	4.68	4.62	4.56	4.53	4.50	4.43
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.06	4.00	3.94	3.87	3.84	3.81	3.74
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.64	3.57	3.51	3.44	3.41	3.38	3.30
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.35	3.28	3.22	3.15	3.12	3.08	3.01
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.14	3.07	3.01	2.94	2.90	2.86	2.79
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.98	2.91	2.85	2.77	2.74	2.70	2.62
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.85	2.79	2.72	2.65	2.61	2.57	2.49
12	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.75	2.69	2.62	2.54	2.51	2.47	2.38
13	4.67	3.81	3.41	3.18	3.03	2.92	2.77	2.67	2.60	2.53	2.46	2.42	2.38	2.30
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.60	2.53	2.46	2.39	2.35	2.31	2.22
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.54	2.48	2.40	2.33	2.29	2.25	2.16
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.49	2.42	2.35	2.28	2.24	2.19	2.11
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.45	2.38	2.31	2.23	2.19	2.15	2.06
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.41	2.34	2.27	2.19	2.15	2.11	2.02
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.38	2.31	2.23	2.16	2.11	2.07	1.98
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.35	2.28	2.20	2.12	2.08	2.04	1.95
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.30	2.23	2.15	2.07	2.03	1.98	1.89
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.25	2.18	2.11	2.03	1.98	1.94	1.84
26	4.23	3.37	2.98	2.74	2.59	2.47	2.32	2.22	2.15	2.07	1.99	1.95	1.90	1.80
28	4.20	3.34	2.95	2.71	2.56	2.45	2.29	2.19	2.12	2.04	1.96	1.91	1.87	1.77
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.16	2.09	2.01	1.93	1.89	1.84	1.74
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.08	2.00	1.92	1.84	1.79	1.74	1.64
60	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.99	1.92	1.84	1.75	1.70	1.65	1.53
120	3.92	3.07	2.68	2.45	2.29	2.18	2.02	1.91	1.83	1.75	1.66	1.61	1.55	1.43
1000	3.85	3.00	2.61	2.38	2.22	2.11	1.95	1.84	1.76	1.68	1.58	1.53	1.47	1.33

F Distribution, $\alpha = 0.01$ (upper tail)

df_2	1	2	3	4	5	6	8	10	12	15	20	24	30	60
	Upper-tail $\alpha = 0.010$													
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5981.07	6055.85	6106.32	6157.28	6208.73	6234.63	6260.65	6313.03
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.40	99.42	99.43	99.45	99.46	99.47	99.48
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.23	27.05	26.87	26.69	26.60	26.50	26.32
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.55	14.37	14.20	14.02	13.93	13.84	13.65
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	10.05	9.89	9.72	9.55	9.47	9.38	9.20
6	13.75	10.92	9.78	9.15	8.75	8.47	8.10	7.87	7.72	7.56	7.40	7.31	7.23	7.06
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.62	6.47	6.31	6.16	6.07	5.99	5.82
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.81	5.67	5.52	5.36	5.28	5.20	5.03
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.26	5.11	4.96	4.81	4.73	4.65	4.48
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.85	4.71	4.56	4.41	4.33	4.25	4.08
11	9.65	7.21	6.22	5.67	5.32	5.07	4.74	4.54	4.40	4.25	4.10	4.02	3.94	3.78
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.30	4.16	4.01	3.86	3.78	3.70	3.54
13	9.07	6.70	5.74	5.21	4.86	4.62	4.30	4.10	3.96	3.82	3.66	3.59	3.51	3.34
14	8.86	6.51	5.56	5.04	4.69	4.46	4.14	3.94	3.80	3.66	3.51	3.43	3.35	3.18
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.80	3.67	3.52	3.37	3.29	3.21	3.05
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.69	3.55	3.41	3.26	3.18	3.10	2.93
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.59	3.46	3.31	3.16	3.08	3.00	2.83
18	8.29	6.01	5.09	4.58	4.25	4.01	3.71	3.51	3.37	3.23	3.08	3.00	2.92	2.75
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.43	3.30	3.15	3.00	2.92	2.84	2.67
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.37	3.23	3.09	2.94	2.86	2.78	2.61
22	7.95	5.72	4.82	4.31	3.99	3.76	3.45	3.26	3.12	2.98	2.83	2.75	2.67	2.50
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.17	3.03	2.89	2.74	2.66	2.58	2.40
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	3.09	2.96	2.81	2.66	2.58	2.50	2.33
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	3.03	2.90	2.75	2.60	2.52	2.44	2.26
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.98	2.84	2.70	2.55	2.47	2.39	2.21
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.80	2.66	2.52	2.37	2.29	2.20	2.02
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.63	2.50	2.35	2.20	2.12	2.03	1.84
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.47	2.34	2.19	2.03	1.95	1.86	1.66
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.53	2.34	2.20	2.06	1.90	1.81	1.72	1.50

D

Datasets Used in This Book

This appendix catalogues every dataset referenced in an exercise or R example in the book. All datasets are distributed with the course's electronic resources. Paths below are relative to the course's shared data folder.

`lmu_class_data_cleaned.csv`

Source. Anonymized responses collected from MATH 205 students, cleaned by the instructor. Used as a running demonstration dataset throughout Chapters 1–3 and Ch. 12.

Variables.

- `height_cm` – height in centimetres (numeric)
- `weight_kg` – body mass in kilograms (numeric)
- `sex` – self-reported sex (categorical)
- `major` – declared academic major (categorical)
- `study_hours` – self-reported weekly study hours (numeric)
- additional demographic and academic variables (see the original `R_Quick_Reference_Guide.pdf` in the data folder)

`Dataset1_Finch_Beaks.csv`

Source. Final-project dataset based on Grant & Grant's long-running study of Galapagos finches. Used in Chapters 11, 12, 16, and 17.

Variables. `species`, `mass_g`, `beak_length_mm`, `beak_depth_mm`, `wing_length_mm`, `year`.

`Dataset2_Mammal_Brain_Body.csv`

Source. Comparative-mammalogy dataset of brain vs. body mass across mammalian species. Used in regression exercises (Ch. 17) to illustrate log transforms.

Variables. `species`, `body_mass_kg`, `brain_mass_g`, `order`.

Dataset3_Plant_CO2.csv

Source. Plant-physiology dataset measuring CO_2 uptake under different temperature and chilling treatments. Referenced in Chapters 11, 15, and the final project.

Variables. `Plant`, `Type` (Quebec / Mississippi), `Treatment` (chilled / nonchilled), `conc` (CO_2 concentration), `uptake` (μmol/m^2/s).

Dataset4_Global_Health.csv

Source. WHO-derived public-health indicators at the country level. Used in the final project.

Variables. `country`, `region`, `gdp_per_capita`, `life_expectancy`, `infant_mortality`, `vaccination_rate`.

Dataset5_Antibiotic_Resistance.csv

Source. Hospital-ward antibiotic-resistance isolates. Referenced in Chapter 9 (contingency analysis) and the final project.

Variables. `isolate_id`, `ward` (ICU / surgical / medical), `resistance` (resistant / susceptible), `pathogen`, `year`.

Example datasets referenced in text

Several smaller datasets appear in worked examples and exercises within the chapters. They are available as CSV alongside this book:

- `juno_hatch.csv` – green sea turtle hatchling masses (Ch. 11 worked example, 18 observations).
- `lizard_speed.csv` – side-blotched lizard sprint speeds at two elevations (Ch. 12 worked example).
- `bp_pre_post.csv` – pre/post systolic blood pressures on 12 patients (Ch. 12 paired example).
- `fish_biomass.csv` – sampling-site biomass (Ch. 13, log-transform example).
- `reaction_time.csv` – reaction times by noise condition (Ch. 13, nonparametric example).
- `seedling_growth.csv` – stem lengths under three light regimes (Ch. 15 ANOVA example).
- `finch_measurements.csv` – mass and wing length (Ch. 16 correlation, Ch. 17 regression).

All datasets are distributed in CSV format with a header row, no row names, and UTF-8 encoding. `read_csv()` from readr (part of tidyverse) handles them without further options.