MA22004 / MA52008 — Statistics and Data Analysis

Dr Eric Hall • ehall001@dundee.ac.uk

2026-01-15

University of Dundee

Table of contents

Welcome

Welcome to MA22004/MA52008 Statistics and Data Analysis at the University of Dundee.

This module covers the basics of statistical inference including point estimation, interval estimation, hypothesis testing, linear regression, and simple goodness-of-fit tests. The appendix contains a list of curated content for you to investigate.

These notes are available at dundeemath.github.io/MA22004/ and also as a PDF (visit the page and click on the PDF icon to download).

Licence

Part I

Module Introduction

Syllabus

MA22004/MA52008 Syllabus (AY2025-2026)

Module Information

- Duration: 11 weeks (10 lecutre weeks)
- Contact Hours: 2 lecture hours + 2 tutorial hours per week + 5 hours self-paced labs
- Software: R (including RMarkdown/Quarto for reproducibile report writing)

Aims

This module aims to provide students with the skills and knowledge required to analyse data and make statistical inferences using sample data. Students will develop a solid foundation in:

- Exploratory data analysis
- Sampling distributions
- Hypothesis testing
- Analysis of variance
- Regression modeling

Throughout the module, students will use R for data visualization, statistical modeling, hypothesis testing, and creating reproducible reports.

Indicative Content

Data Analysis and Exploration

- Exploratory data analysis (EDA) and visualization using R
- Data characteristics, summary statistics, and descriptive measures
- Correlation measures: Pearson, Spearman, Kendall
- Principal Component Analysis (PCA) for dimensionality reduction

- Reproducible research: RMarkdown/Quarto

Sampling Distributions and Estimation

- Sample statistics and their distributions
- Properties of sampling distributions (mean, variance, shape)
- Applications of the Central Limit Theorem (CLT)
- Random sampling methods and bias considerations

Inferences

- Point estimators and their properties
- Confidence intervals for:
  - Means and variances (Normal and t)
  - Proportions and rates (Binomial and Poisson)
  - Variances (Chi-square and F)
- One and two-sample inference (including paired data and bootstrapping)
- Hypothesis testing framework:
  - Null/alternative hypotheses, significance levels
  - Type I & II errors, sensitivity, specificity, likelihood ratio, power of a test
- Goodness-of-fit tests: One-way and two-way chi-square tests
- One-way analysis of variance (ANOVA)

Linear Regression Models

- Response and explanatory variables
- Simple and multiple linear regression models
- Least squares estimation and interpretation of coefficients
- Assessing model adequacy: Residual analysis and diagnostic tests
- Use of R to fit a regression model and interpret output

Computational Tools and Reproducibility

- Use of R for:
  - Bootstrapping and permutation tests
  - t-tests, ANOVA, goodness-of-fit tests
  - Linear regression analysis
- Reproducible reporting using RMarkdown/Quarto

Intended Learning Outcomes

By the end of the module, students will be able to:

- Explain how properties of populations relate to sample data and describe appropriate sampling techniques.
- Use R to perform exploratory data analysis, generate summary statistics, and visualize data effectively.
- Interpret correlation coefficients and apply Principal Component Analysis (PCA) for high-dimensional data analysis.
- Construct and interpret confidence intervals for means, variances, proportions, and rates in one and two-sample situations.
- Perform hypothesis tests, including goodness-of-fit tests, ANOVA, and regression model assessment.
- Develop and assess simple and multiple linear regression models, including diagnostic testing.
- Evaluate the reliability of statistical models and interpret the power of a test in an inference setting.
- Engage in mathematical and statistical dialogue.
- Use R to conduct analyses, test statistical models, and create reproducible research reports.

Lecture Plan

Week	Topic	Demo
1	Orientation, EDA, Correlation, PCA, Reproducibility	Icebreaker, R PCA
2	Probability Distributions, Sampling Distributions, CLT	Candy sampling

Week	Topic	Demo
3	Estimation, CI, Hypothesis Testing, Errors & Power	Why 0.05?
4	One-Sample Inference (CI & Tests for Means, Proportions, Variances)	How much water?
5	Two-Sample Inference (Unpaired Cases)	UN & Africa
6	Two-Sample Inference (Paired vs. Unpaired, Variance Testing)	Tennis ball challenge
7	One-Way ANOVA, Goodness-of-Fit Tests	R `aov`
8	Simple & Multiple Regression, LINE Assumptions	R `lm`
9	Regression Inference, Model Diagnostics, Chi-square	R `lm`, Chi-square
10	Quality Control, Bias Correction, Final Review	R 3-sigma control charts

Assessment & Coursework

See MyDundee material.

Software and Resources

- R and RStudio (including RMarkdown/Quarto)
- Lecture notes: Statistics and Data Analysis Lecture Notes
- Lab materials: Statistics and Data Analysis Labs
- Recommended Texts: see MyDundee Library resources.

Appendix: Mapping to CS1 Actuarial Statistics Syllabus

This appendix provides a structured mapping of which CS1 topics are addressed in which weeks of the module.

CS1 Section	Topics Covered	Week(s)
1.1.1	Aims of data analysis	1
1.1.2	Stages of data analysis and suitable tools	1
1.1.3	Sources of data and their characteristics	1
1.1.4	Reproducible research methods	1, 10
1.2.1	Summary statistics and exploratory visualizations	1

CS1 Section	Topics Covered	Week(s)
1.2.2	Correlation measures, including Pearson's, Spearman's, and Kendall's coefficients	1
1.2.3	Principal Component Analysis (PCA) for dimensionality reduction	1
2.6.1	Sample statistics and their distributions	2
2.6.2	Properties of sampling distributions (mean, variance, shape)	2
2.6.3	Applications of the Central Limit Theorem	2
2.6.4	Random sampling methods and bias considerations	2
3.2.1	Point estimators and their properties	3
3.2.2	Confidence intervals for means and variances (Normal and t), proportions and rates (Binomial and Poisson), variances (Chi-square and F)	3, 4, 5, 6
3.2.3	Confidence intervals based on one and two-sample situations, including paired data and bootstrapping	3, 4, 5, 6
3.3.1	Hypothesis testing framework, including null/alternative hypotheses, significance levels, errors, and power of a test	3
3.3.2	Hypothesis tests for one and two-sample situations	4, 5, 6
3.3.3	Goodness-of-fit tests, including one-way and two-way chi-square tests	7, 9
4.1.1	Simple and multiple linear regression models	8
4.1.2	Assessing model adequacy, including residual analysis and diagnostic tests	9
4.1.3	Use of R to fit a linear regression model and interpret output	8, 9

Lab Guide

You will learn about the statistical programming language `R` and the software RStudio by working through seven interactive lab tutorials and completing lab reports. The lab reports should answer the exercise questions at the end of each tutorial.

Tutorials and all associated materials (templates, data sets, further instructions, etc.) are available as an `R` package at the GitHub repository `dundeemath/MA22004labs` (i.e., https://github.com/dundeemath/MA22004labs).

Instructions on how to install and access the interactive lab tutorials can be found at:

- https://dundeemath.github.io/MA22004labs/.

The following section contains details about writing lab reports.

Writing Lab Reports

Assessment Criteria

There are seven interactive lab tutorials with accompanying exercises. Each lab tutorial specifies how marks are allocated across the exercises (a maximum of 20 marks available for each lab report).

> **!** Important
>
> Marks are awarded for both **content** and **presentation**.

Content

Please work through the interactive tutorial for each lab. Your lab report should answer the exercises found at the end of each tutorial.

Presentation

Please use Quarto to create your lab report. Further instructions on using Quarto for creating *reproducible* lab reports that combine data analysis and text can be found in Lab 1. This set of lecture notes was authored using Quarto; you can see the source code in the GitHub repository https://github.com/dundeemath/MA22004.

Plots

Plots should be neat and legible, with appropriate aesthetic elements. Please use `ggplot` for creating plots and visualisations. Each plot should be annotated with titles, axis labels, and legends as appropriate. Plot aesthetics should be distinguished, e.g. using colours or line styles that are identified using a legend. Important data points and coordinates should be annotated using labels.

Mathematical formulas

Mathematical formulas should follow the same style rules as the lecture notes. Formulas can be included in Quarto documents using LaTeX syntax. There should be appropriate spacing around operators and equals signs, e.g. $a + b = c$. For punctuation, formulas are treated as part of the text, so they often need to end with a full stop or comma. Important formulas can appear "displayed" on their own line (with line spacing above and below them), e.g.,

$$A = \pi r^2 \, .$$

Structure

Structure should be logical and clear. Organise your writing with suitable headings and sub-headings. For example, provide a solution to each exercise under its own heading.

Writing

Writing should follow the usual rules of good written English, including writing complete sentences and paragraphs that get to the point quickly. Your tone and language should be similar to lecture notes or scientific journal articles. Formal writing does not require unnecessary words, long words or monotonous use of passive voice. I will reward concise and clear communication, so please do not write, "Upon carefully analysing the aforementioned equations, the following mathematical solution was found," when "The solution is" conveys the same thing.

Formatting

Formatting should rely on the *MA22004/MA52008 Lab Report* template. This is available in the `MA22004labs` package, and further instructions can be found in Lab 1.

Part II

Lecture Notes

Preliminaries

This section contains a list of abbreviations, comment on notation, and a (very quick) review of probability.

Abbreviations

In Table 3 we list abbreviations used throughout these lecture notes. These abbreviations are pretty standard and you might encounter them outside the module in other references.

Table 3: Commonly used abbreviations.

Abbreviation	Expanded
pdf	probability density function
cdf	cumulative distribution function
rv	random variable
iid	independent and identically distributed
obs	observations
CI	confidence interval
df	degrees of freedom

Notation

Uppercase roman letters, e.g., X, will typically denote random variables (rvs); lower case letters, e.g., x, will represent a particular value (observation) of a rv. Rvs have probability distributions. Distributions are typically characterised by *parameters* that describe population characteristics. In the present module, we will adopt the (frequentists) view that parameters are fixed real numbers that are often unknown and must be estimated from data. Statistical inference is a tool that will help us to do this.

> ⚠ Variables and parameters
>
> Statistical models comprise both rvs and parameters. Be careful not to confuse them!

For a random variable X that has a distribution F depending on a set of parameters Θ, we will write $X \sim F(\Theta)$.

> ⚠ Specifying a probability distribution
>
> We write $X \sim F(\Theta)$ to indicate X has distribution function $F(\Theta)$. This is **not** read as "X is approximately $F(\Theta)$"!

Sample space, events, probabilities

A sample space Ω is a set of possible outcomes of an experiment. Points $\omega \in \Omega$ are sample outcomes or realizations. Subsets $A \subset \Omega$ are called events.

Example 0.1 (Sample space)**.** Consider an experiment where we measure the petal widths from a randomly sampled cyclamen flowers. Before we observe the petal width, there is uncertainty that we can model using a sample space of events. The sample space is $\Omega = (0, \infty)$, since measurements of length should be positive (practically, the lengths will have a finite size, too). Each $\omega \in \Omega$ is a measurement of petal width for a cyclamen flower. Consider an event $A = (5, 12]$; this is the event that the petal width is larger than 5 but less than or equal to 12. Remember, we use probability to model uncertainty *before* we observe the petal width — after we take a measurement, the petal width is no longer uncertain (we have collected a statistic).

As sample spaces and events are described using sets, we recall the following notations, definitions, and laws about set theory. Let A, B, and $A_1, A_2, \ldots$ be events in a sample space Ω.

- complement: $A^c = \{\omega \in \Omega : \omega \notin A\}$.
- null event: $\varnothing = \Omega^c$.
- intersection: $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$. In particular, for $A_1, A_2, \ldots$, then

$$\bigcap_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for all } i\}.$$

- difference: $A \setminus B = \{\omega \in \Omega : \omega \in A, \omega \notin B\}$.
- size: $|A|$ denotes the number of elements in A.
- disjoint: $A_i \cap A_j = \varnothing$, for $i \neq j$.
- partition: disjoint $A_1, A_2, \ldots$ such that $\bigcup_{i=1}^{\infty} A_i = \Omega$.
- indicator: $I_A(\omega) = I(\omega \in A) = \{1 \text{ if } \omega \in A; 0 \text{ if } \omega \notin A\}$.
- monotone increasing: $A_1 \subset A_2 \subset \ldots$ and define limit

$$\lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i.$$

- monotone decreasing: $A_1 \supset A_2 \supset \ldots$ and define limit

$$\lim_{n\to\infty} A_n = \bigcap_{i=1}^{\infty} A_i.$$

- distributive laws:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

- De Morgan's laws:

$$(A \cap B)^c = A^c \cup B^c\,,$$
$$(A \cup B)^c = A^c \cap B^c\,.$$

We assign probabilities to events in our sample space.

Definition 0.1 (Probability distribution)**.** A probability distribution is a function $P : \Omega \to \mathbf{R}$ satisfying three axioms:

1. $P(A) \geq 0$ for every $A \subset \Omega$ (positivity),
2. $P(\Omega) = 1$ (totality),
3. if $A_1, A_2, \ldots$ are disjoint subsets of Ω, then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)\,.$$

Perspectives

We can interpret $P(A)$ as representing:

- **frequency**, i.e., the long-run proportion of times A is true (the *frequentist perspective*),
- **degrees of belief**, i.e., as a measure of the observer's strength of belief that A is true (the *Bayesian perspective*).

Theorem 0.1 (PIE)**.** *The principle of inclusion-exclusion,*

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

Theorem 0.1 follows from the definition of a probability distribution and facts about set theory.

Definition 0.2 (Probability of an event)**.** For events A from finite sample spaces Ω, we assign probabilities according to:

$$P(A) = \frac{|A|}{|\Omega|}\,.$$

For finite sample spaces, we assign probabilities according to their long-run frequency of occurring. For an event A, this is the ratio of the size of A (number of ways A can happen) to the size of Ω (number of total outcomes).

Definition 0.3 (Independent events)**.** Events A and B are independent, i.e., $A \perp\!\!\!\perp B$, iff $P(A \cap B) = P(A)P(B)$.

That is, events A and B are independent if and only if the probability of A and B occurring is equal to the probability A occurring times the probability of B occurring.

Definition 0.4 (Conditional probability)**.** If $P(B) > 0$, then

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

Note that:

- $P(\cdot \mid B)$ satisfies the axioms of probability, for fixed B,
- in general, $P(A \mid \cdot)$ is not a probability for fixed A, and,
- in general, $P(A \mid B) \neq P(B \mid A)$.

Theorem 0.2 (Bayes Theorem)**.** *Let events $A_1, \ldots, A_k$ partition Ω, with $P(A_i) > 0$.*

If $P(B) > 0$, then

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_j P(B \mid A_j)P(A_j)}.$$

Generally, it is not feasible to assign probabilities to *all* subsets of Ω (e.g., if $\square$is infinite). In that case, we restrict to our attention to a σ-algebra $\mathscr{A}$ (also called, σ-field), which is a collection of sets satisfying:

1. $\varnothing \in \mathscr{A}$,
2. if $A_1, A_2, \ldots, \in \mathscr{A}$ then $\cup_{i=1}^{\infty} A_i \in \mathscr{A}$, $3. A \in \mathscr{A} \implies A^c \in \mathscr{A}$.

Sets in $\mathscr{A}$ are said to be *measurable* and $(\Omega, \mathscr{A})$ is a measure space. If P is a probability defined on $\mathscr{A}$, then $(\Omega, \mathscr{A}, P)$ is called a *probability space*.

E.g., when $\Omega \equiv \mathbf{R}$, we take $\mathscr{A}$ to be the smallest σ-field containing all open subsets of $\mathbf{R}$, which is called the Borel σ-field. If you find these details interesting, take: MA42008 Mathematical Statistics!

Random variables

How do we link sample spaces and events to data?

We use random variables to link sample spaces and events to data.

Definition 0.5 (Random variables)**.** A random variable (rv) is a mapping $X : \Omega \to \mathbf{R}$ that maps $\omega \in \Omega \mapsto X(\omega)$.

Example 0.2. Consider a coin flipping experiment where you flip a fair coin eight times. Let X be the number of heads in the sequence. If three heads occur, e.g., $\omega = HTTTTTHH$, then $X(\omega) = 3$.

Example 0.3. Consider an experiment where you draw a point a random from the unit disk. Then $\Omega = \{(x, y) : x^2 + y^2 \leq 1\}$ and a typical outcome will be the pair $\omega = (x, y)$. Some random variables to consider are $X(\omega) = x$, $Y(\omega) = y$, $Z(\omega) = x + y$, and $W(\omega) = \sqrt{x^2 + y^2}$.

Definition 0.6 (Assigning probabilities to rvs)**.** Given X and $A \subset \mathbf{R}$, we define

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$$

and let

$$P(X \in A) = P(X^{-1}(A)) = P(\{\omega \in \Omega : X(\omega) \in A\}),$$

e.g., $P(X = x) = P(X^{-1}(x)) = P(\{\omega \in \Omega : X(\omega) = x\})$.

⚠ Observations vs rvs

X denotes a rv and x denotes a particular value of X.

❗ We measure probabilities of events

A rv X by itself is not an event. You would never write $P(X)$, would you!?

Example 0.4. Consider a coin flipping experiment where you flip a fair coin twice. Let X be the number of heads. Then

$$P(X = 0) = P(\{TT\}) = \frac{1}{4},$$

$$P(X = 1) = P(\{HT\} \cup \{TH\}) = P(\{HT\}) + P(\{TH\}) = \frac{1}{2},$$

$$P(X = 2) = P(\{HH\}) = \frac{1}{4}.$$

Definition 0.7 (Cdf)**.** The cumulative distribution function (cdf), $F_X : \mathbf{R} \to [0, 1]$, is defined by $F_X(x) = P(X \leq x)$.

Figure 1 displays the cdf for the coin flip experiment considered in Example 0.4. The cdf $F_X(x)$ jumps at $x = 0$, $x = 1$, and $x = 2$. The height of the jumps are given by $P(X = x)$. We observe as well that $F_X(x) = 0$ for $x < 0$, as no probability has been accumulated; recall that probabilities are always non-negative, so a function that accumulates probability will always be non-negative. Further, $F_X(x) = 1$ for $x \geq 2$, as all the probability has been accumulated; remember that the total probability that can be assigned over the whole sample space must sum to one.

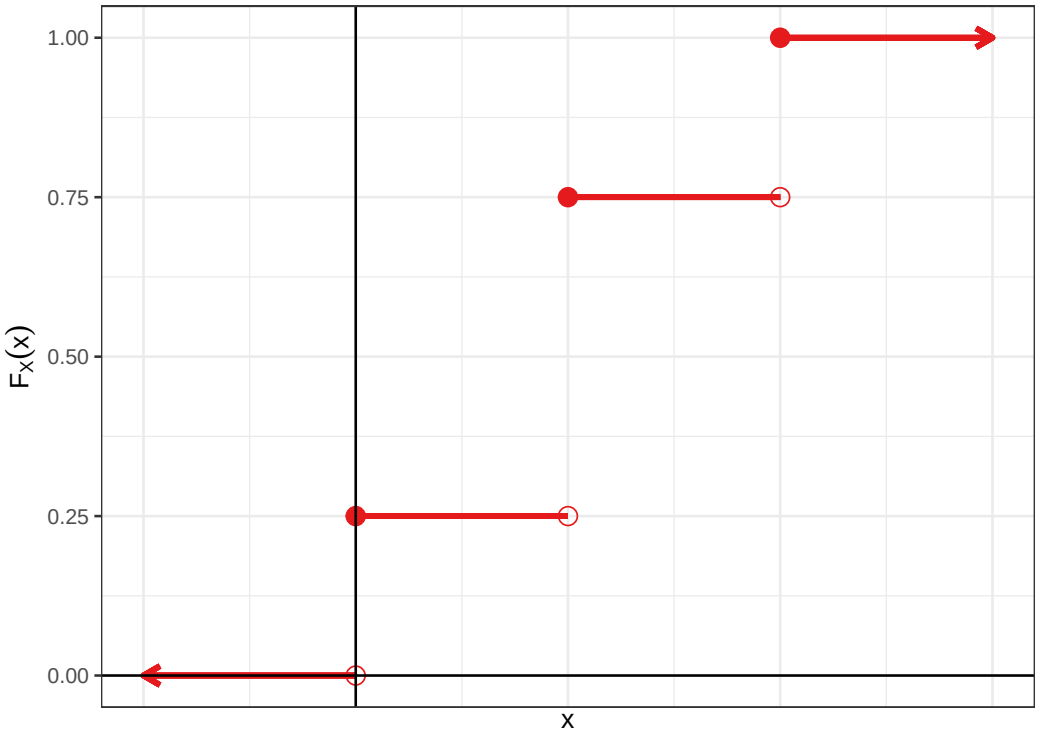


Figure 1: The cdf for the two coin flip example.

Note that a cdf completely determines the distribution of a random variable. This statement is captured in Theorem 0.3.

Theorem 0.3. *Let X have cdf F and Y have cdf G. If $F(x) = G(x)$ for all x, then $P(X \in A) = P(Y \in A) \forall A \in \mathbf{R}$.*

Since cdfs determine or characterize a probability distribution, it is useful to know the key properties of cdfs, which are listed below in Theorem 0.4.

Theorem 0.4 (Properties of cdfs)**.** *$F : \mathbf{R} \to [0, 1]$ is a cdf for some P iff,*

1. *F is nondecreasing (i.e., $x_1 < x_2 \implies F(x_1) \leq F(x_2)$),*
2. *F is normalized to $[0, 1]$ (i.e., $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$),*
3. *F is right-continuous (i.e., $F(x) = F(x^*) \forall x$ where $F(x^*) = \lim_{y > x; y \to x} F(y)$).*

For a rv X we say X is *discrete* if it assumes at most a *countable* number of (discrete) values. For a discrete sample space, the collection of all probabilities of $X(\omega)$ gives us a probability distribution.

Definition 0.8 (Pmf)**.** A pdf for a discrete rv X is $f_X(x) = P(X = x)$. Since this density function places a "point mass" at each x, it is sometimes referred to as a probability mass function (pmf).

Figure 2 displays the pmf for the coin flip experiment considered in Example 0.4. The pmf is a histogram with point masses at $x = 0$, $x = 1$, and $x = 2$. The mass placed at these points is given by $P(X = x)$. Since the pmf is a pdf for a discrete random variable, recall from the axioms of probability that the pmf therefore satisfies $f(x) \geq 0$, $\forall x \in \mathbf{R}$, and $\sum_i f(x_i) = 1$. This fact can be observed in Figure 2: $f_X(0) + f_X(1) + f_X(2) = 0.25 + 0.5 + 0.25 = 1$.

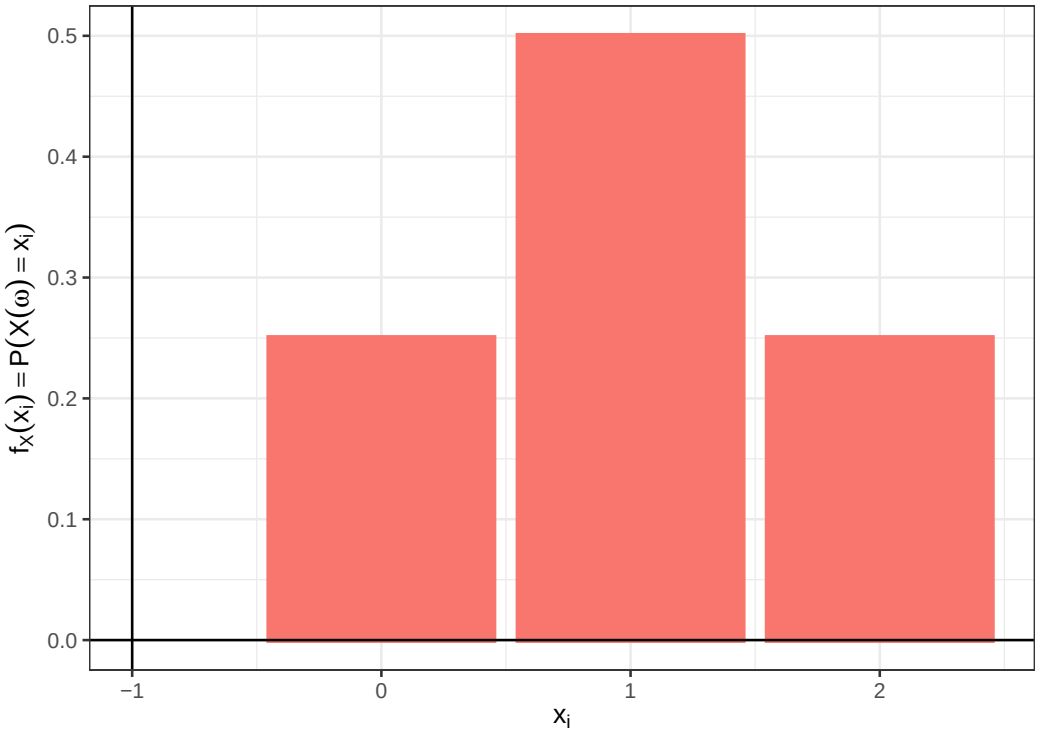


Figure 2: The histogram (pmf) for the two coin flip example.

A rv X is *continuous* if there exists a continuous function f_X such that,

1. $f_X(x) \geq 0 \forall x$,
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$ and
3. $P(a < X < b) = \int_a^b f_X(x)dx$, for $a \leq b$.

Definition 0.9 (Pdf)**.** A f_X satisfying the three properties above is a pdf for the continous rv X.

> **!** Events of probability zero
>
> If X is continuous, then $P(X = x) = 0$ for every x. That is,
>
> $$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b),.$$

The cdf is related to the pdf by the derivative (difference). If X is continuous:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

and $f_X(x) = F'_X(x)$ at all x at which F_X is differentiable. (Likewise, if X is discrete, then we replace the integral with a sum $F_X(x) = P(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$.)

Definition 0.10 (Quantile function)**.** Let X be a rv with cdf F. The inverse cdf, or quantile function, is defined by

$$F^{-1}(q) = \inf\{x : F(x) > q\}$$

for $q \in [0, 1]$. If F is monotonic increasing and continuous then $F^{-1}(q)$ is the unique real number x such that $F(x) = q$.

Some quantiles get used more than others (and therefore get names). Important quantiles include, $F^{-1}(\frac{1}{4})$ is the first quantile, $F^{-1}(\frac{1}{2})$ is the median, and $F^{-1}(\frac{3}{4})$ is the third quantile.

Definition 0.11 (Equality in distribution)**.** We say X and Y are equal in distribution, $X \equiv Y$, if $F_X(x) = F_Y(x)$ for all x.

> **!** Equality in distribution versus equality of rvs
>
> Note that equality in distribution does not mean that the random variables are the same. Rather, probability statements are the same.
> Consider the following example. Suppose
>
> $$P(X = 1) = P(X = -1) = \frac{1}{2}.$$
>
> Let $Y = -X$. Then
>
> $$P(Y = 1) = P(Y = -1) = \frac{1}{2}.$$
>
> Thus,
>
> $$X \equiv Y,$$
>
> but X and Y are not equal! In fact, $P(X = Y) = 0$.

We sometimes consider more than one random variable, taken to together. This leads to the concept of a joint and marginal densities.

Definition 0.12 (Joint pdf)**.** A joint pdf for (X, Y) satisfies

1. $f(x, y) \geq 0 \; \forall x, y$,
2. $\iint_{-\infty}^{\infty} f(x, y) dx dy = 1$,
3. for $A \in \mathbf{R} \times \mathbf{R}$, $P((X, Y) \in A) = \iint_A f(x, y) dx dy$.

Definition 0.13 (Joint cdf)**.** A joint cdf is given by $F(x, y) = P(X \leq x, Y \leq y)$.

Definition 0.14 (marginal pdf)**.** For X, Y with joint pdf $f(x, y)$, we define the marginals for X and Y as $f_X(x) \int f(x, y) dy$ and $f_Y(y) = \int f(x, y) dx$, respectively.

We also have a notion of independence for two rvs.

Definition 0.15 (Independence of rvs)**.** Rvs X and Y are independent if $P(X \in A, Y \in B) = P(X \in A) P(Y \in B)$.

Theorem 0.5. *Let X, Y have joint f_{XY}. Then X and Y are independent iff $f_{XY} = f_X \cdot f_Y$ for all x, y.*

If $X_1, \ldots X_n$ are independent and each as the same marginal distribution with cdf F, we say $X_1, \ldots, X_n$ are iid and write $X_1, \ldots, X_n \sim F$ iid. We also write $X_1, \ldots, X_n \sim f$ if F has corresponding density f, when no confusion arises. We will often consider collections of iid random variables.

Definition 0.16 (Random sample)**.** $X_1, \ldots, X_n \sim F$ iid is a random sample of size n from a distribution F.

We also consider the expected value of a rv.

Definition 0.17 (Expectation)**.** For a discrete rv X with possible outcomes $x_1, x_2, \ldots$ and corresponding probabilities $p_1, p_2, \ldots$, the expectation is defined by

$$\mathbf{E}[X] = \sum_{i=1}^{\infty} x_i p_i \, .$$

For a continuous rv X with pdf f, the expectation is defined by

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f(x) dx \, .$$

For both discrete and continuous rvs, we refer to various statistics relating to expected values as moments of the distribution.

Definition 0.18 (n-th raw moment)**.** For a rv X, the n-th raw moment is given by $\mathbf{E}[X^n]$.

Definition 0.19 (n-th central moment)**.** For a rv X with $\mu = \mathbf{E}[X]$, the n-th central moment is defined as $\mathbf{E}[(X - \mu)^n]$.

Table 4: First few moments for a rv X with mean $\mu = \mathbf{E}[X]$.

(a) Quantities related to raw moments

Expression	Name
$\mathbf{E}[X]$	mean
$\mathbf{E}[X^2]$	—
$\mathbf{E}[X^3]$	—
$\mathbf{E}[X^4]$	—

(b) Quantities related to central moments

Expression	Name
$\mathbf{E}[(X-\mu)]$	—
$\mathbf{E}[(X-\mu)^2]$	variance
$\mathbf{E}[(X-\mu)^3/\sigma^3]$	(Fisher's) skewness
$\mathbf{E}[(X-\mu)^4/\sigma^4]$	kurtosis

The *mean* of a distribution is the first raw moment. The *variance* of a distribution is the second central moment. Quantities related to higher order central moments are also of interest; Table 4 lists some of these with associated "names" that you might encounter. Variance is a measure of dispersion about the mean. Skewness is a measure of the lopsidedness of a distribution. If a distribution is symmetric (and its third central moment is defined) then it will have skewness equal to zero. A distribution that is skewed to the left (i.e., the tail of the distribution is longer on the left) will have negative skewness and a distribution that is skewed to the right (i.e., the tail of the distribution is longer on the right) will have positive skewness. Kurtosis is a measure of how "fat" or "heavy" the tails of a distribution are; distributions with heavy tails will have high kurtosis values. Since variance and kurtosis are related to the even-powered central moments, they will always be non-negative.

i Its all Greek … when it comes to kurtosis

The root of kurtosis comes from the Greek word for "bulging" or "convex". You may see a heavy-tailed or high kurtosis distributions described as *leptokurtic* ("narrow" + "bulging") and a light-tailed or low kurtosis distributions described as *platykurtic* ("broad" or "flat" + "bulging"). The "high" and "low" qualifications are made in relation to the tails of the normal distribution; a distribution having the same kurtosis as the normal distribution can be described as *mesokurtic* ("middle" + "bulging").

1 Exploratory data analysis

Exploratory data analysis (EDA) is the process of understanding a data set before performing formal inference or fitting a statistical model. The goal is to identify the important features of the data (and any problems) early.

In practice, EDA answers questions like:

- What are the variables and units?
- How much missing data is there?
- What do the marginal distributions look like?
- Are there outliers or data errors?
- Which variables appear associated?

1.1 A first look in R

Cyclamen Data

The **Cyclamen Data** contain 150 obs. of six features (sequence [], petal length [mm], petal width [mm], aperture diameter [mm], color [pink, white], and group [alpha, beta, gamma, delta, epsilon]) for *cyclamen hederifolium*. These measurements were taken by small groups of former MA22004 students using the same measurement protocol at the University Botanic Gardens in the autumn of 2024.

```
cyclamen |> glimpse()
```

```
Rows: 150
Columns: 6
$ sequence    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, ~
$ petal.width  <dbl> 9, 7, 11, 13, 11, 10, 12, 14, 9, 9, 9, 9, 11, 10, 15, 14, 11, 11, 9~
$ petal.length <dbl> 16, 13, 23, 19, 16, 18, 20, 20, 15, 25, 17, 19, 21, 23, 20, 23, 19,~
$ aperture    <dbl> 5, 4, 6, 6, 7, 6, 7, 8, 7, 6, 6, 6, 7, 8, 9, 7, 9, 8, 6, 7, 6, 7, 7~
$ colour     <fct> white, pink, pink, white, pink, white, pink, white, white, pink, wh~
$ group      <fct> epsilon, epsilon, epsilon, epsilon, epsilon, epsilon, epsilon, epsi~
```

A useful first step is to compute basic summaries and check that the categorical variables have the expected levels.

```
cyclamen |> summary()
```

```
    sequence      petal.width     petal.length     aperture        colour       group
```

```
 Min.   : 1.0   Min.   : 4.00   Min.   : 9.0   Min.   : 4.000   pink :83   alpha  :30
 1st Qu.: 8.0   1st Qu.: 9.00   1st Qu.:18.0   1st Qu.: 6.000   white:67   beta   :30
 Median :15.5   Median :10.00   Median :21.0   Median : 7.000              delta  :30
 Mean   :15.5   Mean   :10.29   Mean   :20.5   Mean   : 7.173              epsilon:30
 3rd Qu.:23.0   3rd Qu.:12.00   3rd Qu.:23.0   3rd Qu.: 8.000              gamma  :30
 Max.   :30.0   Max.   :19.00   Max.   :31.0   Max.   :11.000
```

The summary above includes the sample mean of the continuous variables (e.g., `petal.width`) and the counts in each category for factors (e.g., `group`).

> **Missingness**
>
> In R, most functions will silently return `NA` if you have missing values. When computing summaries, decide in advance how you want to handle missingness (for example, removing incomplete rows for a particular calculation, or imputing missing values).

A simple missingness check is:

```
cyclamen |> summarise(across(everything(), ~ sum(is.na(.))))
```

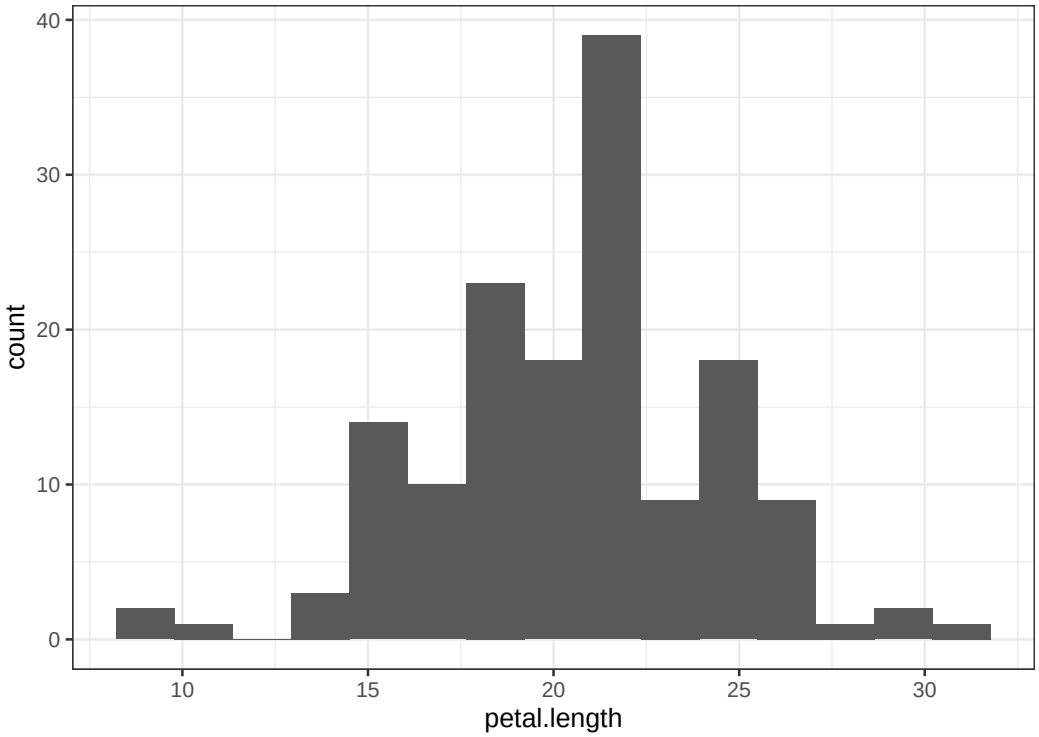
```
# A tibble: 1 x 6
  sequence petal.width petal.length aperture colour group
     <int>       <int>        <int>    <int>  <int> <int>
1        0           0            0        0      0     0
```

In EDA we typically combine numerical summaries with visual summaries. The next few sections outline basic visualisations.

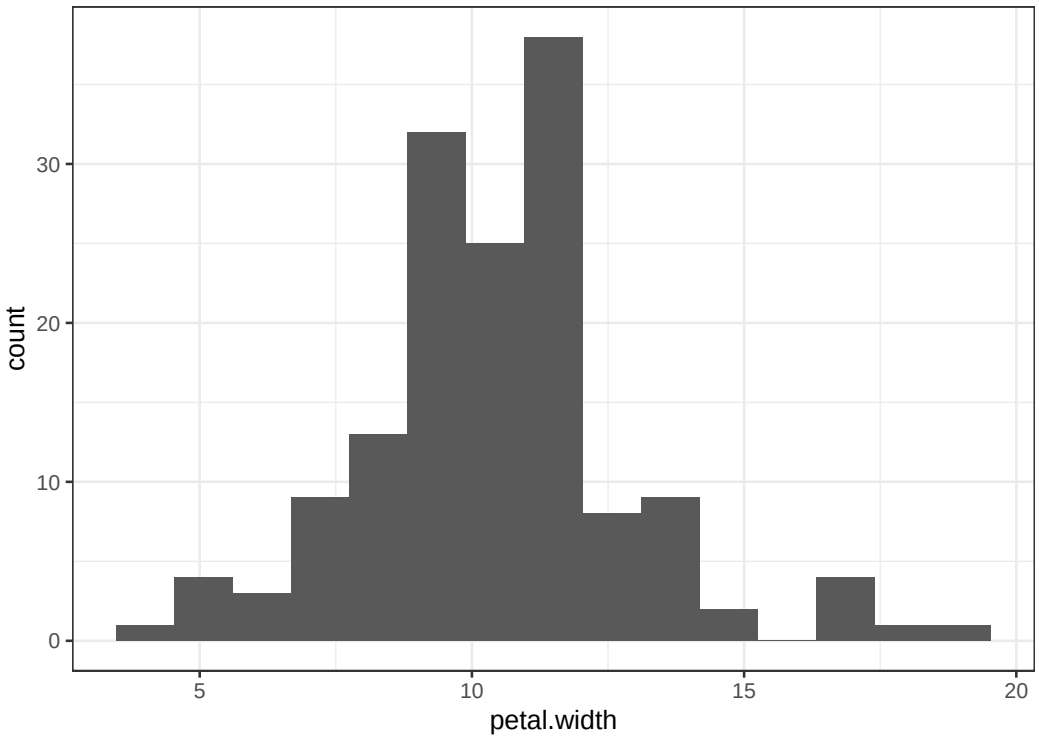
1.2 Histograms and boxplots

Below are some standard plots for petal measurements. Histograms help you see shape (symmetry, skewness, multimodality), and boxplots highlight spread and potential outliers.

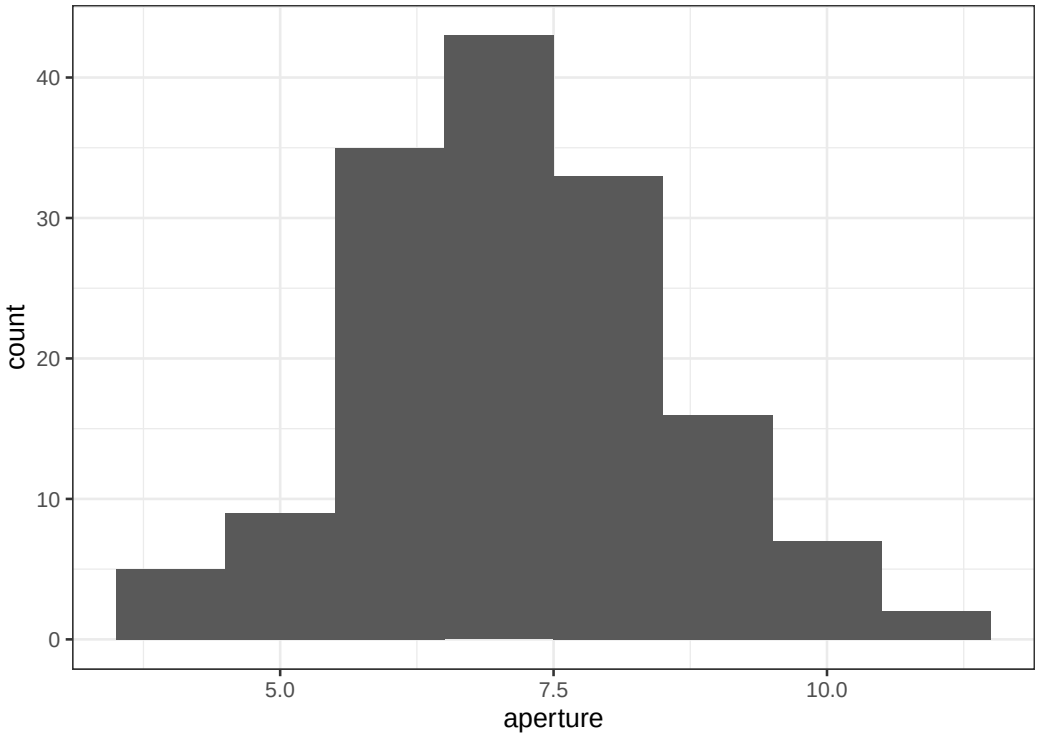
```
cyclamen |>
    ggplot(aes(x = petal.length)) +
    geom_histogram(bins = 15)
```



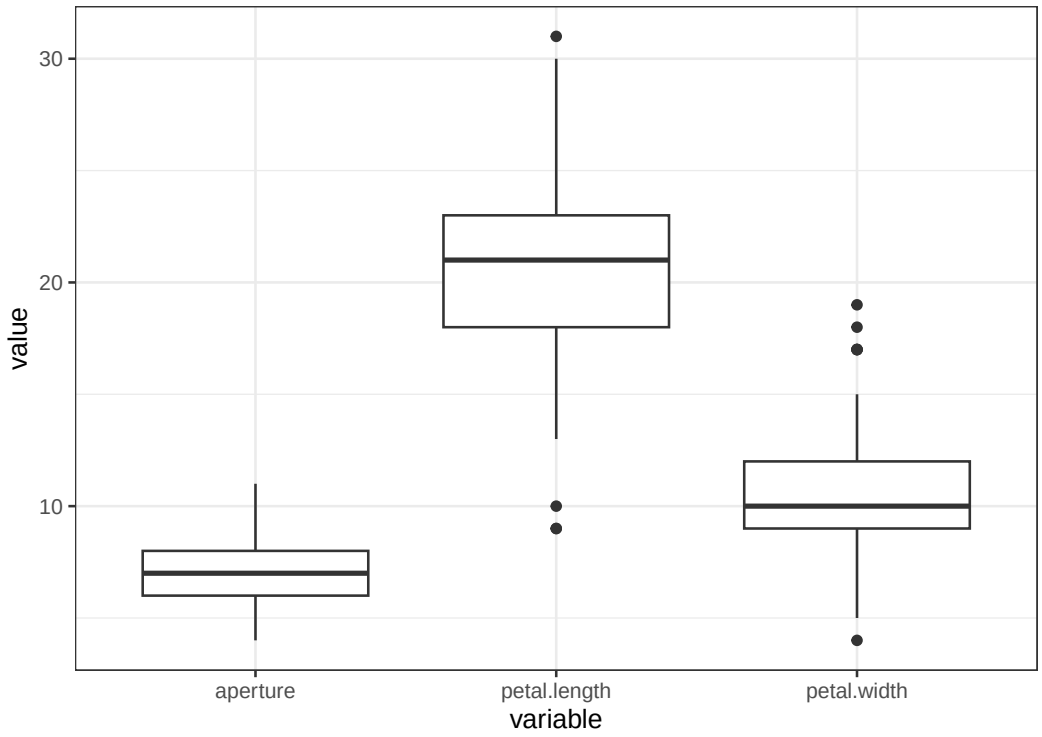

```
cyclamen |>
    ggplot(aes(x = petal.width)) +
    geom_histogram(bins = 15)
```




```
cyclamen |>
    ggplot(aes(x = aperture)) +
    geom_histogram(bins = 8)
```




```
cyclamen |>
    pivot_longer(cols = c(petal.length, petal.width, aperture),
        names_to = "variable",
        values_to = "value") |>
            ggplot(aes(x = variable, y = value)) +
            geom_boxplot()
```

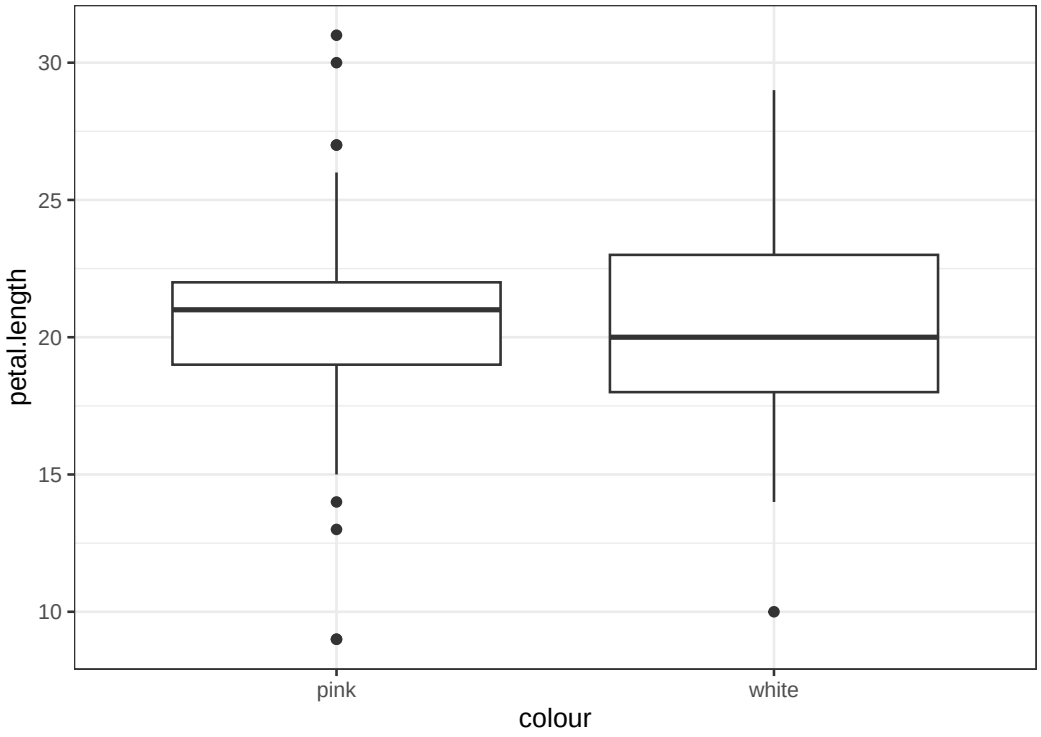


> **Outliers are not necessarily errors**
>
> A point that looks like an outlier is not automatically a data *error*. It may be a legitimate observation (for example, an unusually large flower), but it is always worth checking whether the measurement could have been recorded incorrectly (for example, were the correct units used).

1.3 Comparing groups

A common EDA question is whether distributions differ across a group variable. Here, color is a natural grouping to start with.

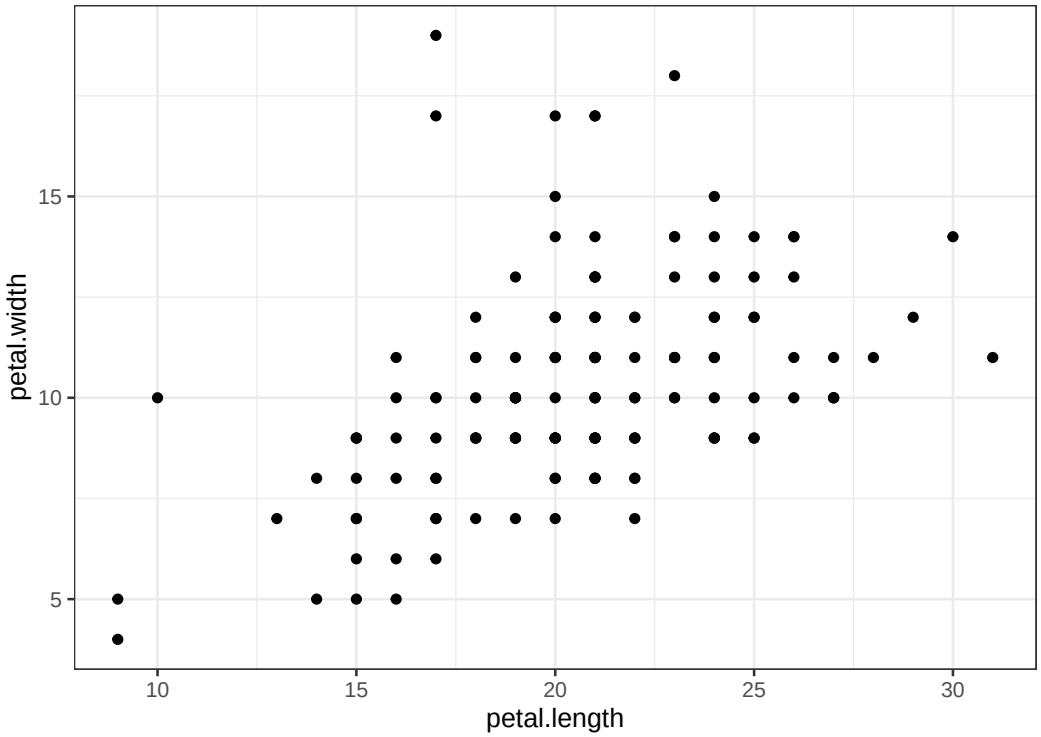
```
cyclamen |>
    ggplot(aes(x = colour, y = petal.length)) +
    geom_boxplot()
```



1.4 Relationships between variables

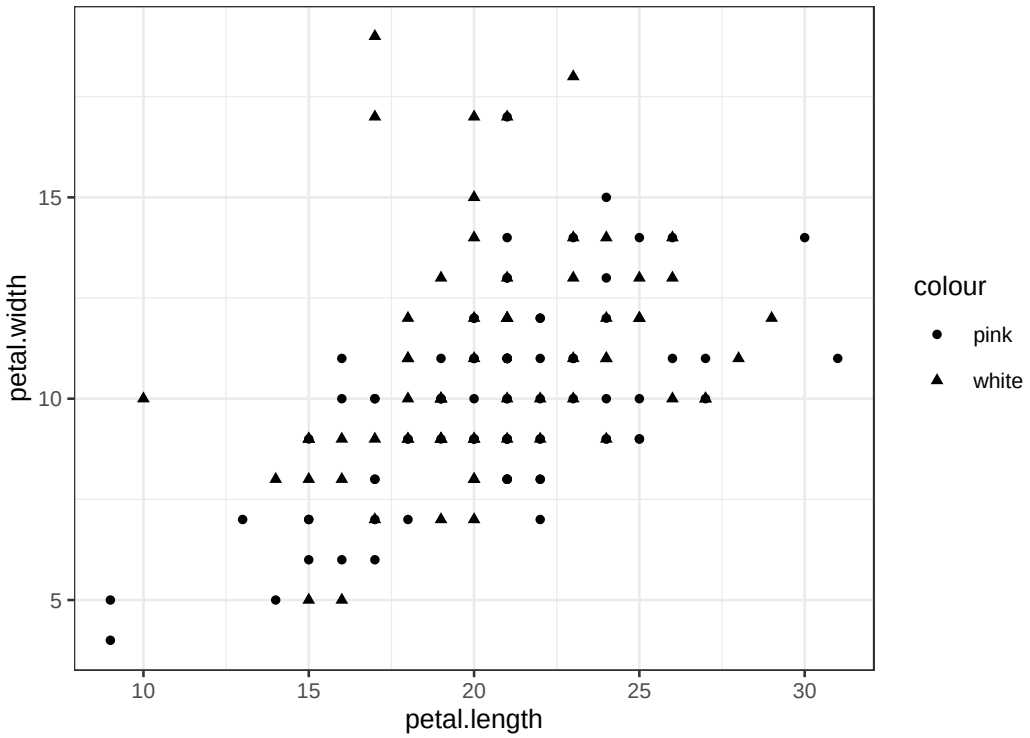
For two numerical variables, a scatter plot is usually the first plot to make.

```
cyclamen |>
    ggplot(aes(x = petal.length, y = petal.width)) +
    geom_point()
```



Sometimes it is helpful to add a grouping variable (here: `colour`) to see whether patterns differ between groups.

```
cyclamen |>
    ggplot(aes(x = petal.length, y = petal.width, shape = colour)) +
    geom_point()
```



1.5 Correlation as a summary of association

Correlation coefficients give a one-number summary of association between two numerical variables. Correlation does not imply causation!

In this module, we will use three common correlation coefficients:

- **Pearson** correlation measures linear association (it is the default and is most directly connected to linear regression).
- **Spearman** correlation measures monotone association by applying Pearson correlation to the ranks of the data.
- **Kendall** correlation (Kendall's τ) measures association using concordant and discordant pairs, and also measures monotone association.

In EDA, it is common to compute Pearson, Spearman, and Kendall correlations together. If Pearson is small but Spearman/Kendall are large (in magnitude), that often suggests a relationship that is monotone but not well-approximated by a straight line.

> A good EDA habit
>
> 1. Make a scatter plot first.
> 2. Compute Pearson correlation.
> 3. If the relationship looks monotone but not linear, or if outliers look influential, also compute Spearman and Kendall as a sensitivity check.

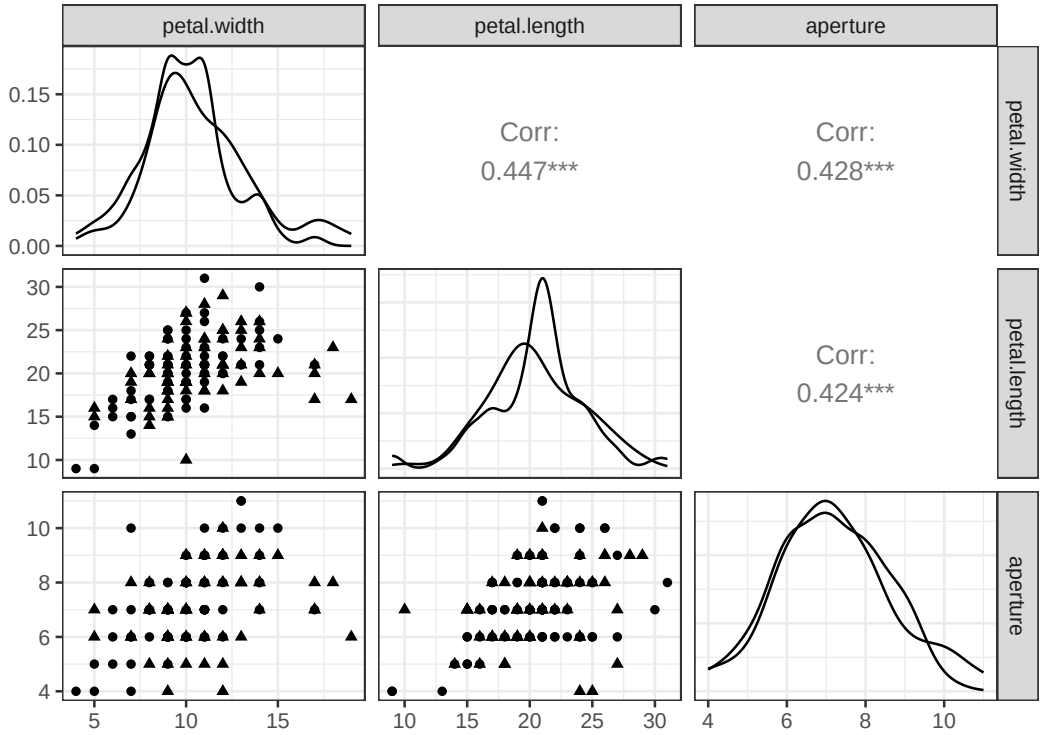


Figure 1.1: Pairwise scatterplot matrix for cyclamen petal width, petal length, and aperture diameter (using `GGally:ggpairs`). Points are shaped by flower colour. The diagonal panels show each variable's distribution (one density per colour); the upper panels summarise pairwise association using the *Pearson* correlation.

The built-in function `cor` can be used to compute different correlation coefficients by specifying the method. For the cyclamen data, we will compute correlations between the three numerical flower measurements (here we exclude `sequence`, which is an index rather than a flower measurement).

```
cyclamen_num <- cyclamen |>
    select(petal.width, petal.length, aperture)
```

1.5.1 Pearson correlation matrix

```
cor(cyclamen_num, use = "complete.obs", method = "pearson")
```

```
             petal.width petal.length  aperture
petal.width    1.0000000    0.4474288 0.4277849
petal.length   0.4474288    1.0000000 0.4238784
aperture       0.4277849    0.4238784 1.0000000
```

1.5.2 Spearman correlation matrix

```
cor(cyclamen_num, use = "complete.obs", method = "spearman")
```

```
             petal.width petal.length  aperture
petal.width    1.0000000    0.4852075 0.4734670
petal.length   0.4852075    1.0000000 0.4154867
aperture       0.4734670    0.4154867 1.0000000
```

1.5.3 Kendall correlation matrix

```
cor(cyclamen_num, use = "complete.obs", method = "kendall")
```

```
             petal.width petal.length  aperture
petal.width    1.0000000    0.3689397 0.3751712
petal.length   0.3689397    1.0000000 0.3353495
aperture       0.3751712    0.3353495 1.0000000
```

If Pearson is close to 0 but Spearman/Kendall are not, that often suggests a relationship that is monotone but not well-approximated by a straight line. If Pearson differs substantially from Spearman/Kendall, it is also worth checking for outliers.

1.6 Principal Component Analysis

Principal Component Analysis (PCA) is an exploratory technique for understanding variation in multivariate numerical data. It constructs new variables (principal components) which are linear combinations of the original variables and which capture decreasing amounts of variation.

For cyclamen, we will apply PCA to the three numerical flower measurements: petal width, petal length, and aperture diameter.

> ⚠ Rescaling
>
> PCA is sensitive to the scale of the variables. When variables are measured on different scales (or have very different spreads), it is standard practice to *standardise* them before doing PCA. In `prcomp`, this is done with `scale. = TRUE`.

```
cyclamen_pca <- cyclamen |>
    select(petal.width, petal.length, aperture, colour)

pca_fit <- cyclamen_pca |>
    select(!colour) |>
    prcomp(scale. = TRUE)

summary(pca_fit)
```

```
Importance of components:
                          PC1    PC2    PC3
Standard deviation     1.366 0.7625 0.7432
Proportion of Variance 0.622 0.1938 0.1841
Cumulative Proportion  0.622 0.8159 1.0000
```

1.6.1 Understanding the PCA summary

For three variables, PCA produces three principal components (PC1, PC2, PC3). The `summary()` output reports:

- **Standard deviation** of each principal component: this is the spread of the data along that component.
- **Proportion of Variance** explained by each component: the fraction of the total variability captured by that component.
- **Cumulative Proportion** or the fraction captured by the first k (reading left to right across the table).

In particular, the variance explained by PCk is `pca_fit$sdev[k]^2`, and the proportion of variance explained is obtained by dividing by the total:

$$\text{PVE}_k = \frac{\text{sdev}_k^2}{\sum_{j=1}^{3} \text{sdev}_j^2}.$$

Because we used `scale. = TRUE`, each original variable is standardised to have variance 1, so the total variance is approximately 3 (up to rounding). This is why $\sum_{k=1}^{3} \text{sdev}_k^2 \approx 3$.

For the cylcamen data, the summary output shows:

- PC1 explains about 62.2% of the variation in the three measurements.
- PC2 explains about 19.4%, and PC3 about 18.4%.
- The first two components together explain about 81.6% of the variation.

```
pca_var <- pca_fit$sdev^2
pca_pve <- pca_var / sum(pca_var)

pve_tbl <-
    tibble(
        PC = factor(paste0("PC", seq_along(pca_pve)),
        levels = paste0("PC", seq_along(pca_pve))),
        proportion = pca_pve,
        cumulative = cumsum(pca_pve)
        )

pve_tbl
```

```
# A tibble: 3 x 3
  PC    proportion cumulative
  <fct>      <dbl>      <dbl>
1 PC1        0.622      0.622
2 PC2        0.194      0.816
3 PC3        0.184      1
```

1.6.2 PCA scores plot

A useful EDA plot is the projection of the observations onto the first two principal components. Here we colour by `colour` to check whether colour is associated with the main modes of variation in the measurements.

```
scores <-
    as_tibble(pca_fit$x) |>
    bind_cols(cyclamen_pca |> select(colour))

scores |>
    ggplot(aes(x = PC1, y = PC2, shape = colour)) +
    geom_point()
```

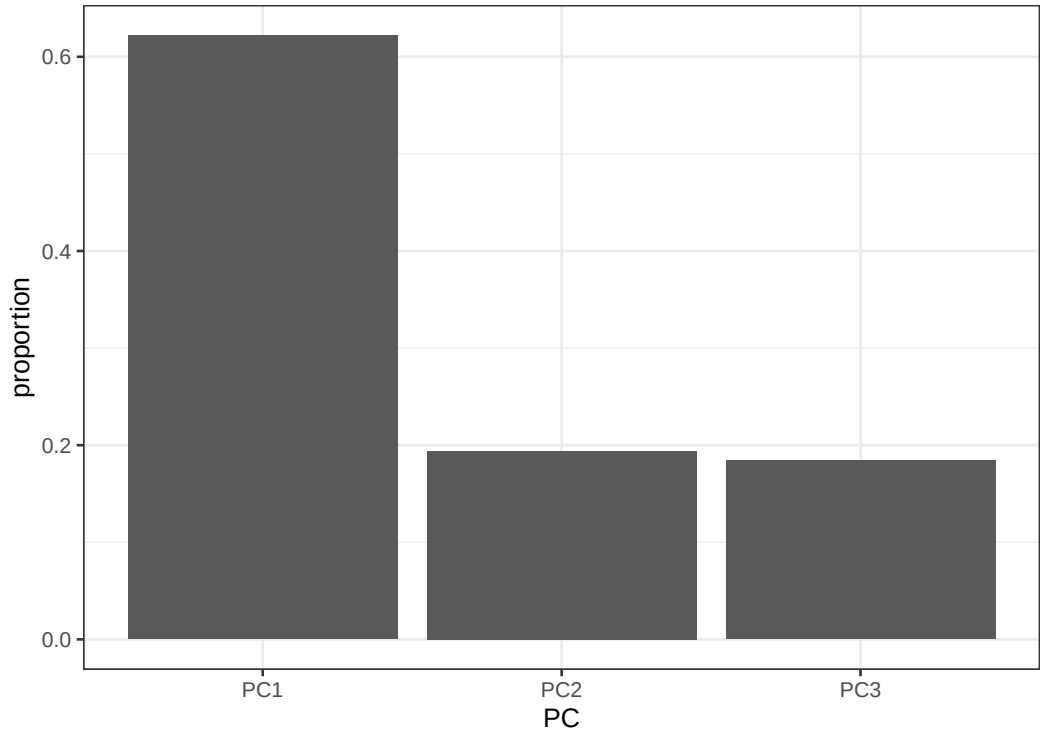
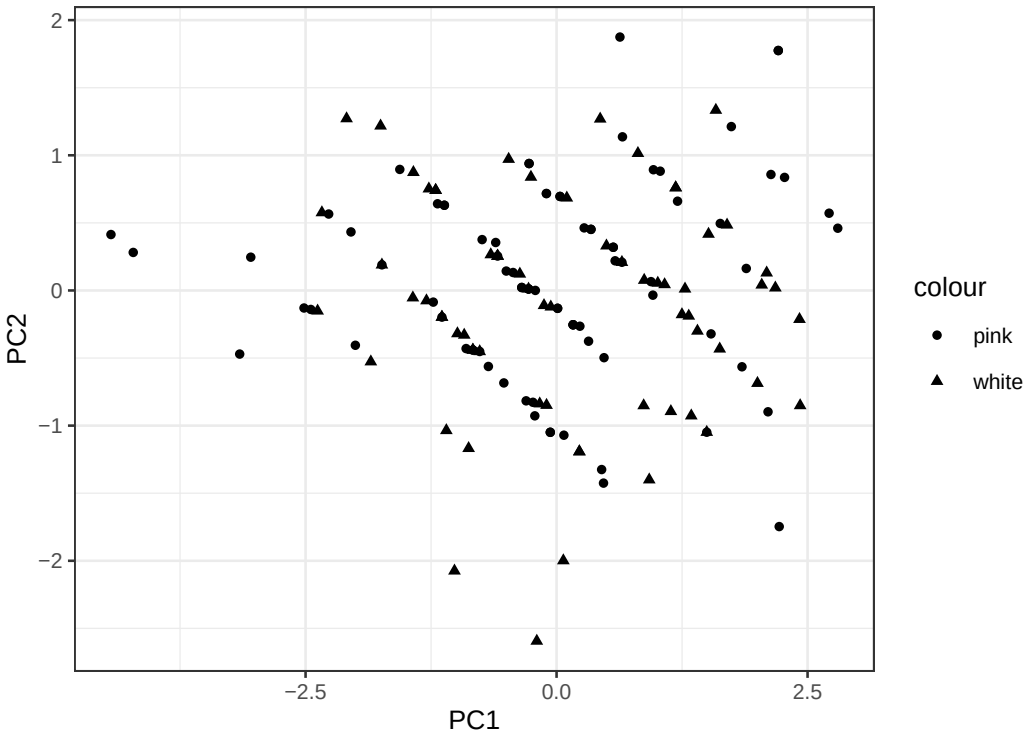


Figure 1.2: **Scree plot**: proportion of variance explained by each principal component in the cyclamen PCA.



The scatter plot above displays cyclamen PCA scores (PC1 vs PC2). Each point is a flower; points are shaped by flower colour.

1.6.3 Interpreting the principal components

To interpret the components, we examine the **loadings** (the coefficients defining each principal component as a linear combination of the original variables). In `prcomp`, these are stored in `pca_fit$rotation`.

```
as_tibble(pca_fit$rotation, rownames = "variable")
```

```
# A tibble: 3 x 4
  variable       PC1    PC2     PC3
  <chr>        <dbl>  <dbl>   <dbl>
1 petal.width  0.581 -0.347 -0.736
2 petal.length 0.580 -0.458  0.674
3 aperture     0.571  0.818  0.0649
```

A loading with larger magnitude means that the corresponding variable contributes more strongly to that principal component.

> ⚠ PCA is not a hypothesis test
>
> PCA is an exploratory tool to help you describe patterns and identify dominant directions of variation. A common outcome is dimension reduction: here, PC1 and PC2 capture about 81.6% of the variation, so a two-dimensional view retains most of the information in the three measurements.

2 Sampling distributions

A *statistic* is a quantity that can be calculated from sample data. Before observing data, a statistic is an unknown quantity and is, therefore, a rv.

Definition 2.1 (Statistic)**.** Let $X_1, \dots, X_n$ be observable rvs and let g be an arbitrary real-valued function of n random variables. The rv

$$T = g(X_1, \dots, X_n)$$

is a statistic.

We refer to the probability distribution for a statistic as a sampling distribution. The sampling distribution illustrates how the statistic will vary across possible sample data. The sampling distribution contains information about the values a statistic is likely to assume and how likely it is to assume those values prior to observing data.

Definition 2.2 (Sampling distribution)**.** Suppose rvs $X_1, \dots, X_n$ are a random sample from $F(\theta)$, a distribution depending on a parameter θ whose value is unknown. Let the rv

$$T = g(X_1, \dots, X_n, \theta)$$

be a function of $X_1, \dots, X_n$ and (possibly) θ. The distribution of T (given θ) is the sampling distribution of T.

The sampling distribution of T is derived from the distribution of the random sample. Often we will be interested in a statistic T that is an estimator for a parameter θ (that is, T will not depend on θ).

In what follows, we review several special families of distributions that are widely used in probability and statistics. These special families of distributions will be indexed by one or more parameters and include discrete distributions (Bernoulli, Binomial, Poisson, and discrete Uniform) as well as continuous distributions (continuous Uniform, Normal, Student's t, χ^2, and F).

2.1 Bernoulli distribution

The Bernoulli distribution describes a single trial with two possible outcomes, often coded as 1 (success) and 0 (failure). It is the basic building block behind proportions and binomial counts.

Definition 2.3 (Bernoulli distribution)**.** A discrete rv X has a Bernoulli distribution with parameter $p \in (0, 1)$ if

$$P(X = x; p) = \begin{cases} p, & x = 1, \\ 1 - p, & x = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Equivalently,

$$P(X = x; p) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\}. \tag{2.1}$$

We write $X \sim \text{Bernoulli}(p)$.

> **Parameter**
>
> The parameter p is the probability of success, i.e. $p = P(X = 1)$.

If $X \sim \text{Bernoulli}(p)$, then it can be shown that

$$\mathbf{E}[X] = p \qquad \text{and} \qquad \text{Var}(X) = p(1-p). \tag{2.2}$$

The pmf for two Bernoulli rvs with different values of p is shown in Figure 2.1.

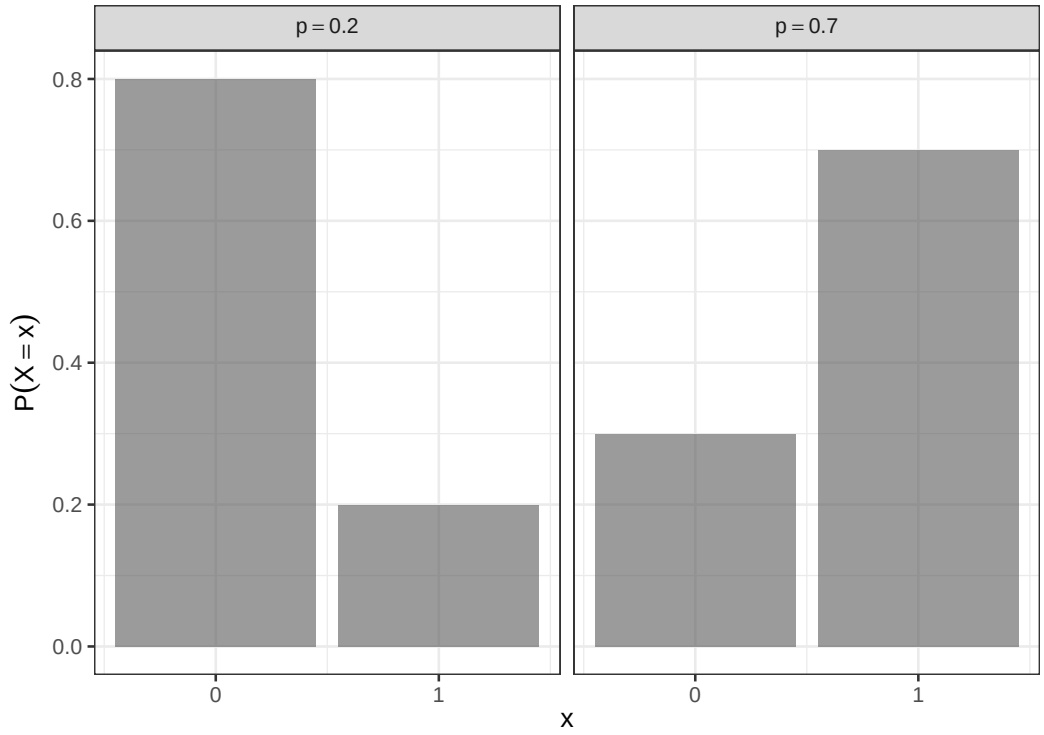


Figure 2.1: The pmf of a Bernoulli rv for two values of the success probability p.

Example 2.1. Suppose that a component passes a quality check with probability $p = 0.95$. Let $X = 1$ if the component passes and $X = 0$ otherwise. Then $X \sim \text{Bernoulli}(0.95)$, and the probability that the component fails is

$$P(X = 0) = 1 - p = 0.05.$$

> **Indicator variables**
>
> A Bernoulli rv is often used as an **indicator**: it takes value 1 when a property holds, and 0 when it does not. This is exactly the setup used for estimating proportions in Section 4.2.1.

2.2 Binomial distribution

The binomial distribution describes the number of successes in k independent Bernoulli trials, each having the same success probability p.

Definition 2.4 (Binomial distribution)**.** Let X denote the number of successes in k independent trials, where each trial is a Bernoulli(p) rv. Then X has a binomial distribution with parameters $k \in \mathbf{N}_>$ and $p \in (0, 1)$ if

$$P(X = x; k, p) = \binom{k}{x} p^x (1-p)^{k-x}, \quad x = 0, 1, \ldots, k. \tag{2.3}$$

We write $X \sim \text{Binomial}(k, p)$.

> **Parameters**
>
> The binomial distribution has **two parameters**: the number of trials k and the success probability p.

If $X \sim \text{Binomial}(k, p)$, then

$$\mathbf{E}[X] = kp \qquad \text{and} \qquad \text{Var}(X) = kp(1-p). \tag{2.4}$$

The pmf for Binomial(10, p) for several values of p is shown in Figure 2.2.

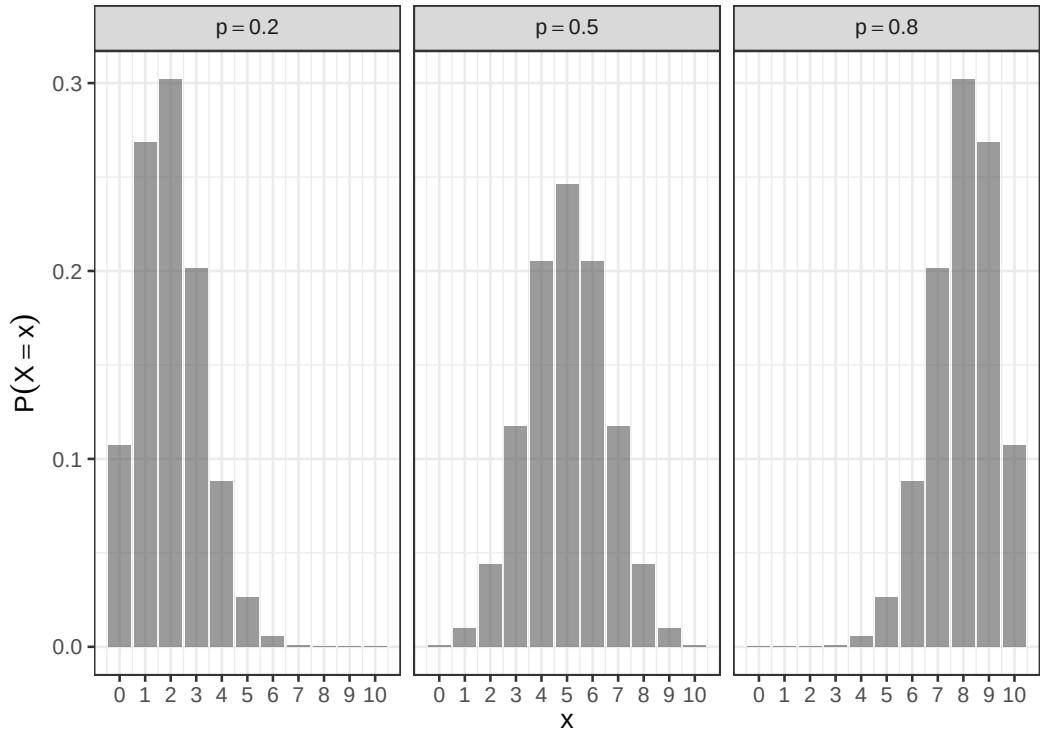


Figure 2.2: The pmf of $X \sim \text{Binomial}(10, p)$ for several values of p.

Example 2.2. A student answers $k = 10$ multiple-choice questions by guessing. Each question has probability $p = 0.25$ of being correct. If X is the number of correct answers, then

$$X \sim \text{Binomial}(10, 0.25).$$

The probability of getting exactly 3 correct answers is

$$P(X = 3) = \binom{10}{3}(0.25)^3(0.75)^7.$$

This can be computed in R as `dbinom(3, size = 10, prob = 0.25)`.

> **Binomial counts are sums of Bernoulli trials**
>
> If $X_1, \ldots, X_k \sim \text{Bernoulli}(p)$ are independent, then
>
> $$X = \sum_{i=1}^{k} X_i \sim \text{Binomial}(k, p).$$
>
> This is why binomial distributions appear naturally when counting the number of times an event occurs in repeated trials.

2.3 Poisson distribution

The Poisson distribution is widely used to model the number of events in a fixed unit of exposure (often time or space), such as the number of arrivals at a desk per hour or the number of faults detected per kilometre.

Definition 2.5 (Poisson distribution)**.** A discrete rv X has a Poisson distribution with parameter $\lambda > 0$ if

$$P(X = x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots \tag{2.5}$$

We write $X \sim \text{Poisson}(\lambda)$.

> **Parameter**
>
> The parameter λ is the mean number of events in the unit of exposure (e.g. "per hour").

If $X \sim \text{Poisson}(\lambda)$, then

$$\mathbf{E}[X] = \lambda \qquad \text{and} \qquad \text{Var}(X) = \lambda. \tag{2.6}$$

That is, the Poisson distribution has the distinctive property that the **mean equals the variance**. The pmf for several values of λ is shown in Figure 2.3.

Example 2.3. Suppose the number of emails you receive in an hour is modelled as $X \sim \text{Poisson}(\lambda)$ with $\lambda = 2.5$. The probability of receiving no emails in the next hour is

$$P(X = 0) = e^{-2.5}\frac{2.5^0}{0!} = e^{-2.5}.$$

In R, this is `dpois(0, lambda = 2.5)`.

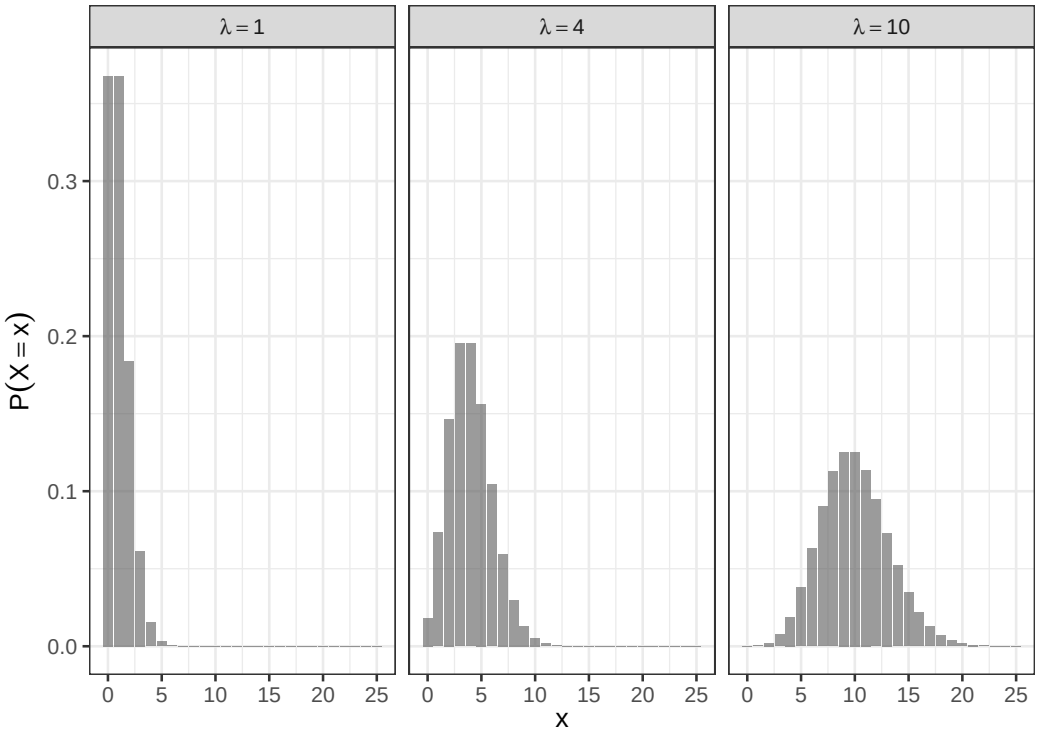


Figure 2.3: The pmf of a Poisson rv for several values of the rate parameter λ.

i When does Poisson modelling make sense?

Poisson models are commonly used when (i) events occur one at a time, (ii) they occur "at random" throughout the exposure period, and (iii) the average event rate is roughly constant over the period being observed.

2.4 Uniform Distribution

The uniform distribution places equal weight on the items being sampled. The items can be discrete or be a continuum.

2.4.1 Discrete uniform distribution

The discrete uniform places equal probability on a *finite* set of possible values.

Definition 2.6 (Discrete uniform distribution)**.** A discrete rv X has a discrete uniform distribution on the integers $\{a, a+1, \ldots, b\}$ with $a < b$, if

$$P(X = x; a, b) = \begin{cases} \frac{1}{b-a+1}, & x \in \{a, a+1, \ldots, b\}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

We write $X \sim \mathsf{Unif}\{a, \ldots, b\}$.

Parameters

The parameters a and b determine the finite support $\{a, a+1, \ldots, b\}$.

If $X \sim \text{Unif}\{a, \ldots, b\}$, then

$$\mathbf{E}[X] = \frac{a+b}{2} \qquad \text{and} \qquad \text{Var}(X) = \frac{(b-a+1)^2 - 1}{12}. \tag{2.8}$$

A familiar example is the outcome of a fair six-sided die: $X \sim \text{Unif}\{1, \ldots, 6\}$. The pmf is shown in Figure 2.4.

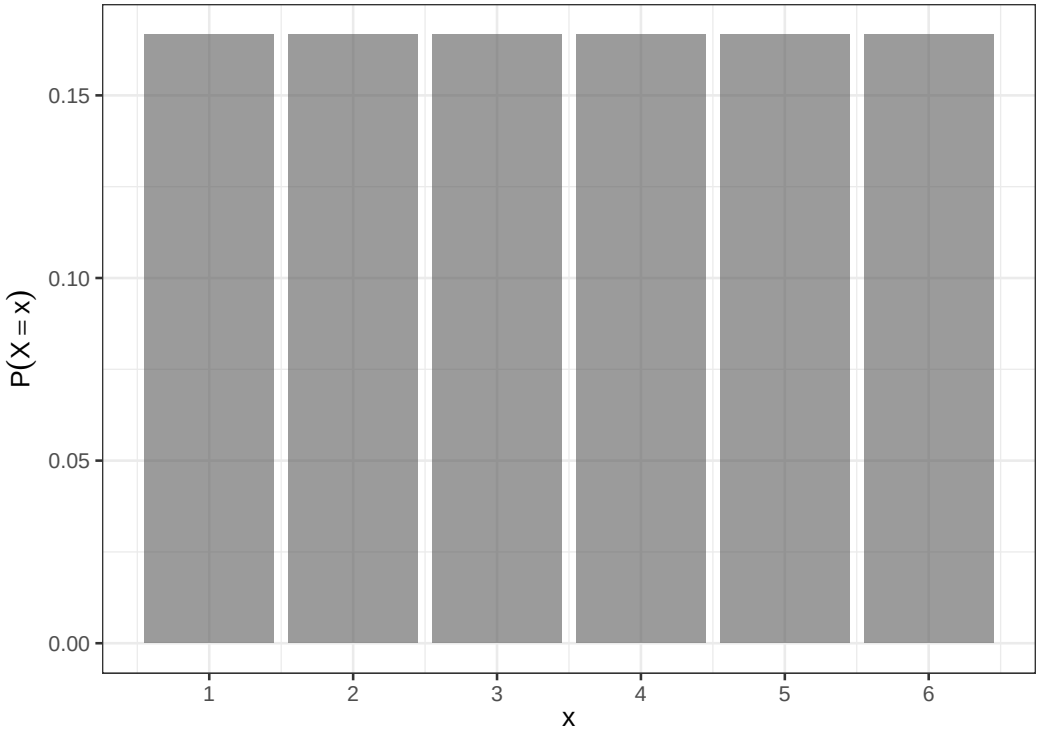


Figure 2.4: The pmf of a discrete uniform rv on $\{1, 2, 3, 4, 5, 6\}$ (a fair die).

Example 2.4. Let X be the outcome of a fair die roll. Then $X \sim \text{Unif}\{1, \ldots, 6\}$. The probability of rolling a value at least 5 is

$$P(X \geq 5) = P(X = 5) + P(X = 6) = \frac{2}{6} = \frac{1}{3}.$$

2.4.2 Continuous uniform distribution

Compared with the discrete version, the continuous uniform distribution places equal weight across an interval.

Definition 2.7 ((Continuous) Uniform distribution)**.** A continuous rv X has a uniform distribution on $[a, b]$ with $a < b$, if X has pdf

$$f(x; a, b) = \frac{1}{b-a}, \quad a < x < b,$$

or zero otherwise. We write $X \sim \text{Unif}(a, b)$.

⚠ Parameters

Note that a and b are parameters in Definition 2.7.

Exercise 2.1. As an exercise, derive the cdf using the definition. Derive a formula for the mean and variance in terms of the parameters a and b.

2.5 Normal distribution

Normal distributions play an important role in probability and statistics as they describe many natural phenomena. For instance, the Central Limit Theorem tells us that the sample mean of a large random sample (size m) of rvs with mean μ and variance σ^2 is approximately normal in distribution with mean μ and variance σ^2/m.

Definition 2.8 (Normal or Gaussian distribution)**.** A continuous rv X has a normal distribution with parameters μ and σ^2, where $-\infty < \mu < \infty$ and $\sigma > 0$, if X has pdf

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty.$$

We write $X \sim \mathsf{N}(\mu, \sigma^2)$.

For $X \sim \mathsf{N}(\mu, \sigma^2)$, it can be shown that $\mathbf{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, that is, μ is the *mean* and σ^2 is the *variance* of X. The pdf forms a bell-shaped curve that is symmetric about μ, as illustrated in Figure 2.5. The value σ (*standard deviation*) is the distance from μ to the inflection points of the curve. As σ increases, the dispersion in the density increases, as illustrated in Figure 2.6. Thus, the distribution's position (location) and spread depend on μ and σ.

Definition 2.9 (Standard normal distribution)**.** We say that X has a standard normal distribution if $\mu = 0$ and $\sigma = 1$ and we will usually denote standard normal rvs by

$$Z \sim \mathsf{N}(0, 1)$$

(why Z? tradition![1]). We denote the cdf of the standard normal by

$$\Phi(z) = P(Z \leq z)$$

and write $\varphi = \Phi'$ for its density function.

❗ Useful facts about normal variates

1. If $X \sim \mathsf{N}(\mu, \sigma^2)$, then
$$Z = (X - \mu)/\sigma \sim \mathsf{N}(0, 1).$$
2. If $Z \sim \mathsf{N}(0, 1)$, then
$$X = \mu + \sigma Z \sim \mathsf{N}(\mu, \sigma^2).$$

[1] "Traditions, traditions… Without our traditions, our lives would be as shaky as a fiddler on the roof!" [https://www.youtube.com/watch?v=gRdfX7ut8gw].

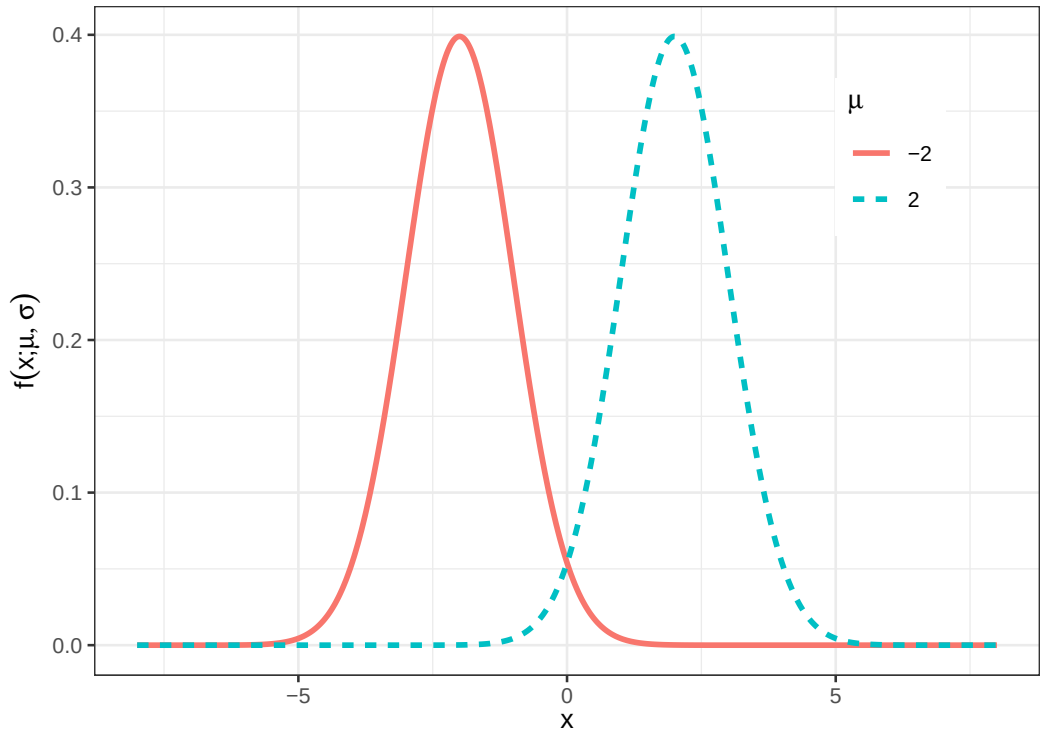


Figure 2.5: The pdfs of two normal rvs, $X_1 \sim \mathsf{N}(-2, 1)$ and $X_2 \sim \mathsf{N}(2, 1)$, with *different means* and the same standard deviations.

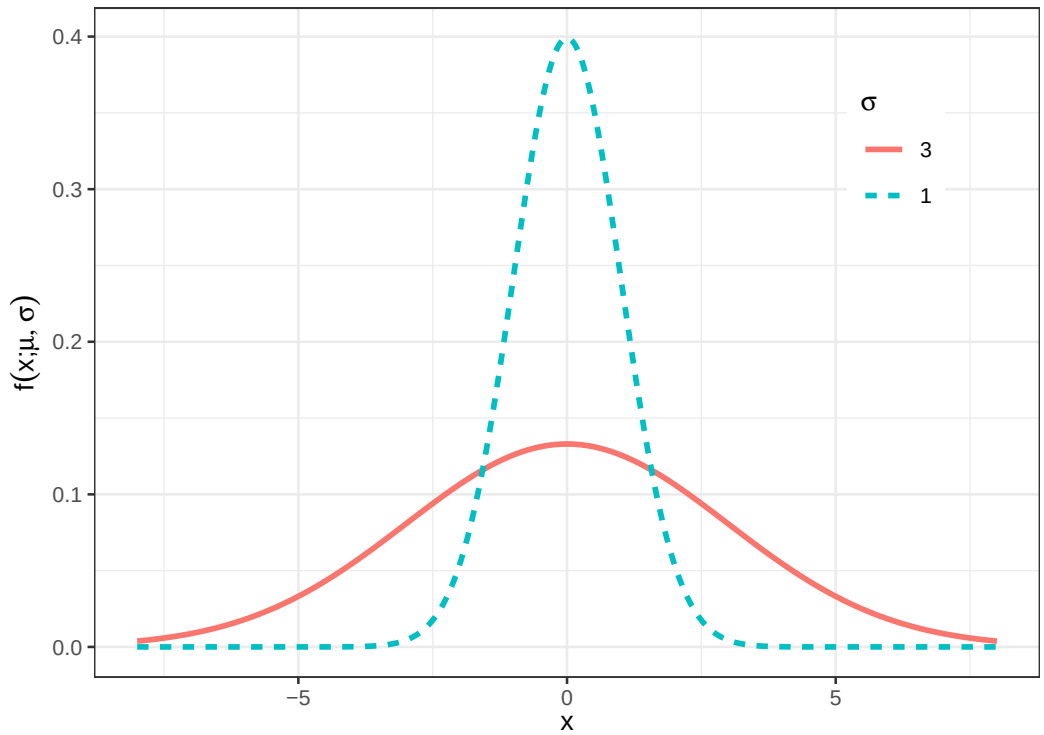


Figure 2.6: The pdfs of two normal rvs, $X_1 \sim \mathsf{N}(0, 9)$ and $X_2 \sim \mathsf{N}(0, 1)$, with the same means and *different standard deviations*.

3. If $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$ are independent rvs, then

$$\sum_{i=1}^{n} X_i \sim \mathsf{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

Variances add

In particular, for differences of independent rvs $X_1 \sim \mathsf{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathsf{N}(\mu_2, \sigma_2^2)$ then the variances add:

$$X_1 - X_2 \sim \mathsf{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2).$$

Probabilities $P(a \leq X \leq b)$ are found by converting the problem in $X \sim \mathsf{N}(\mu, \sigma^2)$ to the *standard normal* distribution $Z \sim \mathsf{N}(0, 1)$ whose probability values $\Phi(z) = P(Z \leq z)$ can then be looked up in a table. From (1.) above,

$$\begin{aligned} P(a < X < b) &= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \end{aligned}$$

This process is often referred to as *standardising* (the normal rv).

Example 2.5. Let $X \sim \mathsf{N}(5, 9)$ and find $P(X \geq 5.5)$.

$$\begin{aligned} P(X \geq 5.5) &= P\left(Z \geq \frac{5.5-5}{3}\right) \\ &= P(Z \geq 0.1667) \\ &= 1 - P(Z \leq 0.1667) \\ &= 1 - \Phi(0.1667) \\ &= 1 - 0.5662 \\ &= 0.4338, \end{aligned}$$

where we look up the value of $\Phi(z) = P(Z \leq z)$ in a table of standard normal curve areas.

The probability corresponds to the shaded area under the normal density $\varphi(x) = \Phi'(x)$ corresponding to $x \geq 5.5$ (see Figure 2.7). To calculate this area, we can also use the R code: `pnorm(5.5, mean = 5, sd = 3, lower.tail = FALSE)`.

Example 2.6. Let $X \sim \mathsf{N}(5, 9)$ and find $P(4 \leq X \leq 5.25)$.

$$\begin{aligned} P(4 \leq X \leq 5.25) &= P\left(\frac{4-5}{3} \leq Z \leq \frac{5.25-5}{3}\right) \\ &= P(-0.3333 \leq Z \leq 0.0833) \\ &= \Phi(0.0833) - \Phi(-0.3333) \\ &= 0.5332 - 0.3694 \\ &= 0.1638. \end{aligned}$$

where we look up the value of $\Phi(z) = P(Z \leq z)$ in a table of standard normal curve areas.

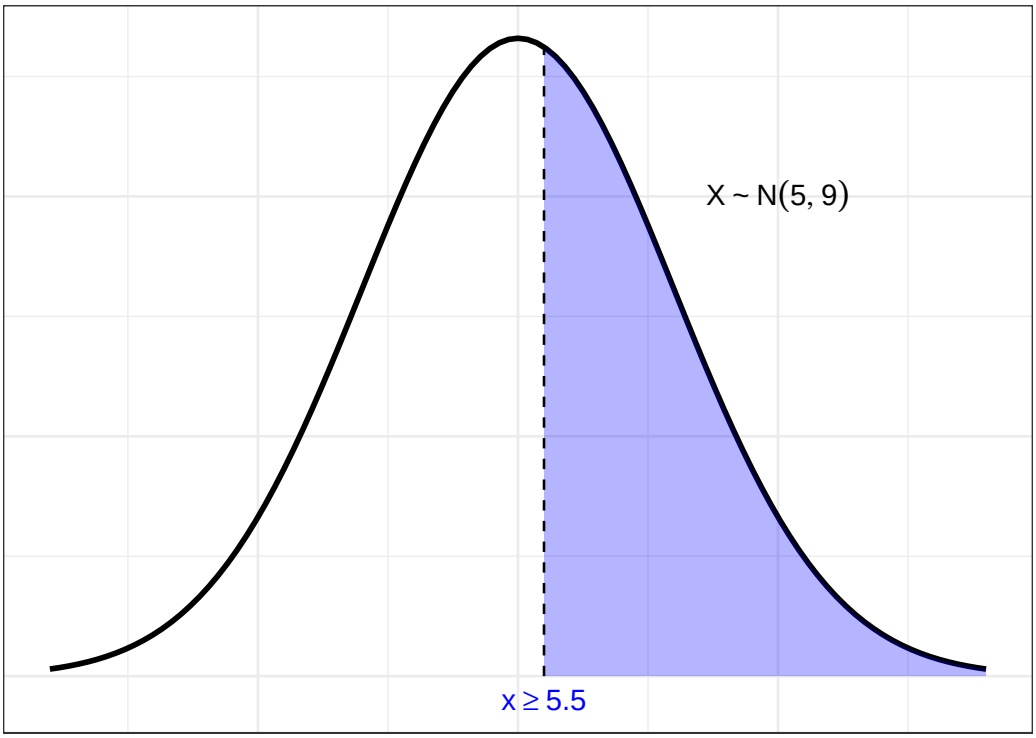


Figure 2.7: The normal density $\mathsf{N}(5, 9)$ with the (one-sided) interval shaded in blue that corresponds to the probability $P(X \geq 5.5)$.

The probability corresponds to the shaded area under the normal density $\varphi(x) = \Phi'(x)$ corresponding to $4 \leq x \leq 5.25$ (see Figure 2.8). To calculate this area, we can use the R code: `pnorm(5.25, mean = 5, sd = 3) - pnorm(4, mean = 5, sd = 3)`.

! Empirical rule (68 – 95 – 99.7 rule)

For samples from a normal distribution, the percentage of values that lie within one, two, and three standard deviations of the mean are 68.27%, 95.45%, and 99.73%, respectively. That is, for $X \sim \mathsf{N}(\mu, \sigma^2)$,

$$P(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx 0.6827\,,$$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.9545\,,$$

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.9973\,.$$

For a normal population, nearly all the values lie within "three sigmas" of the mean.

2.6 Student's t distribution

Student's t distribution gets its peculiar name as it was first published under the pseudonym "Student".[2] This bit of obfuscation was to protect the identity of his employer,[3] and thereby vital trade secrets, in a

[2] William Sealy Gosset (1876–1937) wrote under the pseudonym "Student" [https://mathshistory.st-andrews.ac.uk/Biographies/Gosset/].

[3] Gosset invented the t-test to handle small samples for quality control in brewing, specifically for the Guinness brewery in Dublin [https://www.wikiwand.com/en/Guinness_Brewery].

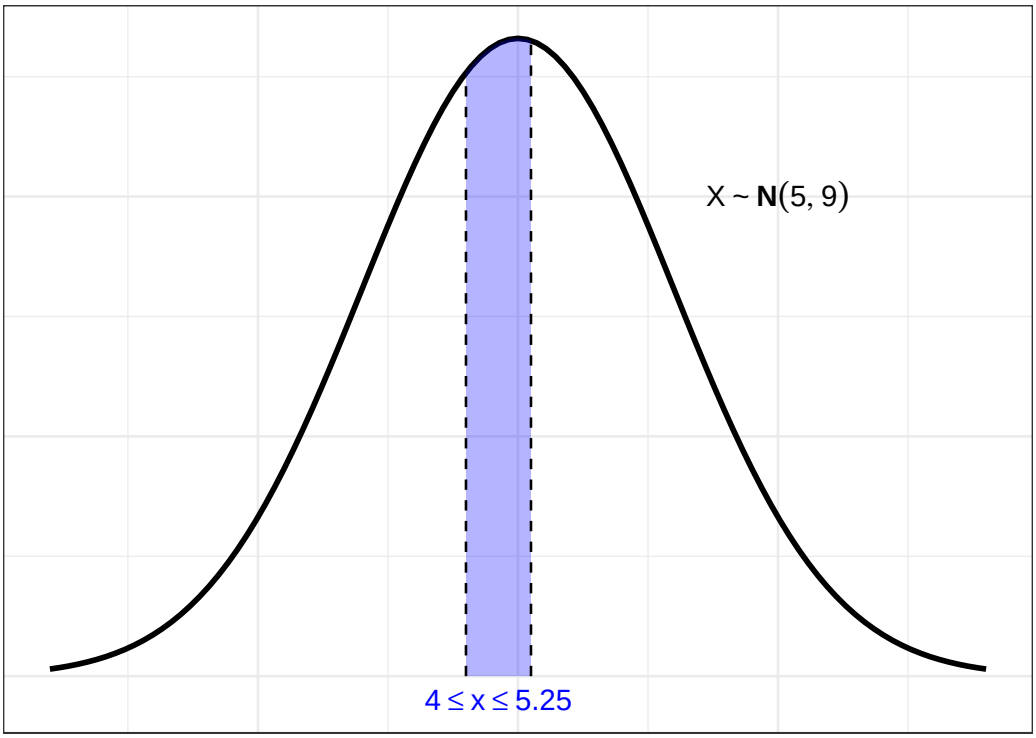


Figure 2.8: The normal density N(5, 9) with the (two-sided) interval shaded in blue that corresponds to the probability $P(4 \leq X \leq 5.25)$.

highly competitive and lucrative industry.

Definition 2.10 (Student's t distribution)**.** A continuous rv X has a t distribution with parameter $\nu > 0$, if X has pdf

$$f(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad -\infty < x < \infty\,.$$

We write $X \sim \mathrm{t}(\nu)$. Note Γ is the standard gamma function.[4]

The density for $\mathrm{t}(\nu)$ for several values of ν are plotted below in Figure 2.9.

! Properties of t distributions

1. The density for $\mathrm{t}(\nu)$ is a bell-shaped curve centred at 0.
2. The density for $\mathrm{t}(\nu)$ is more spread out than the standard normal density (i.e., it has "fatter tails" than the normal).
3. As $\nu \to \infty$, the spread of the corresponding $\mathrm{t}(\nu)$ density converges to the standard normal density (i.e., the spread of the $\mathrm{t}(\nu)$ density decreases relative to the standard normal).

If $X \sim \mathrm{t}(\nu)$, then $\mathbf{E}[X] = 0$ for $\nu > 1$ (otherwise the mean is undefined).

[4]The gamma function is defined by $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$ when the real part of z is positive. For any positive integer n, $\Gamma(n) = (n-1)!$ and for half-integers $\Gamma(\frac{1}{2}+n) = \frac{(2n)!}{4^n n!}\sqrt{\pi}$.

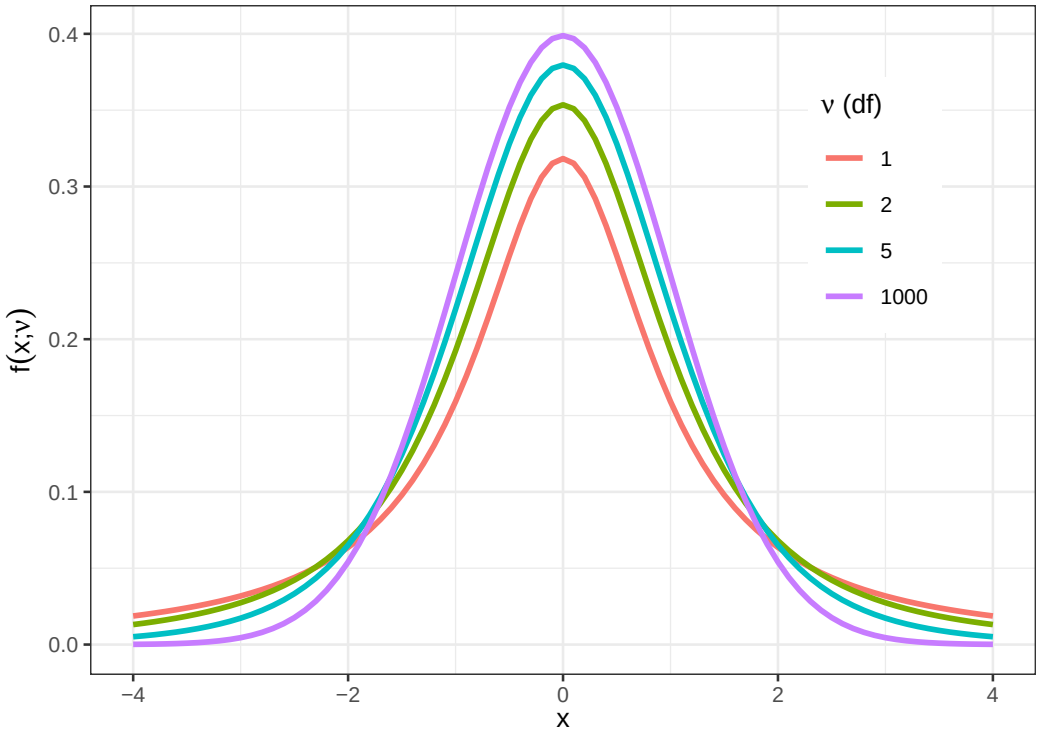


Figure 2.9: The density for t(ν) for several values of ν (df).

Cauchy distribution

A t distributions with $\nu = 1$ has pdf

$$f(x) = \frac{1}{\pi(1+x^2)},$$

and we call this the Cauchy distribution.

2.7 χ^2 distribution

The χ^2 distribution arises as the distribution of a sum of the squares of ν independent standard normal rvs.

Definition 2.11 (χ^2 distribution)**.** A continuous rv X has a χ^2 distribution with parameter $\nu \in \mathbf{N}_>$, if X has pdf

$$f(x;\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2},$$

with support $x \in (0,\infty)$ if $\nu = 1$, otherwise $x \in [0,\infty)$. We write $X \sim \chi^2(\nu)$.

The pdf $f(x;\nu)$ of the $\chi^2(\nu)$ distribution depends on a positive integer ν referred to as the df. The densities for several values of ν are plotted below in Figure 2.10. The density $f(x;\nu)$ is positively skewed, i.e., the right tail is longer, so the mass is concentrated to the figure's left in Figure 2.10. The distribution becomes more symmetric as ν increases. We denote critical values of the $\chi^2(\nu)$ distribution by $\chi^2_{\alpha,\nu}$.

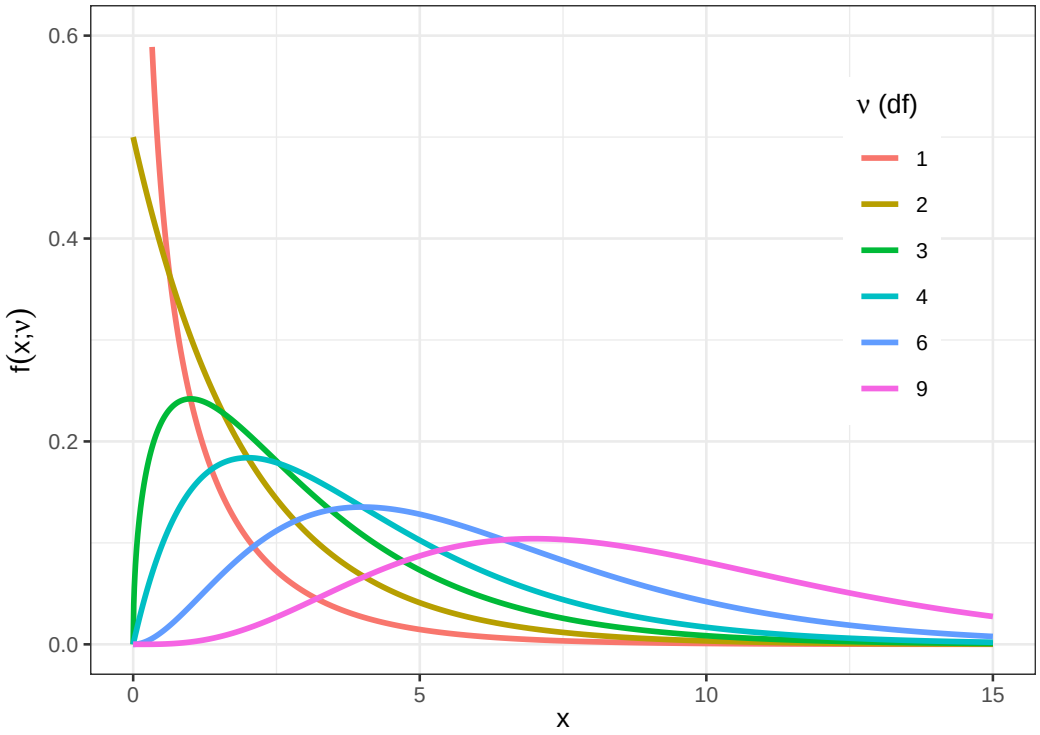


Figure 2.10: The density for $\chi^2(\nu)$ for several values of ν (df).

> **Skew**
>
> Unlike the normal and t distributions, the χ^2 distribution is not symmetric! This means that critical values, e.g.,
>
> $$\chi^2_{.99,\nu} \quad \text{and} \quad \chi^2_{0.01,\nu},$$
>
> are **not** equal. Hence, it will be necessary to look up both values for CIs based on χ^2 critical values.

If $X \sim \chi^2(\nu)$, then $\mathbf{E}[X] = \nu$ and $\text{Var}[X] = 2\nu$.

2.8 F distribution

The F distribution ("F" for Fisher) arises as a test statistic when comparing population variances and in the analysis of variance (see Chapter 6).

Definition 2.12 (F distribution)**.** A continuous rv X has an F distribution with df parameters ν_1 and ν_2, if X has pdf

$$f(x;\nu_1,\nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \frac{x^{\nu_1/2-1}}{(\nu_2+\nu_1 x)^{(\nu_1+\nu_2)/2}}.$$

The pdf $f(x;\nu_1,\nu_2)$ of the $\mathsf{F}(\nu_1,\nu_2)$ distribution depends on two positive integers ν_1 and ν_2 referred to, respectively, as the numerator and denominator df. The density is plotted below for several combinations of (ν_1,ν_2) in Figure 2.11.

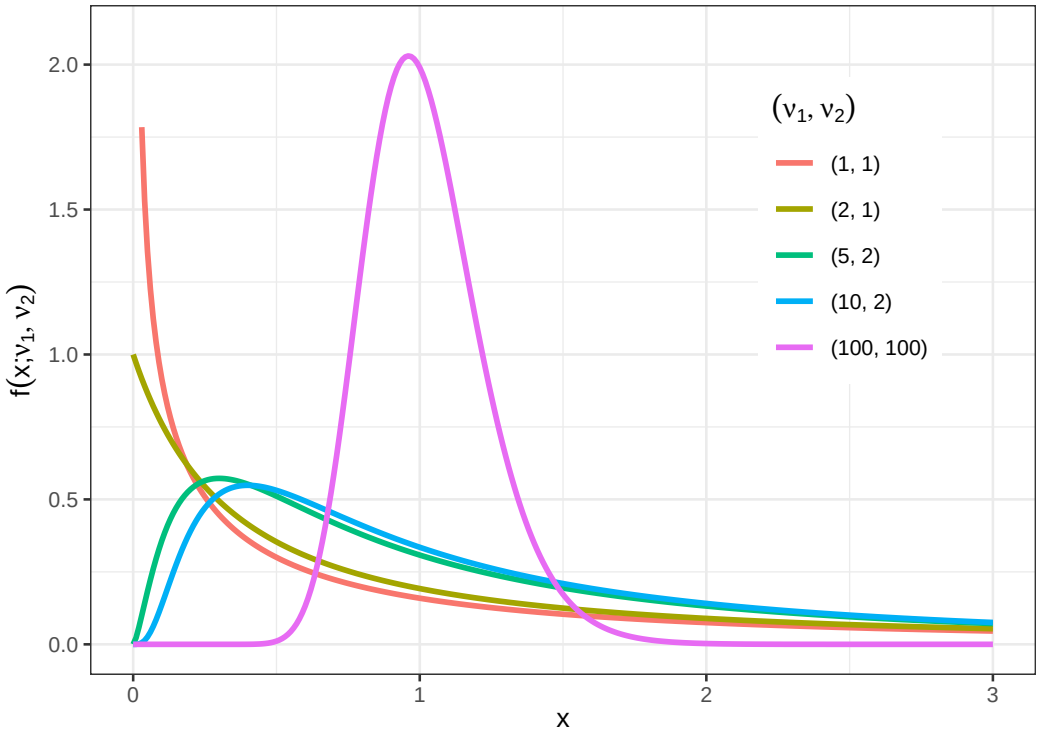


Figure 2.11: The density for $\mathsf{F}(\nu_1, \nu_2)$ for several combinations of (ν_1, ν_2).

Where do the terms numerator and denominator df come from?

The F distribution is related to ratios of χ^2 rvs, as captured in Theorem 2.1.

Theorem 2.1 (Ratio of χ^2 rvs)**.** *If $X_1 \sim \chi^2(\nu_1)$ and $X_2 \sim \chi^2(\nu_2)$ are independent rvs, then the rv*

$$F = \frac{X_1/\nu_1}{X_2/\nu_2} \quad \sim \mathsf{F}(\nu_1, \nu_2)\,,$$

that comprises the ratio of two χ^2 rvs divided by their respective df has an $\mathsf{F}(\nu_1, \nu_2)$ distribution.

3 Basics of statistical inference

We discuss point estimation, confidence intervals, and hypothesis testing in Sections Section 3.1, Section 3.2, and Section 3.3, respectively. These three tools will form the basis for making inferences about a population.

3.1 Point estimation

Statistical inference seeks to draw conclusions about the characteristics of a population from data. For example, suppose we are botanists interested in the taxonomic classification of iris flowers. Let μ denote the true average petal length (in cm) of the *Iris setosa*[1] (AKA the bristle-pointed iris). The parameter μ is a characteristic of the whole population of the *setosa* species. Before we collect data, the petal lengths of m independent *setosa* flowers are denoted by rvs $X_1, X_2, \ldots, X_m$. Any function of the X_i's, such as the sample mean,

$$\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i\,, \tag{3.1}$$

or the sample variance,

$$S^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \overline{X})^2\,, \tag{3.2}$$

is also a rv.

Suppose we actually find and measure the petal length of 50 independent *setosa* flowers resulting in observations $x_1, x_2, \ldots, x_{50}$; the distribution (counts) of 50 such petal length measurements are displayed in Figure 3.1. The sample mean $\overline{x}$ for petal length can then be used to draw a conclusion about the (true) value of the population mean μ. Based on the data in Figure 3.1 and using Equation 3.1, the value of the sample mean is $\overline{x} = 1.462$. The value $\overline{x}$ provides a "best guess" or point estimate for the true value of μ based on the $m = 50$ samples.

> **Loading datasets**
>
> The `datasets` package has a variety of datasets that you can play with. Once installed, data sets can be accessed in `R` by loading `library(datasets)` and then calling, e.g., `data(iris)` to see the `iris` data set. For a full list of available data sets, call `library(help = "datasets")` from the console.

> **Note 1: Iris Data**
>
> The botanist Edgar Anderson's **Iris Data** contains 50 obs. of four features (sepal length [cm], sepal width [cm], petal length [cm], and petal width [cm]) for each of three plant species (*setosa*, *virginica*, *versicolor*) for 150 obs. total.

[1] More about the *Iris setosa* here [https://www.wikiwand.com/en/Iris_setosa].

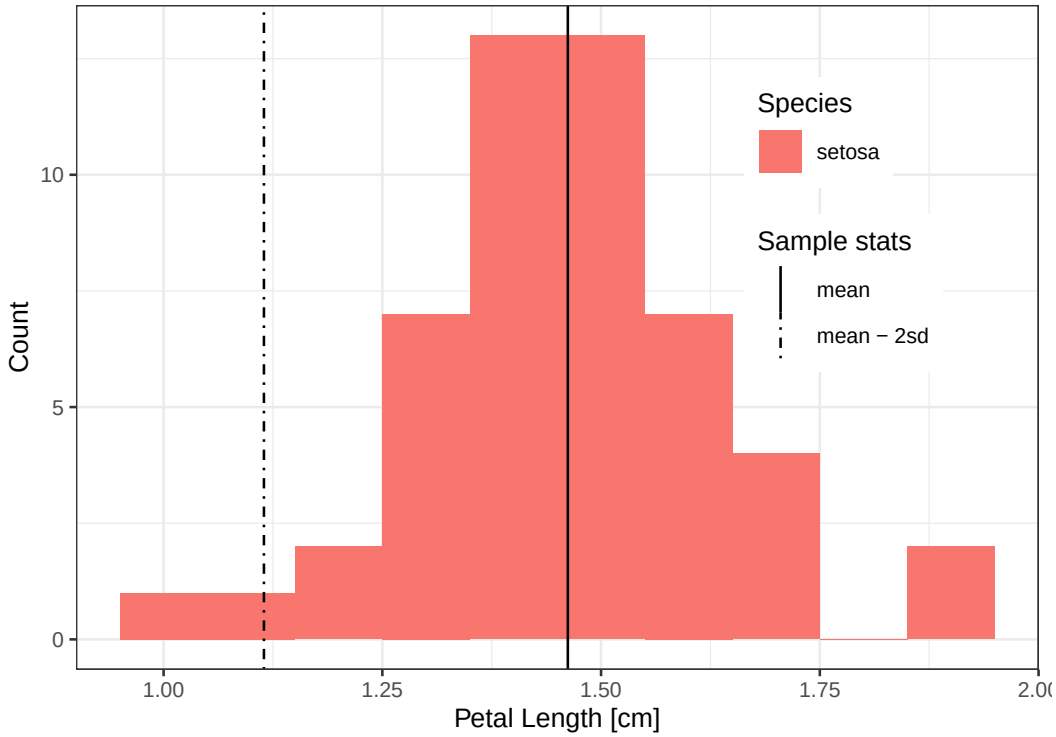


Figure 3.1: The distribution (counts) of $m = 50$ *setosa* petal length measurements.

```
iris |> glimpse()
```

```
Rows: 150
Columns: 5
$ Sepal.Length <dbl> 5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9, 5.4, 4.8, 4.8, 4.~
$ Sepal.Width  <dbl> 3.5, 3.0, 3.2, 3.1, 3.6, 3.9, 3.4, 3.4, 2.9, 3.1, 3.7, 3.4, 3.0, 3.~
$ Petal.Length <dbl> 1.4, 1.4, 1.3, 1.5, 1.4, 1.7, 1.4, 1.5, 1.4, 1.5, 1.5, 1.6, 1.4, 1.~
$ Petal.Width  <dbl> 0.2, 0.2, 0.2, 0.2, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.2, 0.2, 0.1, 0.~
$ Species    <fct> setosa, setosa, setosa, setosa, setosa, setosa, setosa, setosa, set~
```

Definition 3.1 (Point estimate)**.** A point estimate of a parameter θ (recall: a parameter is a fixed, unknown quantity) is a single number that we consider a reasonable value for θ. Consider

$$\text{iid } X_1, X_2, \ldots, X_m \sim F(\theta)\,.$$

A point estimator $\hat{\theta}_m$ of θ is obtained by selecting a suitable statistic g,

$$\hat{\theta}_m = g(X_1, \ldots, X_m)\,.$$

A point estimate $\hat{\theta}_m$ can then be computed from the estimator using sample data.

⚠ Overloaded notation

The symbol $\hat{\theta}_m$ (or simply $\hat{\theta}$ when the sample size m is clear from context) is typically used to denote both the estimator and the point estimate resulting from a given sample.

Table 3.1: Observations of $m = 31$ felled black cherry trees.

Height [in]
63, 64, 65, 66, 69, 70, 71, 72, 72, 74, 74, 75, 75, 75, 76, 76, 77, 78, 79, 80, 80, 80, 80, 80, 81, 81, 82, 83, 85, 86, 87

Best practice for reporting

Writing, e.g., $\hat{\theta} = 42$ does not indicate how the point estimate was obtained. Therefore, it is essential to report both the estimator and the resulting point estimate.

Definition 3.1 does not say how to select an appropriate statistic. For the *setosa* example, the sample mean $\overline{X}$ is suggested as a good estimator of the population mean μ. That is, $\hat{\mu} = \overline{X}$ or:

> "the point estimator of μ is the sample mean $\overline{X}$".

Here, while μ and σ^2 are fixed quantities representing population characteristics, $\overline{X}$ and S^2 are rvs with sampling distributions. If the population is *normally distributed* or if the *sample is large* then the sampling distribution for $\overline{X}$ has a known form:

$$\overline{X} \sim \mathsf{N}(\mu, \sigma^2/m)\,,$$

that is, $\overline{X}$ is normal with mean $\mu_{\overline{X}} = \mu$ and variance $\sigma^2_{\overline{X}} = \sigma^2/m$ where m is the sample size and μ and σ are the (typically unknown) population parameters.

Note 2: Cherry Tree Data

The **Cherry Tree Data** contains 31 obs. of three features (diameter, height, and volume).

```
trees |> glimpse()
```

```
Rows: 31
Columns: 3
$ Girth  <dbl> 8.3, 8.6, 8.8, 10.5, 10.7, 10.8, 11.0, 11.0, 11.1, 11.2, 11.3, 11.4, 11.4~
$ Height <dbl> 70, 65, 63, 72, 81, 83, 66, 75, 80, 75, 79, 76, 76, 69, 75, 74, 85, 86, 7~
$ Volume <dbl> 10.3, 10.3, 10.2, 16.4, 18.8, 19.7, 15.6, 18.2, 22.6, 19.9, 24.2, 21.0, 2~
```

Example 3.1. Let us consider the heights (measured in inches) of 31 black cherry trees (sorted, for your enjoyment) in Table 3.1.

The quantile-quantile plot in Figure 3.2, which compares the quantiles of this data to the quantiles of a normal distribution, is fairly straight. Therefore, we assume that the distribution of black cherry tree heights is (at least approximately) normal with a mean value μ; i.e., that the population of heights is distributed $\mathsf{N}(\mu, \sigma^2)$, where μ is a parameter to be estimated and σ^2 is unknown. The observations $X_1, \ldots, X_{31}$ are then assumed to be a random sample from this normal distribution,

$$\text{iid} \quad X_1, \ldots, X_{31} \sim \mathsf{N}(\mu, \sigma^2)\,.$$

Consider the following three different estimators and the resulting point estimates for μ based on the 31 samples in Table 3.1.

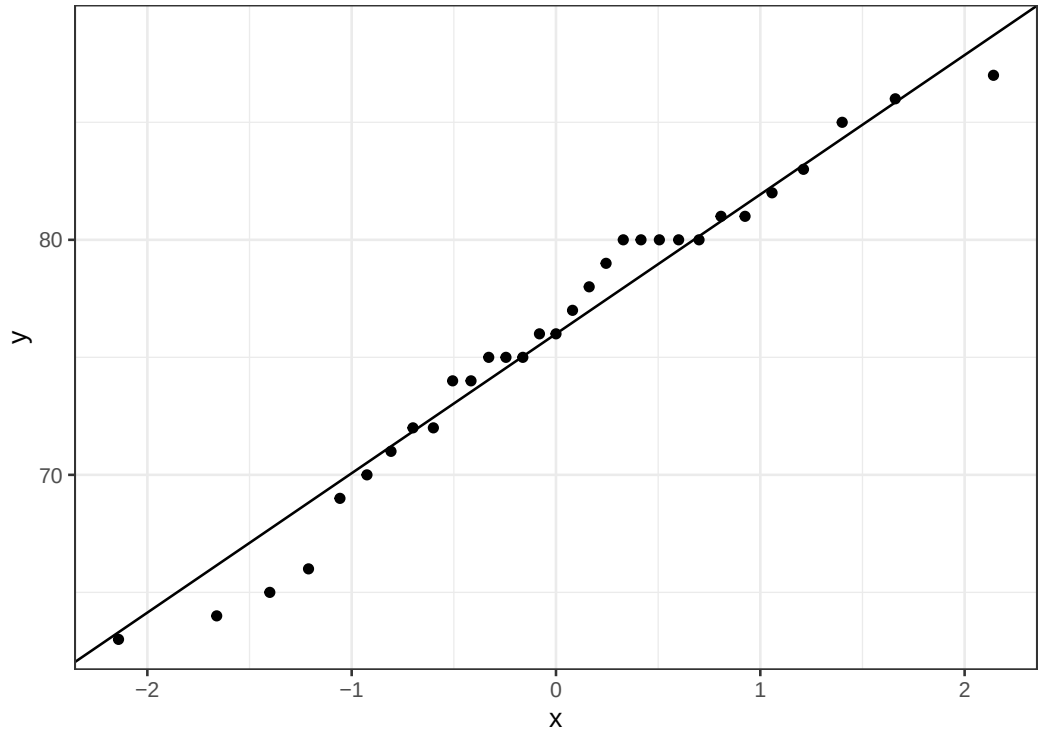


Figure 3.2: Normal quantile-quantile plot for the `Height` variable (feature) in the Cherry Tree Data.

a. Estimator (sample mean) $\overline{X}$ as in Equation 3.1 and estimate $\overline{x} = \sum x_i/n = 2356/31 = 76$.

b. Estimator (average of extreme heights) $\widetilde{X} = [\min(X_i) + \max(X_i)]/2$ and estimate $\tilde{x} = (63 + 87)/2 = 75$.

c. Estimator (10% trimmed mean – i.e., in this instance exclude the smallest and largest three values) $\overline{X}_{\text{tr(10)}}$ and estimate $\overline{x}_{\text{tr(10)}} = (2356 - 63 - 64 - 65 - 87 - 86 - 85)/25 = 76.24$.

Each estimator above uses a different notion of "centre" for the sample data, i.e., represents a different statistic. An interesting question is: which estimator will tend to produce estimates closest to the true parameter value? Will the estimators work universally well for all distributions?

How do we tell whether a population is normal?

Constructing a normal quantile-quantile plot (or QQ plot) is one way of assessing whether a normality assumption is reasonable. A QQ plot compares the quantiles of the sample data x_i against the theoretical standard normal quantiles, see Figure 3.2. If the sample data is consistent with a sample from a normal distribution, the points will lie in a straight line (more or less). The QQ plot in Figure 3.2 compares quantiles of cherry tree heights from Table 3.1 to normal quantiles. It is produced using the following code.

```
trees |> ggplot(aes(sample = Height)) + stat_qq() + stat_qq_line()
```

The data `trees` is piped to the command `ggplot`. For a QQ plot the key aesthetic element is `sample`; in this particular instance we set this to `Height`. The geometry `stat_qq()` adds the data quantiles plotted versus the normal quantiles. The geometry `stat_qq_line()` simply adds the fit line.

Example 3.2. Although probably overkill for this problem, the `infer` package can be used for point estimation using the `specify` and `calculate` commands as follows:

```
trees |>
 specify(response = Height) |>
 calculate(stat = "mean")
```

```
Response: Height (numeric)
# A tibble: 1 x 1
   stat
  <dbl>
1    76
```

The `response` option specifies the variable of interest, and the `stat` option can be changed to several quantities of interest.

In addition to reporting a point estimate and its estimator, some indication of its precision should be given. One measure of the precision of an estimate is its standard error.

Definition 3.2 (Standard error)**.** The standard error of an estimator $\hat{\theta}$ is the standard deviation

$$\sigma_{\hat{\theta}} = \sqrt{\mathrm{Var}(\hat{\theta})}\,.$$

Often, the standard error depends on unknown parameters and must also be estimated. The estimated standard error is denoted by $\widehat{\sigma}_{\hat{\theta}}$ or simply $s_{\hat{\theta}}$.

> Alternative notation
>
> The standard error is sometimes denoted $\mathsf{se} = \mathsf{se}(\hat{\theta})$ and the estimated standard error by $\widehat{\mathsf{se}}$.

3.2 Confidence intervals

An alternative to reporting a point estimate for a parameter is to report an interval estimate suggesting an entire range of plausible values for the parameter of interest. A confidence interval is an estimate that makes a probability statement about the interval's degree of reliability. The first step in computing a confidence interval is to select the confidence level α. A popular choice is a 95% confidence interval which corresponds to level $\alpha = 0.05$.

Definition 3.3 (Confidence interval)**.** A $100(1-\alpha)\%$ confidence interval for a parameter θ is a *random* interval

$$C_m = (L_m, U_m)\,,$$

where $L_m = \ell(X_1, \ldots, X_m)$ and $U_m = u(X_1, \ldots, X_m)$ are functions of the data, such that

$$P_\theta(L_m < \theta < U_m) = 1 - \alpha\,,$$

for all $\theta \in \Theta$.

My favourite interpretation of a confidence interval is due to (Wasserman 2004, p 92):

> *On day 1, you collect data and construct a 95 percent confidence interval for a parameter θ_1. On day 2, you collect new data and construct a 95 percent confidence interval for an unrelated parameter θ_2. On day 3, you collect new data and construct a 95 percent confidence interval for an unrelated parameter θ_3. You continue this way constructing confidence intervals for a sequence of unrelated parameters $\theta_1, \theta_2, \ldots$ Then 95 percent of your intervals will trap the true parameter value. There is no need to introduce the idea of repeating the same experiment over and over.*

This interpretation clarifies that a confidence interval is not a probability statement about the parameter θ. In Definition 3.3, note that θ is fixed (θ is not a rv) and the interval C_m is random. After data has been collected and a point estimator has been calculated, the resulting CIs either contain the true parameter value or do not, as illustrated in Figure 3.3.

3.3 Hypothesis testing

Section 3.1 and Section 3.2 reviewed how to estimate a parameter by a single number (point estimate) or range of plausible values (confidence interval), respectively. Next, we discuss methods for determining which of two contradictory claims, or hypotheses, about a parameter is correct.

Definition 3.4 (Null and alternative)**.** The null hypothesis, denoted by H_0, is a claim we initially assume to be true by default. The alternative hypothesis, denoted by H_a, is an assertion contradictory to H_0.

Typically, we shall consider a hypothesis test concerning a parameter $\theta \in \Theta$, i.e., taking values in a parameter space Θ. The statistical hypotheses are contradictory in that H_0 and H_a divide Θ into two disjoint sets. For example, for a statistical inference regarding the *equality* of a parameter θ with a fixed quantity θ_0, the null and alternative hypotheses will usually take one of the following forms in Table 3.2.

Table 3.2: Typical null hypothesis and corresponding alternative hypothesis.

Null hypothesis	Alternative hypothesis	Test form
$H_0 : \theta = \theta_0$	$H_a : \theta \neq \theta_0$	two-sided test
$H_0 : \theta \leq \theta_0$	$H_a : \theta > \theta_0$	one-sided test
$H_0 : \theta \geq \theta_0$	$H_a : \theta < \theta_0$	one-sided test

These hypothesis pairs are associated with either a one-sided or two-sided test; what this means will become apparent in the sequel. The value θ_0, called the null value, separates the alternative from the null.

Definition 3.5 (Hypothesis test)**.** A hypothesis test asks if the available data provides sufficient evidence to reject H_0. If the observations disagree with H_0, we reject the null hypothesis. If the sample evidence does not strongly contradict H_0, then we continue to believe H_0. The two possible conclusions of a hypothesis test are: *reject* H_0 or *fail to reject* H_0.

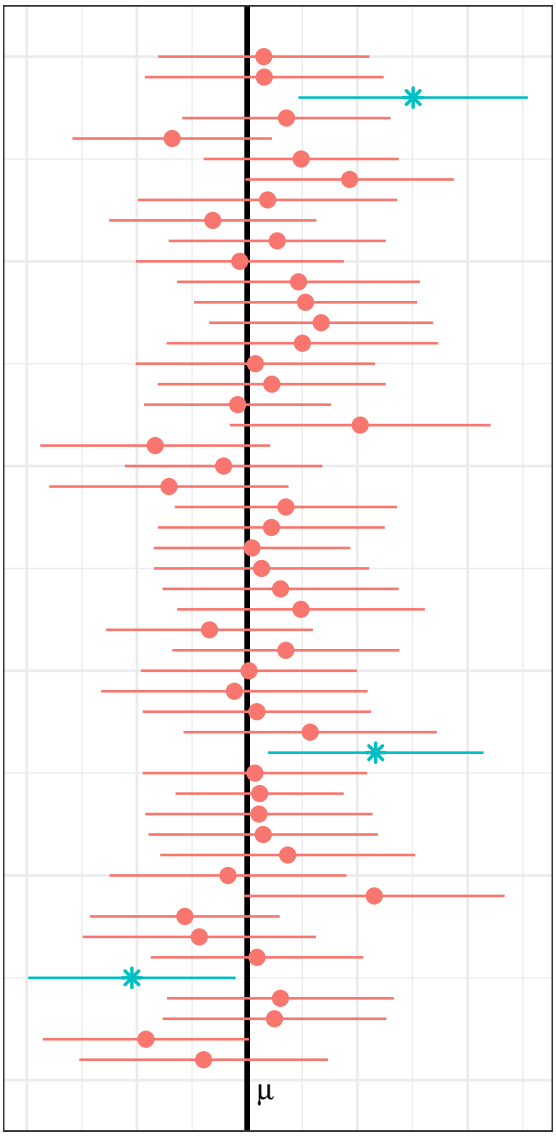


Figure 3.3: Fifty 95% CIs for a population mean μ. After a sample is taken, the computed interval estimate either contains μ or does not (asterisks identify intervals that do not include μ). When drawing such a large number of 95% CIs, we would anticipate that approximately 5% (ca. 2 or 3) would fail to cover the true parameter μ.

> ❗ "Fail to reject" versus "accept"
>
> We comment that *fail to reject* H_0 is sometimes phrased as *retain* H_0 or (perhaps less accurately) *accept* H_0.
> Why not just *accept* the null and move on with our lives?
> Well, if I search the Highlands for the Scottish wildcat (endangered) and fail to find any, does that prove they do not exist?

A procedure for carrying out a hypothesis test is based on specifying two additional items: a test statistic and a corresponding rejection region. A test statistic T is a function of the sample data (like an estimator). The decision to reject or fail to reject H_0 will involve computing the test statistic. The rejection region R is the collection of values of the test statistic for which H_0 is to be rejected in favour of the alternative, e.g.,

$$R = \{x : T(x) > c\} ,$$

where c is referred to as a critical value. If a given sample falls in the rejection region, we reject H_0. If $X \in R$ (e.g., the calculated test statistic exceeds some critical value), we reject H_0. The alternative is that $X \notin R$ and we fail to reject the null in this case.

Two types of errors can be made when carrying out a hypothesis test. The basis for choosing a rejection region involves considering these errors.

Definition 3.6 (Error types)**.** A type I error occurs if H_0 is rejected when H_0 is actually true. A type II error is made if we fail to reject H_0 when H_0 is actually false.

If a test's maximal type I error is fixed at an acceptably small value, then the type II error decreases as the sample size increases. In particular, a conclusion is reached in a hypothesis test by selecting a significance level α for the test linked to the maximal type I error rate. Typically, $\alpha = 0.10, 0.05, 0.01$, or 0.001 is selected for the significance level.

Definition 3.7 (P-value)**.** A P-value is the probability, calculated assuming H_0 is true, of obtaining a value of the test statistic at least as contradictory to H_0 as the value calculated from the sample data.

Smaller P-values indicate stronger evidence against H_0 in favor of H_a. If $P \leq \alpha$ then we reject H_0 at significance level α. If $P \geq \alpha$ we fail to reject H_0 at significance level α.

> ⚠ What a P-value isn't…
>
> The P-value is a probability calculated assuming that H_0 is true. However, the P-value is **not** the probability that:
>
> 1. H_0 is TRUE,
> 2. H_0 is FALSE, or
> 3. a wrong conclusion is reached.

Proposition 3.1. *The hypothesis test procedure that*

$$\begin{cases} \textit{rejects } H_0 & \textit{if } P \leq \alpha, \\ \textit{fails to reject } H_0 & \textit{otherwise}, \end{cases}$$

has $P(\textit{type I error}) = \alpha$.

Example 3.3. Churchill claims that he will receive half the votes for the House of Commons seat for the constituency of Dundee.[2] If we do not believe Churchill's claim and are doubtful of his popularity, we would seek to test an alternative hypothesis. How should we write down our research hypotheses?

If we let p be the fraction of the population voting for Churchill, then we have the null hypothesis,

$$H_0 : p = 0.5\,,$$

and the alternative hypothesis (we believe Churchill is less popular than he claims),

$$H_a : p < 0.5\,.$$

Support for the alternative hypothesis is obtained by showing a lack of support for its converse hypothesis (the null hypothesis).

Example 3.4. Suppose that $m = 15$ voters are selected from Dundee and X, the number favouring Churchill, is recorded. Based on observing X, we construct a rejection region $R = \{x : x \leq k\}$. If k is small compared to m, then the rejection region would provide strong evidence to reject H_0. How should one choose the rejection region?

Assume now that $m = 15$ voters are polled and that we select $k = 2$ to have a rejection region $R = \{x \leq 2\}$. For this choice of k, the rejection region R provides strong support to reject H_0. Assuming the null hypothesis is true, we expect approximately half of the 15 voters (ca. 7) to vote for Churchill. Observing $x = 0$, $x = 1$ or $x = 2$ (the values that would place us in the rejection region) would provide strong evidence *against* H_0.

We can calculate the probability of a type I error. From the definition of type I error,

$$\begin{aligned}\alpha &= P(\text{type I error})\\ &= P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true})\\ &= P(X \in R \text{ when } H_0 \text{ is true})\\ &= P(X \leq 2 \text{ when } p = 0.5)\,.\end{aligned}$$

Since $X \sim \mathsf{Binom}(15, 0.50)$, we calculate that $\alpha = 0.00369$. Thus, for this particular choice of rejection region R, the risk of concluding that Churchill will lose if, in fact, he is the winner is tiny.

For this rejection region, how good is the test at protecting us from type II errors, i.e., concluding that Churchill is the winner if, in fact, he will lose? Suppose that Churchill receives 25 of the votes ($p = 0.25$). The probability of type II error β is,

$$\begin{aligned}\beta &= P(\text{type II error})\\ &= P(\text{fail to reject } H_0 \text{ when } H_0 \text{ false})\\ &= P(X \notin R \text{ when } H_0 \text{ false})\\ &= P(X > 2 \text{ when } p = 0.3)\,.\end{aligned}$$

For $X \sim \mathsf{Binom}(15, 0.25)$, we calculate $\beta = 0.764$. If we use $R = \{x \leq 2\}$, then our test will lead us to conclude that Churchill is the winner with a probability of 0.764 even if p is as low as 0.25!

If we repeat these calculations for $R^* = \{x \leq 5\}$, we find $\alpha = 0.151$ versus $\beta = 0.148$, even if p is as low as 0.25, which is a much better balance between type I and type II errors.

[2] Sir Winston Churchill was Member of Parliament for Dundee from 1908–1922 [https://www.wikiwand.com/en/Winston_Churchill].

⚠ What if the sample size is close to the population size?

In Example 3.4, X is a binomial random variable because it can be modelled as m independent Bernoulli trials each with probability p of success (i.e., votes for Churchill) as long as the sample size m is much smaller than the population of Dundee. If we had the means to canvas nearly the whole population, what goes wrong conceptually?

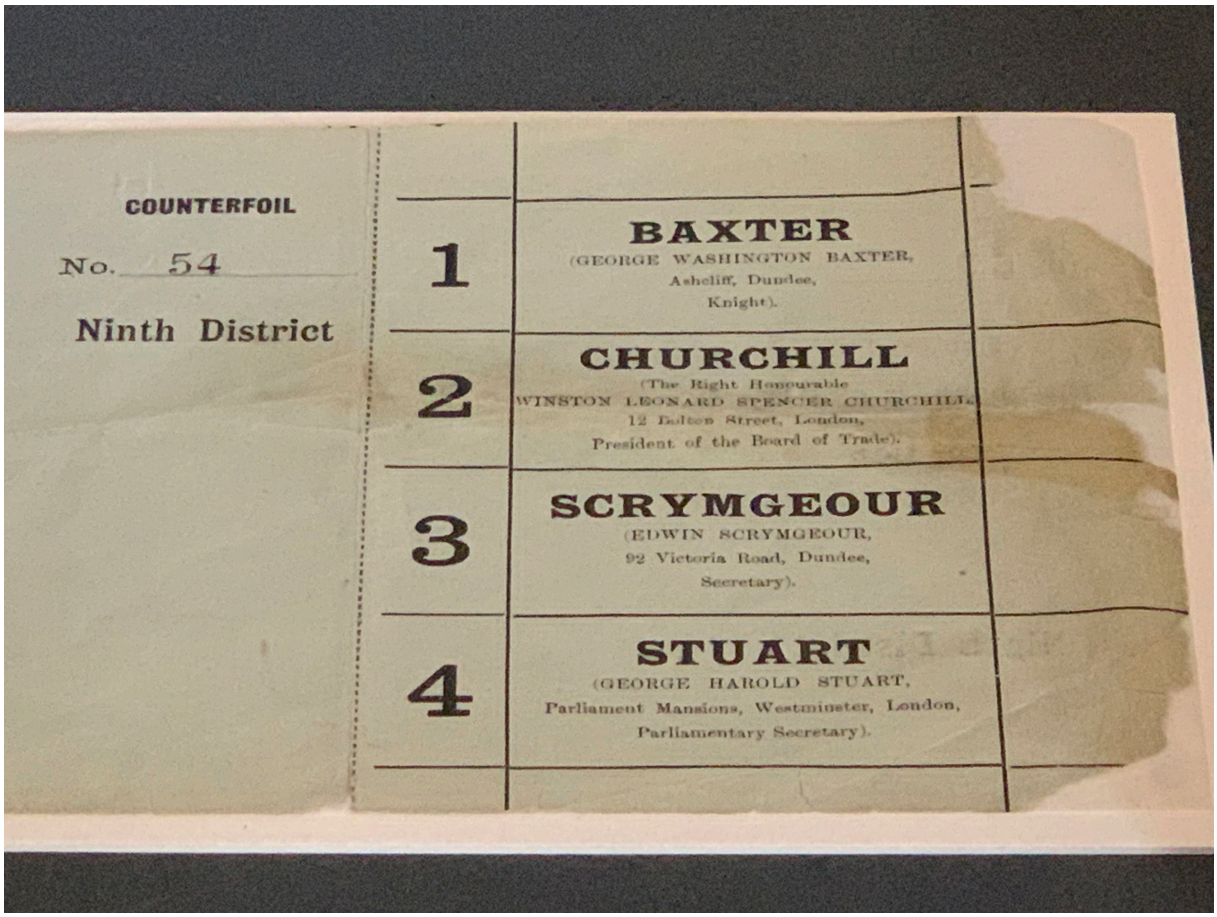


Figure 3.4: Ballot listing Churchill from the collection of the McManus, Dundee. When you take a break from studying, go and see if you can find it! For more information on visiting the McManus visit https://www.mcmanus.co.uk/.

! Elements of a statistical test

A statistical test is based on a null hypothesis (H_0) and an alternative hypothesis (H_a).
An appropriate test statistic T is computed. Then either:

- T is compared to a rejection region (based on significance level α)

OR

- P-value (based on T) is compared to the significance level α.

4 Single sample inferences

In a few situations, we can derive the sampling distribution for the statistic of interest and use this as the basis for constructing confidence intervals and hypothesis tests. Presently we estimate population means μ in Section 4.1, population proportions p in Section 4.2.1, and population variances σ^2 in Section 4.3 in some special cases.

4.1 Estimating means

If the parameter of interest is the population mean $\theta = \mu$, then what can be said about the distribution of the sample mean estimator $\hat{\theta} = \overline{X}$ in Equation 3.1? We will consider three cases,

1. normal population with known σ^2,
2. any population with unknown σ^2, when the sample size m is large, and
3. normal population with unknown σ^2, when the sample size m is small.

In each, the form of the confidence interval and hypothesis test statistic for μ can be derived using the approximate normality of the sample mean.

In general, the confidence intervals for the mean based on normality theory will have the form:

$$\text{point estimate } \mu \pm (\text{critical value}) \cdot (\text{precision of point estimate}), \tag{4.1}$$

where the reference distribution will be the standard normal (for 1. and 2.) and the Student's t distribution (for 3.). The critical value corresponds to the value under the reference distribution that yields the two-sided (symmetric) tail areas summing to $1 - \alpha$.

4.1.1 Mean of a normal population with known variance

When sampling from a normal population with a known mean and variance, the estimator for the sample mean is also normal with mean μ and variance σ^2/m where m is the sample size. Standardising,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{m}} \quad \sim \mathsf{N}(0, 1) \tag{4.2}$$

we see that

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{m}} < z_{\alpha/2}\right) = 1 - \alpha\,.$$

Based on knowing the estimator's sampling distribution, we state the following CI.

Definition 4.1 (Confidence interval for mean of normal population)**.** A $100(1-\alpha)\%$ confidence interval for the mean μ of a normal population when the value of σ^2 is known is given by

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{m}}, \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{m}}\right), \tag{4.3}$$

or $\overline{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{m}$, where m is the sample size.

The CI for the mean Equation 4.3 can be expressed (cf. Equation 4.1) as

$$\text{point estimate } \mu \pm (z \text{ critical value}) \cdot (\text{standard error of mean}).$$

The z critical value is related to the tail areas under the standard normal curve; we need to find the z-score having a cumulative probability equal to $1-\alpha$ according to Definition @ref(def:confidence-interval-gen).

Example 4.1. Consider 400 samples from a normal population with a known standard deviation $\sigma = 17000$ with mean $\overline{x} = 20992$ as depicted in Figure 4.1. How do we construct a 95% confidence interval for μ?

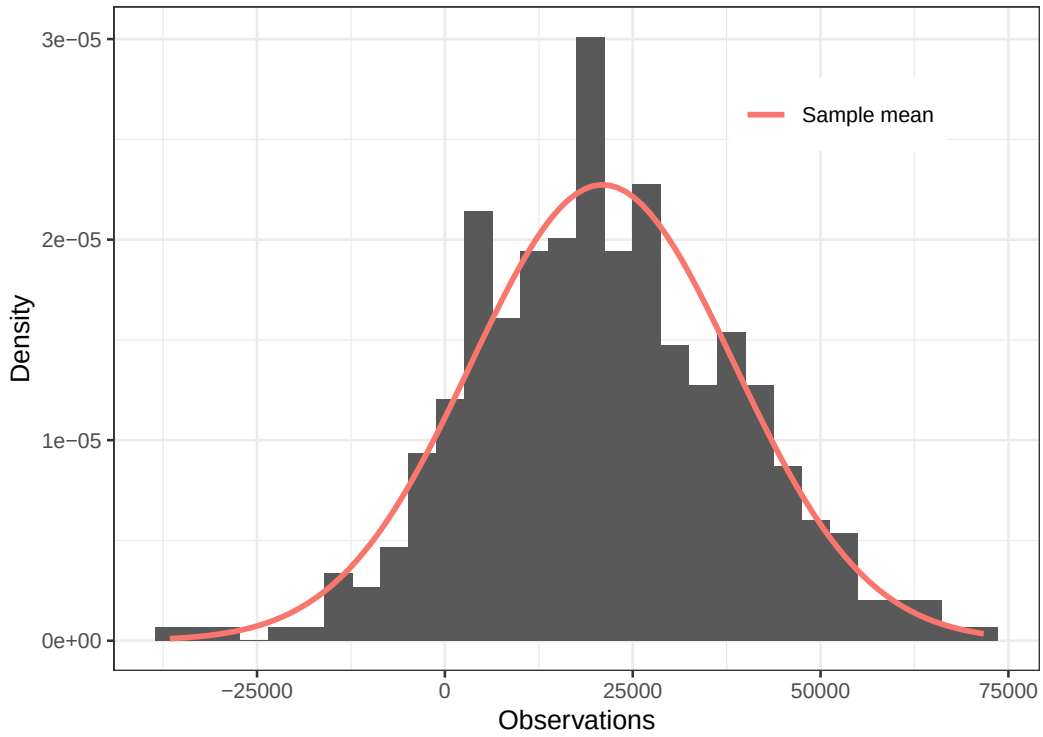


Figure 4.1: 400 samples from a normal population with known variance $\sigma = 17000$ together with the corresponding (normal) sampling distribution for the observed mean.

For $\alpha = 0.05$, the critical value $z_{0.025} = 1.96$; this value can be found by looking in a table of critical z values or using the `R` code `qnorm(1-.05/2)`. From Definition 4.1,

$$\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{m}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{m}}\right) = \left(20992 - 1.96\frac{17000}{\sqrt{400}}, 20992 + 1.96\frac{17000}{\sqrt{400}}\right)$$
$$= (19326, 22658).$$

The data above was generated with a true population parameter $\mu = 21500$, and the CI contains the parameter value (incidentally).

As noted in Equation 4.1 and Equation 4.3, the width of a CI is related to the estimator's precision. The confidence level (or reliability) is inversely related to this precision. When the population is normal and the variance is known, determining the sample size necessary to achieve a desired confidence level and precision is an appealing strategy. A general formula for the sample size m^* necessary to achieve an interval width w is obtained at confidence level α by equating

$$w = 2z_{\alpha/2} \cdot \sigma/\sqrt{m^*}$$

and then solving for m^*.

Proposition 4.1. *The sample size m required to achieve a CI for μ with width w at level α is given by,*

$$m^* = \left(2z_{\alpha/2} \cdot \frac{\sigma}{w}\right)^2 .$$

From Proposition 4.1, we see that the smaller the desired w, the larger m^* must be (and subsequently, the more effort that must be allocated to data collection).

Example 4.2. In Example 4.1 we identified a 95% confidence interval for a normal population with known variance. The range (width) of that interval was $22658 - 19326 = 3332$. How much would m need to increase to halve the interval width?

Using Proposition 4.1,

$$m = \left(2 \cdot 1.96 \cdot \frac{17000}{1666}\right)^2 = (40)^2 = 1600 .$$

Thus, we find that for the same level $\alpha = 0.05$, we would need to quadruple our original sample size to halve the interval.

> ⚠ You heard it here first…
>
> As Example 4.2 shows, it is expensive to reduce uncertainty!

Suppose now that we would like to consider a hypothesis test for the population mean, such as $H_0 : \mu = \mu_0$. Starting from Equation 4.2 and assuming that the null hypothesis is true, we find

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{m}} .$$

The statistic Z measures the distance (measured in units of $\text{sd}[\overline{X}]$) between $\overline{X}$ and its expected value under the null hypothesis. We will use the statistic Z to determine if there is substantial evidence against H_0, i.e. if the distance is too far in a direction consistent with H_a.

Proposition 4.2. *Assume that we sample $X_1, \ldots, X_m$ from a normal population with mean μ and known variance σ^2.*

Consider $H_0 : \mu = \mu_0$. The test statistic is

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{m}} . \tag{4.4}$$

For a hypothesis test at level α, we use the following procedure:

If $H_a : \mu > \mu_0$, *then* $P = 1 - \Phi(z)$, *i.e., upper-tail* $R = \{z > z_\alpha\}$.

If $H_a : \mu < \mu_0$, *then* $P = \Phi(z)$, *i.e., lower-tail* $R = \{z < -z_\alpha\}$.

If $H_a : \mu \neq \mu_0$, *then* $P = 2(1 - \Phi(|z|))$, *i.e., two-tailed* $R = \{|z| > z_{\alpha/2}\}$.

We recall that $\Phi(z)$ is the area in the lower tail of the standard normal density, i.e., to the *left* of the calculated value of z. Thus $1 - \Phi(z)$ is the area in the upper-tail, and $2(1 - \Phi(|z|))$ is twice the area captured in the upper-tail by $|z|$, i.e. the sum of the area in the tails corresponding to $\pm z$. If $P < \alpha$, then we reject H_0 at level α as the data provides sufficient evidence at the α level against the null hypothesis.

Example 4.3. Let's return to the data in Example 4.1, where we sample from a normal population with a known standard deviation $\sigma = 17000$. Suppose that someone claims the true mean is $\mu_0 = 20000$. Does our sample mean $\overline{x} = 20992$ based on $m = 400$ samples provide evidence to contradict this claim at the $\alpha = 0.05$ level?

The first thing to record is our parameter of interest: μ, the true population mean. The null hypothesis, which we assume to be true, is a statement about the value of μ,

$$H_0 : \mu = 20000\,,$$

and the alternative hypothesis is

$$H_a : \mu \neq 20000\,,$$

since we are concerned with a deviation in either direction from $\mu_0 = 20000$.

Since the population is normal with known variance, we compute the test statistic:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{m}} = \frac{20992 - 20000}{17000/\sqrt{400}} = 1.167\,.$$

That is, the observed sample mean $\overline{x}$ is slightly more than 1 standard deviation than what we expect under H_0. Consulting Proposition 4.2, we see that a two-tailed test is indicated for this particular H_a (i.e., containing "$\neq$"). The P-value is the area,

$$P = 2(1 - \Phi(1.167)) = 2(0.1216052) = 0.2432.$$

Thus, since $P = 0.2432 > 0.05 = \alpha$, we fail to reject H_0 at the level 0.05. The data does not support the claim that the true population mean differs from the value 20000 at the 0.05 level.

> **Recall**
>
> Note $\Phi(z) = P(Z \leq z)$ is found by calling `pnorm(z)` in `R` or by looking up the value in a Z table.

4.1.2 Mean of a population with unknown variance (large-sample)

Consider samples $X_1, \ldots, X_m$ from a population with mean μ and variance σ^2. Provided that m is large enough, the Central Limit Theorem implies that the estimator for the sample mean $\overline{X}$ in Equation 3.1 has *approximately* a normal distribution. Then

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{m}} < z_{\alpha/2}\right) \approx 1 - \alpha\,,$$

since the transformed variable has approximately a standard normal distribution. Thus, computing a point estimate based on a large m of samples yields a CI for the population parameter μ at an *approximate* confidence level α. However, it is often the case that the variance is unknown. When m is large, replacing the population variance σ^2 by the sample variance S^2 in Equation 3.2 will not typically introduce too much additional variability.

Proposition 4.3. *For a large sample size m, an approximate* $100(1-\alpha)\%$ *confidence interval for the mean* μ *of any population when the variance is unknown is given by*

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{m}}, \overline{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{m}}\right), \tag{4.5}$$

or $\overline{x} \pm z_{\alpha/2} \cdot s/\sqrt{m}$.

The CI for the mean Equation 4.5 applies regardless of the shape of the population distribution so long as the number of samples is large. A rule of thumb is that $m > 40$ is sufficient. In words, the CI Equation 4.5 can be expressed (cf. Equation 4.1) as

$$\text{point estimate } \mu \pm (z \text{ critical value}) \cdot (\text{estimated standard error of mean})\,.$$

Typically, a large-sample CI for a general parameter θ holds that is similar to Equation 4.5 for any estimator $\hat{\theta}$ that satisfies: (1) approximately normal in distribution, (2) approximately unbiased, and (3) an expression for the standard error is available.

To conduct a large-sample hypothesis test regarding the population mean μ, we consider the test statistic

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{m}}$$

under the null hypothesis, i.e., we replace the population standard deviation σ with the sample standard deviation S. When the number of samples m is large (say $m > 40$), then Z will be approximately normal. Substituting this test statistic Z for Equation 4.4, we follow Proposition 4.2 to determine how to calculate the P-value.

> **Rule of thumb**
>
> For estimating means, we consider a sample size of $m > 40$ to be large.
> However, 'large' depends on the context: for example, the level of support for the evidence that you are seeking. For $m > 20$, the interval estimate
>
> $$\text{point estimate } \pm 2\,\text{sd}$$
>
> has 95% coverage and is surprisingly robust, i.e. applies to a wide variety of population distributions

including the normal. However, this rule of thumb won't apply if you want to consider some different level, say 80% (Belle 2008, sec. 1).

Example 4.4. Let's consider the **Iris Data** from Note 1 and use the `infer` package to make inferences. In particular, consider whether there is evidence at the 0.05 level to support the statement that the true mean petal length of Iris flowers exceeds 3.5 cm.

Recall that the **Iris Data** contains $m = 150$ measurements of petal length across three species of Iris flowers and that the true variance is unknown. We are interested in testing the null hypothesis,

$$H_0 : \mu \leq 3.5\,,$$

against the alternative,

$$H_a : \mu > 3.5\,,$$

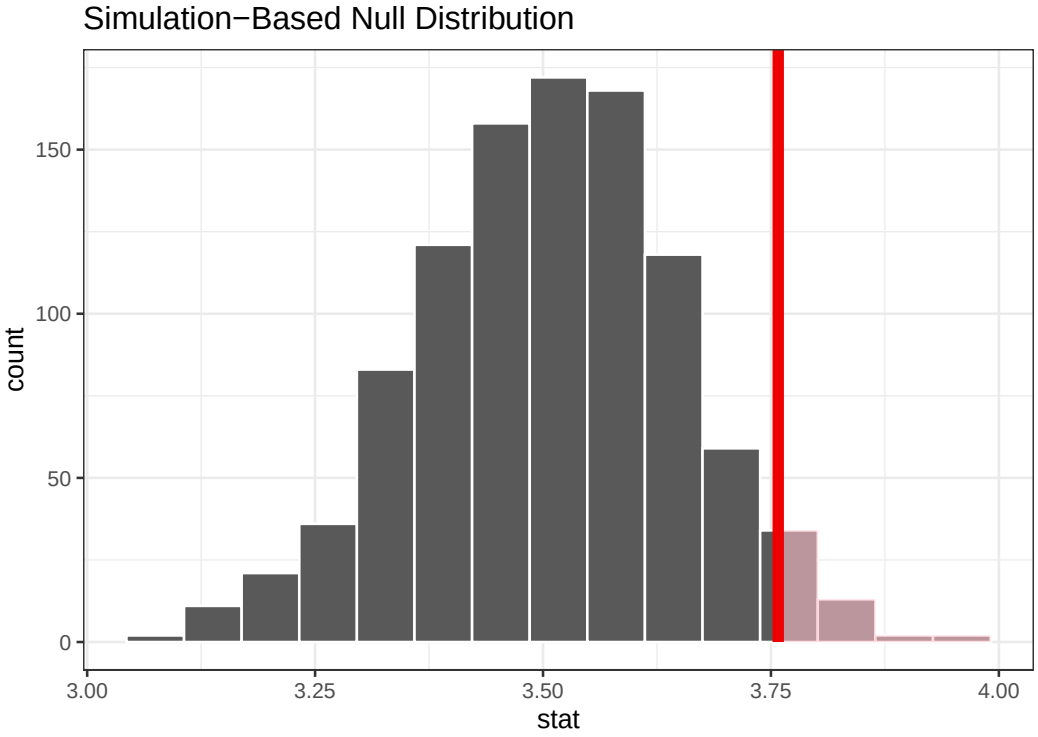
i.e., a one-sided test.

We first compute the observed statistic (sample mean) $\hat{\mu}$. We use the `infer` package to construct a null distribution *computationally* for the response variable (petal length). We specify that the hypothesis test is for the parameter based on a point estimate and that we are testing for equality with the value $\mu_0 = 3.5$. The null distribution is generated by computing 1000 bootstrap replications of the sample mean, i.e., the sample mean is generated 1000 times by drawing 150 values at random with replacement from the original corpus of $m = 150$ samples. (Note that we obtain the null distribution computationally, so we do not need to standardise to Z.)

```
mu_hat <- mean(iris$Petal.Length)

null_dist <- iris |>
 specify(response = Petal.Length) |>
 hypothesise(null = "point", mu = 3.5) |>
 generate(reps = 1000, type = "bootstrap") |>
 calculate(stat = "mean")

null_dist |>
 visualise() +
 shade_p_value(obs_stat = mu_hat, direction = "greater")
```



The bootstrapped null distribution is plotted using the `visualise` command, and the regions of the null distribution that are as extreme (or more extreme) than the observed statistic $\widehat{\mu}$ can be highlighted using the `shade_p_value` command.

```
p_val <- null_dist |>
 get_p_value(obs_stat = mu_hat, direction = "greater")
p_val
```

```
# A tibble: 1 x 1
  p_value
    <dbl>
1   0.037
```

The test yields a P-value of $P = 0.037$. This values is quite small; if $\mu \leq 3.5$, then the probability of obtaining the sample mean value $\widehat{\mu} = 3.758$ is only 0.037! Thus, the data provide sufficient evidence at the 0.05 level against the hypothesis that the true mean petal length is at most 3.5 cm.

4.1.3 Mean of a normal population with unknown variance

In Section 4.1.1, we considered samples $X_1, \ldots, X_m$ from a normal population with a known μ and σ^2. In contrast, here, we consider samples from a normal population and assume the population parameters μ and σ^2 are unknown. If the number of samples is large, the discussion in Section 4.1.2 indicates that the rv

$$Z = (\overline{X} - \mu)\sqrt{m}/S$$

has approximately a standard normal distribution. However, if m is not sufficiently large then the transformed variable will be more spread out than a standard normal distribution.

Theorem 4.1. *For the sample mean $\overline{X}$ based on m samples from a normal distribution with mean μ, the rv*

$$T = \frac{\overline{X} - \mu}{S/\sqrt{m}} \quad \sim \mathsf{t}(m-1)\,, \tag{4.6}$$

that is, T has Student's t *distribution with $\nu = m - 1$ df.*

This leads us to consider a CI for the population parameter μ based on critical values of the t distribution.

Proposition 4.4. *A $100(1-\alpha)\%$ confidence interval for the mean μ of a normal population, when σ^2 is unknown, is given by*

$$\left(\overline{x} - t_{\alpha/2,m-1} \cdot \frac{s}{\sqrt{m}}\,, \overline{x} + t_{\alpha/2,m-1} \cdot \frac{s}{\sqrt{m}}\right), \tag{4.7}$$

or $\overline{x} \pm t_{\alpha/2,m-1} \cdot s/\sqrt{m}$. Here $\overline{x}$ and s are the sample mean and sample standard deviation, respectively.

Example 4.5. Let us return to the height of 31 felled black cherry trees from the **Cherry Tree Data** in Note 2. Give a 99% CI for the population mean μ.

For $m = 31$, the critical value of the reference distribution is

$$t_{0.005,30} \approx 2.7499\,,$$

which can looked up in a table of critical values for $\mathsf{t}(\nu = 30)$ or found using the R command `qt(1-0.01/2, df = 31-1)`. The sample mean $\overline{x} = 76$ (computed in Example 3.1) is combined with the sample standard deviation,

$$\begin{aligned} s &= \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2} \\ &= \sqrt{\frac{1}{30}\left((63-76)^2 + \cdots + (87-76)^2\right)} \\ &= 6.372\,, \end{aligned}$$

to form the interval estimate

$$\begin{aligned} &\left(\overline{x} - t_{\alpha/2,m-1} \cdot \frac{s}{\sqrt{m}}\,, \overline{x} + t_{\alpha/2,m-1} \cdot \frac{s}{\sqrt{m}}\right) \\ &\quad = \left(76 - 2.750 \cdot \frac{6.372}{\sqrt{31}}\,, 76 + 2.750 \cdot \frac{6.372}{\sqrt{31}}\right) \\ &\quad = (72.85\,, 79.15)\,. \end{aligned}$$

For comparison, the critical value $t_{.01/2,\nu}$ for $\nu = 14, \ldots, 40$ can be recalled with the following command.

```
qt(1-0.01/2, df = seq(14,40))
```

```
 [1] 2.976843 2.946713 2.920782 2.898231 2.878440 2.860935 2.845340 2.831360 2.818756
[10] 2.807336 2.796940 2.787436 2.778715 2.770683 2.763262 2.756386 2.749996 2.744042
[19] 2.738481 2.733277 2.728394 2.723806 2.719485 2.715409 2.711558 2.707913 2.704459
```

Table 4.1: Observations of $m = 31$ felled black cherry trees.

Volume [cu ft]
10.2, 10.3, 10.3, 15.6, 16.4, 18.2, 18.8, 19.1, 19.7, 19.9, 21.0, 21.3, 21.4, 22.2, 22.6, 24.2, 24.9, 25.7, 27.4, 31.7, 33.8, 34.5, 36.3, 38.3, 42.6, 51.0, 51.5, 55.4, 55.7, 58.3, 77.0

Note that these critical values can deviate significantly from the corresponding $z_{0.01/2} = 2.575829$. In particular, if we had erroneously used the large sample estimate Equation 4.5, then we would have obtained a 99% CI (73.05 , 78.95) which might give us a false sense of security as it is narrower.

In contrast to Proposition 4.1, it is difficult to select the sample size m to control the width of the t-based CI as the width involves the unknown (before the sample is acquired) s and because m also enters through $t_{\alpha/2,m-1}$. A one-sample t test based on Equation 4.6 can be used to test a hypothesis about the population mean when the population is normal and σ^2 is unknown.

Proposition 4.5. *Assume that we sample $X_1, \dots, X_m$ from a normal population with mean μ and unknown variance σ^2.*

Consider $H_0 : \mu = \mu_0$. The test statistic is

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{m}}\,.$$

For a hypothesis test at level α, we use the following procedure:

If $H_a : \mu > \mu_0$, then P-value is the area under $\mathsf{t}(m-1)$ to the right of t.

If $H_a : \mu < \mu_0$, then P-value is the area under $\mathsf{t}(m-1)$ to the left of t.

If $H_a : \mu \neq \mu_0$, then P-value is twice the area under $\mathsf{t}(m-1)$ to the right of $|t|$.

Example 4.6. Let's consider the **Cherry Tree Data** in Note 2. The average timber volume is given in Table 4.1. The distribution for this data is approximately normal. We might ask if the data provide compelling evidence, say at level 0.05, for concluding that the true average timber volume exceeds 21.3 cubic feet.[1]

Let's carry out a significance test for the true average volume of timber μ at level $\alpha = 0.05$. We assume the null hypothesis

$$H_0 : \mu = 21.3\,.$$

An appropriate alternative hypothesis is

$$H_a : \mu > 21.3\,,$$

that is, we will adopt the stance that the true average exceeds $\mu_0 = 21.3$ only if the null is rejected.

[1] How much wood is that? About a sixth of a cord. A full cord of chopped firewood in the US is 124 cu ft; about enough to keep you warm through a New England winter (according to my mother-in-law).

From our $m = 31$ samples, we find that $\overline{x} = 30.17$ and that $s = 16.44$. The computed value of the one-sample t-statistic is given by

$$\begin{aligned} t &= \frac{\overline{x} - \mu_0}{s/\sqrt{m}} \\ &= \frac{30.17 - 21.3}{16.44/\sqrt{31}} \\ &= 3\,. \end{aligned}$$

The test is based on $\nu = 31 - 1$ df, and $P = 0.002663$. This is the upper-tail area, i.e., the area to the right of t in Figure 4.2. Since $P \ll \alpha$, we reject the null hypothesis that the population mean is 21.3. The data provide sufficient evidence that the population mean differs from 21.3.

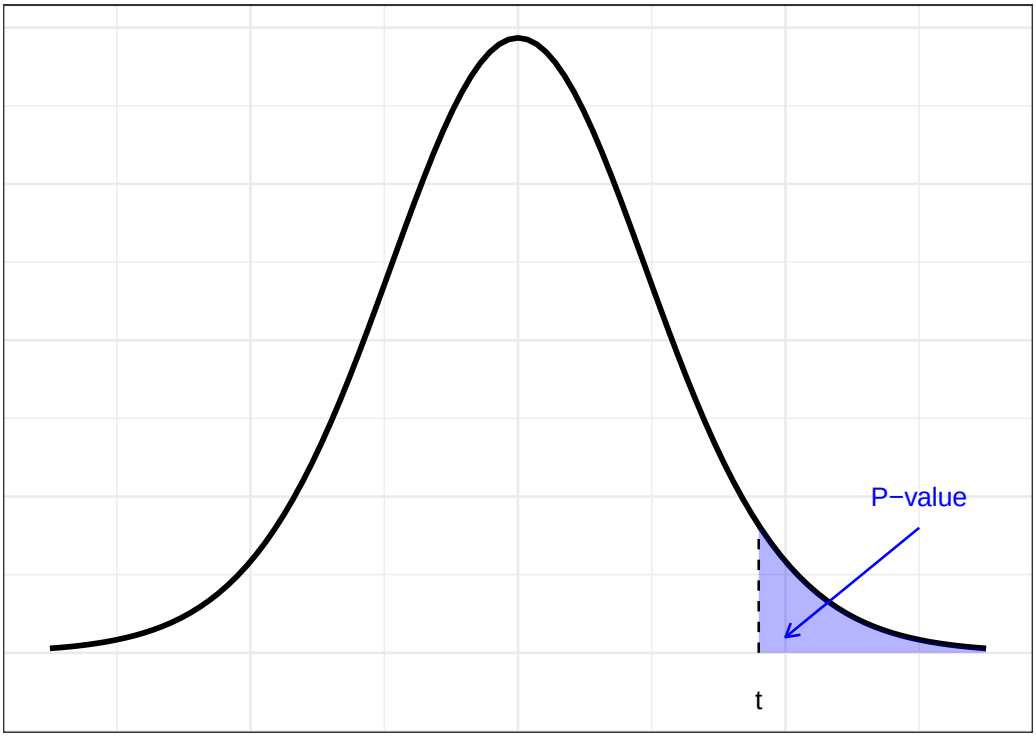


Figure 4.2: For this test, the reference distribution is t($\nu = 30$) (**not** a Normal distribution) and the P-value is the upper-tail area, i.e., to the right of the computed statistic t.

! Shapiro-Wilk normality test

We can assess the normality of the sample by examining the normal quantile-quantile plot as in Example 3.1. For the data in Example 4.6 recall that this is done using the following R code.

```
trees |> ggplot(aes(sample = Volume)) + stat_qq() + stat_qq_line()
```

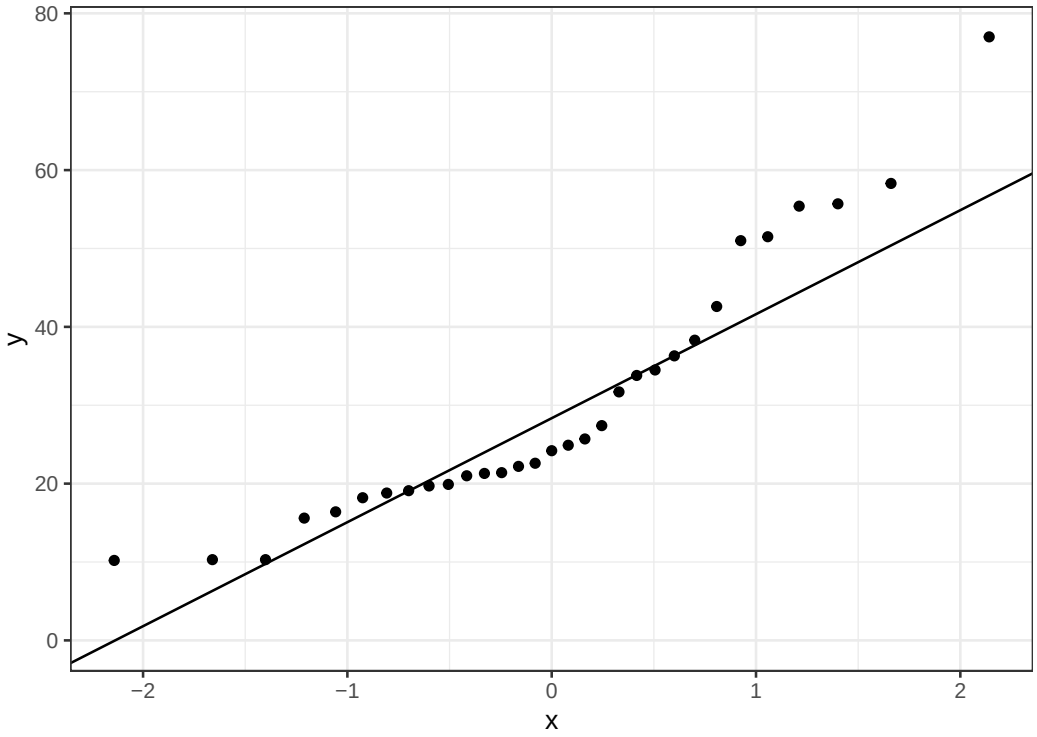


Figure 4.3: Normal quantile-quantile plot for the `Volume` variable (feature) in the Cherry Tree Data.

The data deviates quite a bit in the centre and in the tails of the distribution, indicating that there might be a moderate departure from normality.
It is also possible to test the null hypothesis that the data is consistent with a normal distribution versus the alternative that the data is not normal. This is called a Shapiro-Wilk normality test.

```
shapiro.test(trees$Height)
```

```
    Shapiro-Wilk normality test

data:  trees$Height
W = 0.96545, p-value = 0.4034
```

At level 0.05, the Shapiro-Wilk test yields a P-value $P = 0.4034 > 0.05$, and therefore we fail to reject the null hypothesis. We cannot exclude that the data is drawn from a normal population. This "prove" the data is drawn from a normal distribution, but it does tell us that for this particular example, an inference based on a normal distribution instead of a t distribution will probably be reasonable. It is always good to view the QQ plot as well — sometimes if the number of samples is very large then the Shapiro-Wilk test will reject the null for trivial deviations from normality.

4.2 Estimating proportions and rates

4.2.1 Estimating proportions

Consider a population of size M in which each member either satisfies a given property or does not (i.e. a binary classification). The proportion $p \in (0, 1)$ of the population satisfying the given property is a parameter characterising the population we might be interested in estimating. A sample of classified observations, $X_1, \ldots, X_m \sim \mathsf{Bernoulli}(p)$, from the population contains a proportion,

$$\hat{p} = \frac{1}{m} \sum_{i=1}^{m} X_i \,, \tag{4.8}$$

satisfying the given property. The estimator $\hat{p}$ varies with the sample, and for large m, its sampling distribution has the following properties:

$$\mu_{\hat{p}} = \mathbf{E}[X_i] = p \tag{4.9}$$

and

$$\sigma^2_{\hat{p}} = \frac{\mathrm{Var}[X_i]}{m} = \frac{p(1-p)}{m} \,, \tag{4.10}$$

provided that m is small relative to M. Moreover, by invoking the Central Limit Theorem, we have the distribution of $\hat{p}$ is approximately normal for sufficiently large m as Equation 4.8 is a sample mean. Indeed, this normal approximation works well for moderately large m as long as p is not too close to zero or one; a rule of thumb is that $mp > 5$ and $m(1-p) > 5$.

> **Rule of thumb**
>
> For estimating proportions, a rule of thumb is $m \leq 0.05M$.
> Note that if m is large relative to M ($m > 0.05M$) then the variance Equation 4.10 must be adjusted by a factor (related to the hypergeometric distribution):
>
> $$\sigma^2_{\hat{p}} = \frac{p(1-p)}{m} \frac{M-m}{M-1} \,,$$
>
> where for fixed m the factor converges to 1 as $M \to \infty$.

Proposition 4.6. *For large samples n, a $100(1-\alpha)\%$ confidence interval for the parameter p is given by*

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{m}} \,. \tag{4.11}$$

This follows from Proposition 4.3 by observing that Equation 4.8 is a sample mean and replacing the standard error $\sigma_{\hat{p}}$ from Equation 4.10 by the estimated standard error,

$$\widehat{\mathsf{se}}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{m}} \,;$$

recall the s in Equation 4.5) is the sample variance for the *population* and $s/\sqrt{m} = \mathsf{se}$ is the standard error of the point estimator.

Proposition 4.7. *Let X be the count of members with a given property based on a sample of size m from a population where a proportion p shares the property. Then $\hat{p} = X/m$ is an estimator of p. Assume $mp_0 \geq 10$ and $m(1-p_0) \geq 10$.*

Consider $H_0 : p = p_0$. The test statistic is

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/m}}\,.$$

For a hypothesis test at level α, we use the following procedure:

If $H_a : p > p_0$, then P-value is the area under $\mathsf{N}(0,1)$ to the right of z.

If $H_a : p < p_0$, then P-value is the area under $\mathsf{N}(0,1)$ to the left of z.

If $H_a : p \neq p_0$, then P-value is twice the area under $\mathsf{N}(0,1)$ to the right of $|z|$.

Example 4.7. Let us revisit Example 3.3, where we considered Churchill's claim that he would receive half the votes for the House of Commons seat for the constituency of Dundee. We are sceptical that he is as popular as he says. Suppose 116 out of 263 Dundonians polled claimed they intended to vote for Churchill. Can it be concluded at a significance level of 0.10 that more than half of all eligible Dundonains will vote for Churchill?

The parameter of interest is p, the proportion of votes for Churchill. The null hypothesis is

$$H_0 : p = 0.5\,,$$

and the alternative hypothesis is,

$$H_a : p < 0.5\,.$$

The test is oneside (i.e. $H_a : p < 0.5$) since we are interested in testing support for "more than half". Since $263(0.5) = 131.5 > 10$, we satisfy the assumptions stated in Proposition 4.7.

Based on the sample, $\hat{p} = 116/263 = 0.4411$. The test statistic value is

$$\begin{aligned} z &= \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/m}} \\ &= \frac{0.4411 - 0.5}{\sqrt{0.5(1-0.5)/263}} \\ &= -1.91\,. \end{aligned}$$

The P-value for this lower-tailed z test is $P = \Phi(-1.91) = 0.028$. Since $P < 0.10 = \alpha$, we reject the null hypothesis at the 0.1 level. The evidence for concluding that the true proportion is different from $p_0 = 0.5$ at the 0.10 level is compelling.[2]

[2]Churchill took ca. 44% of the vote in the 1908 by-election to become MP for Dundee [https://www.wikiwand.com/en/1908_Dundee_by-election].

4.2.2 Estimating rates

In many applications, the quantity of interest is not a proportion, but a rate: the average number of events per unit of exposure (often time). For example,

- the rate of accidents per month at a road junction,
- the rate of customer complaints per day, or
- the rate of cases per week in a hospital ward.

A standard model for event counts is the Poisson distribution. Suppose that we observe events over an exposure period of length t (e.g. t days), and let X be the total number of observed events. We model

$$X \sim \mathsf{Poisson}(\lambda t),$$

where $\lambda > 0$ is the event rate per unit exposure.

A natural estimator for the rate is

$$\hat{\lambda} = \frac{X}{t}. \tag{4.12}$$

Since $\mathbf{E}[X] = \lambda t$ and $\mathrm{Var}[X] = \lambda t$ for a Poisson random variable, we obtain

$$\mathbf{E}[\hat{\lambda}] = \lambda \qquad \text{and} \qquad \mathrm{Var}[\hat{\lambda}] = \frac{\lambda}{t}. \tag{4.13}$$

For sufficiently large expected counts (so that the Poisson distribution is not too skewed), the sampling distribution of $\hat{\lambda}$ is approximately normal,

$$\hat{\lambda} \approx \mathsf{N}\Big(\lambda,\ \frac{\lambda}{t}\Big).$$

> **Rule of thumb**
>
> A common rule of thumb for normal approximations with Poisson counts is that the expected count is "large enough", e.g.
>
> $$\lambda t \gtrsim 10.$$
>
> In practice, we often check this using the observed count X, i.e. we look for $X \geq 10$.

To form an interval estimate, we replace the unknown λ in the variance by the point estimate $\hat{\lambda}$, giving the estimated standard error

$$\widehat{\mathsf{se}}(\hat{\lambda}) = \sqrt{\frac{\hat{\lambda}}{t}}.$$

Proposition 4.8. *For a large observed count, an approximate* $100(1-\alpha)\%$ *confidence interval for the Poisson rate* λ *is*

$$\hat{\lambda} \pm z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{t}}. \tag{4.14}$$

> **Exact intervals for small counts**
>
> When the observed count X is small, the normal approximation used in Equation 4.14 can be unreliable. In that case, an exact confidence interval for the Poisson rate λ can be constructed using χ^2

critical values (via the exact sampling distribution of the Poisson count).

We can also test hypotheses about a rate. Under the null hypothesis $H_0 : \lambda = \lambda_0$, the expected count is $\lambda_0 t$ and the standard deviation of X is $\sqrt{\lambda_0 t}$. Thus a natural large-sample test statistic is

$$Z = \frac{X - \lambda_0 t}{\sqrt{\lambda_0 t}} \approx \mathsf{N}(0,1)\,. \tag{4.15}$$

Proposition 4.9. *Assume $X \sim \mathsf{Poisson}(\lambda t)$ and that $\lambda_0 t \geq 10$.*

Consider $H_0 : \lambda = \lambda_0$. The test statistic is

$$Z = \frac{X - \lambda_0 t}{\sqrt{\lambda_0 t}}\,.$$

For a hypothesis test at level α, we use the following procedure:

If $H_a : \lambda > \lambda_0$, then P-value is the area under $\mathsf{N}(0,1)$ to the right of z.

If $H_a : \lambda < \lambda_0$, then P-value is the area under $\mathsf{N}(0,1)$ to the left of z.

If $H_a : \lambda \neq \lambda_0$, then P-value is twice the area under $\mathsf{N}(0,1)$ to the right of $|z|$.

Example 4.8. Suppose a hospital ward records $X = 27$ asthma admissions over $t = 18$ weeks. Estimate the weekly admission rate λ and construct a 95% confidence interval. Then test (at level 0.05) whether the data provide evidence that the rate exceeds $\lambda_0 = 1$ admission per week.

First, the point estimate is

$$\hat{\lambda} = \frac{X}{t} = \frac{27}{18} = 1.5 \quad \text{admissions per week.}$$

Using Proposition 4.8 with $\alpha = 0.05$ and $z_{0.025} = 1.96$, the estimated standard error is

$$\widehat{\mathsf{se}}(\hat{\lambda}) = \sqrt{\frac{\hat{\lambda}}{t}} = \sqrt{\frac{1.5}{18}} = 0.289,$$

so the 95% CI is

$$1.5 \pm 1.96(0.289) = (0.93,\ 2.07).$$

We interpret this as: a plausible range for the true weekly admission rate is between about 0.93 and 2.07 admissions per week.

Now test $H_0 : \lambda = 1$ against $H_a : \lambda > 1$. Under H_0, the expected count over $t = 18$ weeks is $\lambda_0 t = 18$, and the z statistic from Equation 4.15 is

$$z = \frac{27 - 18}{\sqrt{18}} = 2.12.$$

The P-value is the upper-tail area $P = 1 - \Phi(2.12) \approx 0.017$. Since $P < 0.05$, we reject H_0 at the 0.05 level. The data provide evidence that the admission rate exceeds 1 admission per week.

The calculations above can be reproduced using the following R code.

```r
x <- 27
t <- 18

lambda_hat <- x / t
se_hat <- sqrt(lambda_hat / t)

alpha <- 0.05
zcrit <- qnorm(1 - alpha/2)

ci <- lambda_hat + c(-1, 1) * zcrit * se_hat
lambda_hat
```

```
[1] 1.5
```

```r
ci
```

```
[1] 0.9342071 2.0657929
```

For the one-sided test of $H_0 : \lambda = 1$ versus $H_a : \lambda > 1$:

```r
lambda0 <- 1
z <- (x - lambda0 * t) / sqrt(lambda0 * t)
p_val <- 1 - pnorm(z)

z
```

```
[1] 2.12132
```

```r
p_val
```

```
[1] 0.01694743
```

4.3 Estimating variances

Next, we consider estimates of the population variance (and standard deviation) when the population is assumed to have a normal distribution. In this case, the sample variance S^2 in Equation 3.2 provides the basis for inferences. Consider iid samples $X_1, \ldots, X_m \sim \mathsf{N}(\mu, \sigma^2)$. We provide the following theorem without proof.

Theorem 4.2. *For the sample variance S^2 based on m samples from a normal distribution with variance σ^2, the rv*

$$V = \frac{(m-1)S^2}{\sigma^2} = \frac{\sum_i (X_i - \overline{X})^2}{\sigma^2} \qquad \sim \chi^2_{m-1},$$

that is, V has a χ^2 distribution with $\nu = m - 1$ df.

Based on Theorem 4.2,

$$P\left(\chi^2_{1-\alpha/2,m-1} < \frac{(m-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,m-1}\right) = 1-\alpha\,,$$

i.e., the area captured between the right and left tail critical χ^2 values is $1-\alpha$. The expression above can be further manipulated to obtain an interval for the unknown parameter σ^2:

$$P\left(\frac{(m-1)s^2}{\chi^2_{\alpha/2,m-1}} < \sigma^2 < \frac{(m-1)s^2}{\chi^2_{1-\alpha/2,m-1}}\right) = 1-\alpha\,,$$

where we substitute the computed value of the point estimate s^2 for the estimator into the limits to give a CI for σ^2. If we take square roots in the inequality above, we obtain a CI for the population standard deviation σ.

Proposition 4.10. *A $100(1-\alpha)\%$ confidence interval for the variance of a normal population is*

$$\left((m-1)s^2/\chi^2_{\alpha/2,m-1}\,,(m-1)s^2/\chi^2_{1-\alpha/2,m-1}\right)\,. \tag{4.16}$$

A $100(1-\alpha)\%$ confidence interval for the standard deviation σ of a normal population is given by taking the square roots of the lower and upper limits in Equation 4.16.

Example 4.9. For the **Cherry Tree Data** in Table 4.1 concerning the timber volume of 31 felled black cherry trees, give a 95 CI for the variance.

We are interested in estimating the true variance σ^2 of the timber volume based on $m = 31$ samples. Recall that the mean of our data is $\overline{x} = 30.17$ cu ft and that the sample variance is $s^2 = 270.2$ using the estimator Equation 3.2. The critical values for the $\chi^2_{.975,30} = 16.7908$ and $\chi^2.025,30 = 46.9792$ can be found by checking a table of critical values of the $\chi^2(\nu = 30)$ distribution or by using the `R` code `qchisq(1-0.05/2, df=30, lower.tail = FALSE)` and `qchisq(0.05/2, df=df, lower.tail = FALSE)`, respectively, see Figure 4.4 below.

Pulling everything together, a 95% CI for the population variance is given by

$$\begin{aligned}&\left((m-1)s^2/\chi^2_{\alpha/2,m-1}\,,(m-1)s^2/\chi^2_{1-\alpha/2,m-1}\right)\\ &\quad= ((30)270.2/46.9792\,,(30)270.2/16.7908)\\ &\quad= (172.5\,,482.8)\,.\end{aligned}$$

Note the position of the critical values—don't swap them around.

Example 4.10. Let's Revisit Example 4.9 and use the `infer` package to construct a 95% confidence interval for the true standard deviation of the timber volume of black cherry trees based on the available measurements in the **Cherry Tree Data**, Table 4.1.

```
s <- sd(trees$Volume)

null_dist <- trees |>
 specify(response = Volume) |>
 generate(reps = 1000, type = "bootstrap") |>
 calculate(stat = "sd")
```

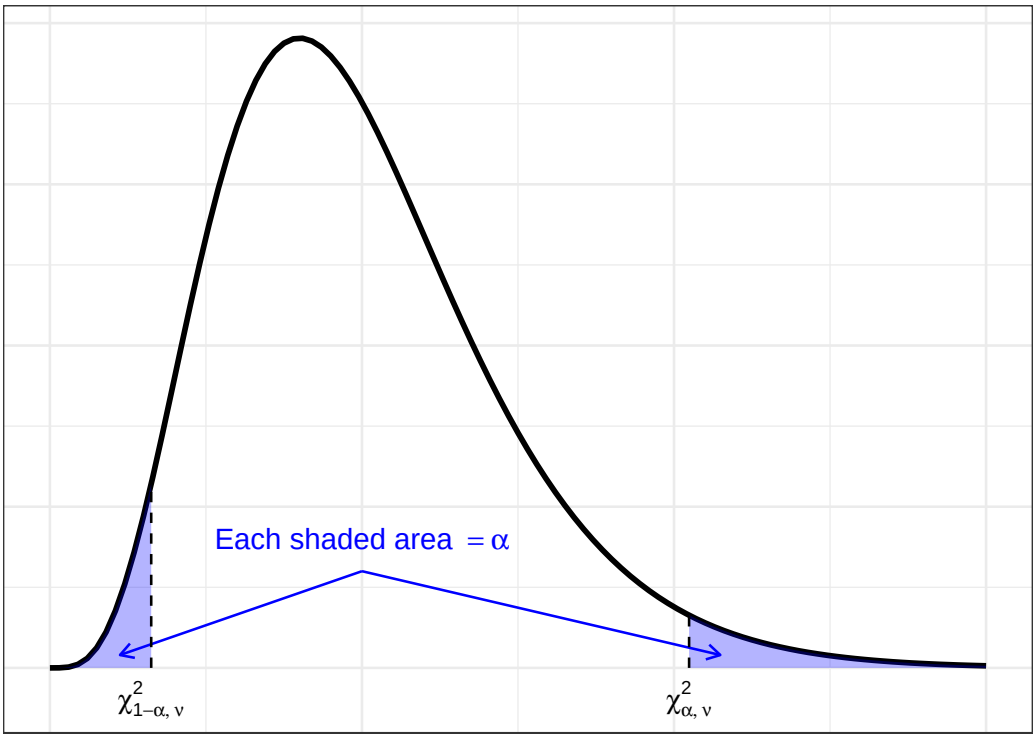
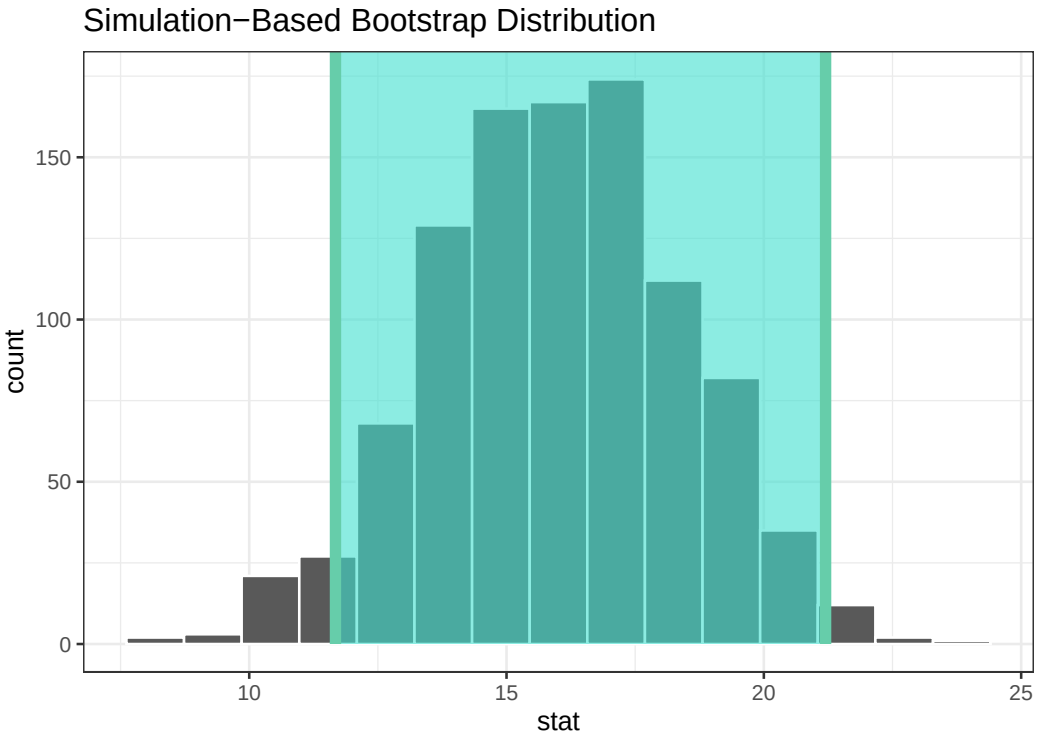


Figure 4.4: As the χ^2 distribution is not symmetric, the upper and lower critical values will not be the same (the shaded areas are equal).

```
ci <- null_dist |>
 get_confidence_interval(point_estimate = s, level = 0.95, type = "se")

null_dist |>
 visualise() + shade_ci(ci)
```



We plot the 95% confidence interval for the standard deviation based on the computational null distribution obtained using 1000 bootstrap replications; note the interval estimate is in good agreement with the values obtained in Example 4.9.

```
ci^2
```

```
  lower_ci upper_ci
1 136.3271 449.4304
```

Due to the computational nature, the bootstrapped interval estimate is not precisely the same as the theoretical interval estimate and rerunning the code will yield a slightly different interval.

5 Two-sample inferences

We consider inferences—estimators, confidence intervals, and hypothesis testing—for comparing means, proportions, and variances based on two independent samples from different populations, respectively, in Section 5.1, Section 5.3, Section 5.4. We also consider inferences when the samples are not independent, so-called paired samples, in Section 5.2.

5.1 Comparing means

Let us assume that we have two normal populations with iid samples

$$X_1, \ldots, X_m \sim \mathsf{N}(\mu_X, \sigma_X^2)$$

and

$$Y_1, \ldots, Y_n \sim \mathsf{N}(\mu_Y, \sigma_Y^2)$$

and, moreover, that the X and Y samples are independent of one another. When comparing the means of two populations, the quantity of interest is the difference: $\mu_X - \mu_Y$.

Proposition 5.1. *If we consider the sample means $\overline{X}$ and $\overline{Y}$, then the mean of the variable $\overline{X} - \overline{Y}$ is,*

$$\mu_{\overline{X}-\overline{Y}} = \mathbf{E}\left[\overline{X} - \overline{Y}\right] = \mu_X - \mu_Y\,,$$

and the variance is,

$$\sigma^2_{\overline{X}-\overline{Y}} = \mathrm{Var}\left[\overline{X} - \overline{Y}\right] = \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\,.$$

Proposition 5.1 follows directly from the definition of the sample mean in Equation 3.1 and properties of expectation and variance. If our parameter of interest is

$$\theta = \mu_1 - \mu_2\,,$$

then its estimator,

$$\hat{\theta} = \overline{X} - \overline{Y}\,,$$

is normally distributed with mean and variance given by Proposition 5.1. If the sample sizes m and n are large, then the estimator is approximately normally distributed by the Central Limit Theorem regardless of the population. We now discuss CIs and hypothesis tests for comparing population means $\theta = \mu_X - \mu_Y$. We consider three cases when comparing means:

1. normal populations when the variances σ_X^2 and σ_Y^2 are known,
2. any populations with unknown variances σ_X^2 and σ_Y^2, when the sample sizes m and n are large,
3. normal populations when the variances σ_X^2 and σ_Y^2 are unknown, when the sample sizes m and n are small,

noting that the development primarily reflects that of Section 4.1.

5.1.1 Comparing means of normal populations when variances are known

When σ_X^2 and σ_Y^2 are known, standardizing $\overline{X} - \overline{Y}$ yields the standard normal variable:

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \quad \sim \mathsf{N}(0,1)\,. \tag{5.1}$$

Inferences proceed by treating the parameter of interest θ as in the single sample case using the test statistic Equation 5.1.

Proposition 5.2. *A* $100(1-\alpha)\%$ *CI for the parameter* $\theta = \mu_X - \mu_Y$ *based on samples of size* m *from a normal population* $\mathsf{N}(\mu_X, \sigma_X^2)$ *and of size* n *from* $\mathsf{N}(\mu_Y, \sigma_Y^2)$ *with known variances, is given by*

$$(\overline{x} - \overline{y}) \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}\,.$$

Proposition 5.3. *Assume that we sample iid* $X_1, \ldots, X_m \sim \mathsf{N}(\mu_X, \sigma_X^2)$ *and iid* $Y_1, \ldots, Y_n \sim \mathsf{N}(\mu_Y, \sigma_Y^2)$ *and that the* X *and* Y *samples are independent.*

Consider $H_0 : \mu_X - \mu_Y = \theta_0$. *The test statistic is*

$$Z = \frac{\overline{X} - \overline{Y} - \theta_0}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}\,. \tag{5.2}$$

For a hypothesis test at level α, *we use the following procedure:*

If $H_a : \mu_X - \mu_Y > \theta_0$, *then* $P = 1 - \Phi(z)$, *i.e., upper-tail* $R = \{z > z_\alpha\}$.

If $H_a : \mu_X - \mu_Y < \theta_0$, *then* $P = \Phi(z)$, *i.e., lower-tail* $R = \{z < -z_\alpha\}$.

If $H_a : \mu_X - \mu_Y \neq \theta_0$, *then* $P = 2(1 - \Phi(|z|))$, *i.e., two-tailed* $R = \{|z| > z_{\alpha/2}\}$.

5.1.2 Comparing means when the sample sizes are large

When the samples are large, the assumptions about the normality of the populations and knowledge of the variances σ_X^2 and σ_Y^2 can be relaxed. For sufficiently large m and n, the difference of the sample means, $\overline{X} - \overline{Y}$, has approximately a normal distribution for any underlying population distributions by the Central Limit Theorem. Moreover, if m and n are large enough, replacing the population variances with the sample variances S_X^2 and S_Y^2 will not increase the variability of the estimator or the test statistic too much.

Proposition 5.4. *For* m *and* n *sufficiently large, an approximate* $100(1-\alpha)\%$ *CI for* $\mu_X - \mu_Y$ *for two samples from populations with any underlying distribution is given by*

$$(\overline{x} - \overline{y}) \pm z_{\alpha/2} \cdot \sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}$$

Proposition 5.5. *Under the same assumptions and procedures as in Proposition 5.3, a large-sample, i.e., $m > 40$ and $n > 40$, test statistic,*

$$Z = \frac{\overline{X} - \overline{Y} - \theta_0}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}},$$

can be used in place of Equation 5.2 for hypothesis testing.

5.1.3 Comparing means of normal populations when variances are unknown and the sample size is small

If σ_X and σ_Y are unknown and either sample is small (e.g., $m < 30$ or $n < 30$), but both populations are normally distributed, then we can use Student's t distribution to make inferences. We provide the following theorem without proof.

Theorem 5.1. *When both population distributions are normal, the standardised variable*

$$T = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}} \sim \mathsf{t}(\nu),$$

where the df ν is estimated from the data. Namely, ν is given by (round ν down to the nearest integer):

$$\nu = \frac{\left(\frac{s_X^2}{m} + \frac{s_Y^2}{n}\right)^2}{\frac{(s_X^2/m)^2}{m-1} + \frac{(s_Y^2/n)^2}{n-1}} = \frac{\left(s_{\overline{X}}^2 + s_{\overline{Y}}^2\right)^2}{\frac{s_{\overline{X}}^4}{m-1} + \frac{s_{\overline{Y}}^4}{n-1}}, \tag{5.3}$$

where s_X^2 and s_Y^2 are point estimators of the sample variances; alternatively, we see that the formula Equation 5.3 can also be written in terms of the standard error of the sample means:

$$s_{\overline{X}} = \frac{s_X}{\sqrt{m}} \quad \textit{and} \quad s_{\overline{Y}} = \frac{s_Y}{\sqrt{n}}.$$

The formula Equation 5.3 for the data-driven choice of ν calls for the computation of the standard error of the sample means.

Proposition 5.6. *A $100(1-\alpha)\%$ CI for $\mu_X - \mu_Y$ for two samples of size m and n from normal populations where the variances are unknown is given by*

$$(\overline{x} - \overline{y}) \pm t_{\alpha/2,\nu} \sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}},$$

where we recall that $t_{\alpha/2,\nu}$ is the $\alpha/2$ critical value of $\mathsf{t}(\nu)$ with ν given by Equation 5.3.

Proposition 5.7. *Assume that we sample iid $X_1, \ldots, X_m$ and iid $Y_1, \ldots, Y_n$ from normal populations with unknown variances and means μ_X and μ_Y, respectively, and that the X and Y samples are independent.*

Consider $H_0 : \mu_X - \mu_Y = \theta_0$. The test statistic is

$$T = \frac{\overline{X} - \overline{Y} - \theta_0}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}. \tag{5.4}$$

For a hypothesis test at level α, we use the following procedure:

If $H_a : \mu_X - \mu_Y > \theta_0$, then P-value is the area under $\mathsf{t}(\nu)$ to the right of t, i.e., upper-tail $R = \{t > t_{\alpha,\nu}\}$.

If $H_a : \mu_X - \mu_Y < \theta_0$, then P-value is the area under $\mathsf{t}(\nu)$ to the left of t, i.e., lower-tail $R = \{t < -t_{\alpha,\nu}\}$.

If $H_a : \mu_X - \mu_Y \neq \theta_0$, then P-value is twice the area under $\mathsf{t}(\nu)$ to the right of $|t|$, i.e., two-tailed $R = \{|t| > t_{\alpha/2,\nu}\}$.

Here ν is given by Equation 5.3.

If the variances of the normal populations are unknown but are the same, $\sigma_X^2 = \sigma_Y^2$, then deriving CIs and test statistics for comparing the means can be simplified by considering a combined or pooled estimator for the single parameter σ^2. If we have two samples from populations with variance σ^2, each sample provides an estimate for σ^2. That is, S_X^2, based on the m observations of the first sample, is one estimator for σ^2 and another is given by S_Y^2, based on n observations of the second sample. The correct way to combine these two estimators into a single estimator for the sample variance is to consider the pooled estimator of σ^2,

$$S_{\mathsf{p}}^2 = \frac{m-1}{m+n-2} S_X^2 + \frac{n-1}{m+n-2} S_Y^2 \,. \tag{5.5}$$

The pooled estimator is a weighted average that adjusts for differences between the sample sizes m and n.

> ⚠ Why a weighted average?
>
> If $m \neq n$, then the estimator with *more* samples will contain *more* information about the parameter σ^2. Thus, the simple average $(S_X^2 + S_Y^2)/2$ wouldn't be fair, would it?

Proposition 5.8. *A $100(1-\alpha)\%$ CI for $\mu_X - \mu_Y$ for two samples of size m and n from normal populations where the variance σ^2 is unknown is given by*

$$(\overline{x} - \overline{y}) \pm t_{\alpha/2,m+n-2} \cdot \sqrt{s_{\mathsf{p}}^2 \left(\frac{1}{m} + \frac{1}{n}\right)},$$

where we recall that $t_{\alpha/2,m+n-2}$ is the $\alpha/2$ critical value of the $\mathsf{t}(\nu)$ with $\nu = m + n - 2$ df.

Similarly, one can consider a pooled t test, i.e., a hypothesis test based on the pooled estimator for the variance as opposed to the two-sample t test in Proposition 5.7. In the case of a pooled t test, the test statistic

$$T = \frac{\overline{X} - \overline{Y} - \theta_0}{\sqrt{S_{\mathsf{p}}^2 \left(\frac{1}{m} + \frac{1}{n}\right)}},$$

with the pooled estimator of the variance, replaces Equation 5.4 in Proposition 5.7 and the same procedures are followed for determining the P-value with $\nu = m + n - 2$ in place of Equation 5.3. If you have reasons to believe that $\sigma_X^2 = \sigma_Y^2$, these pooled t procedures are appealing because ν is very easy to compute.

> ⚠ Robustness
>
> Pooled t procedures are not robust if the assumption of equal variance is violated. Theoretically, you could first carry out a statistical test $H_0 : \sigma_X^2 = \sigma_Y^2$ on the equality of variances and then use a

pooled t procedure if the null hypothesis is not rejected. However, there is no free lunch: the typical F test for equal variances (see Section 5.4) is sensitive to normality assumptions. The two sample t procedures, with the data-driven choice of ν in Equation 5.3, are therefore recommended unless, of course, you have a very compelling reason to believe $\sigma_X^2 = \sigma_Y^2$.

5.2 Comparing paired samples

The preceding analysis for comparing population means was based on the assumption that a random sample $X_1, \ldots, X_n$ is drawn from a distribution with mean μ_X and that a completely independent random sample $Y_1, \ldots, Y_n$ is drawn from a distribution with mean μ_Y. Some situations, e.g., comparing observations before and after a treatment or exposure, necessitate the consideration of paired values.

Consider a random sample of iid pairs,

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

with $\mathbf{E}[X_i] = \mu_X$ and $\mathbf{E}[Y_i] = \mu_Y$. If we are interested in making inferences about the difference $\mu_X - \mu_Y$, then the paired differences

$$D_i = X_i - Y_i, \quad i = 1, \ldots, n,$$

constitute a sample with mean $\mu_D = \mu_X - \mu_Y$ that can be treated using single-sample CIs and tests, e.g., see Section 4.1.3.

5.3 Comparing proportions

Consider a population containing a proportion p_X of individuals satisfying a given property. For a sample of size m from this population, we denote the sample proportion by $\hat{p}_X$. Likewise, we consider a population containing a proportion p_Y of individuals satisfying the same given property. For a sample of size n from this population, we denote the sample proportion by $\hat{p}_Y$. We assume the samples from the X and Y populations are independent. The natural estimator for the difference in population proportions $p_X - p_Y$ is the difference in the sample proportions $\hat{p}_X - \hat{p}_Y$.

Provided the samples are much smaller than the population sizes (i.e., the populations are about 20 times larger than the samples),

$$\mu_{(\hat{p}_X - \hat{p}_Y)} = \mathbf{E}[\hat{p}_X - \hat{p}_Y] = p_X - p_Y,$$

and

$$\sigma^2_{(\hat{p}_X - \hat{p}_Y)} = \mathrm{Var}[\hat{p}_X - \hat{p}_Y] = \frac{p_X(1-p_X)}{m} + \frac{p_Y(1-p_Y)}{n},$$

because the count of individuals satisfying the given property in each population will be independent draws from $\mathsf{Binom}(m, p_X)$ and $\mathsf{Binom}(n, p_Y)$, respectively. Further, if m and n are large (e.g., $m \geq 30$ and $n \geq 30$), then $\hat{p}_X$ and $\hat{p}_Y$ are (approximately) normally distributed. Standardizing $\hat{p}_X - \hat{p}_Y$,

$$Z = \frac{\hat{p}_X - \hat{p}_Y - (p_X - p_Y)}{\sqrt{\frac{p_X(1-p_X)}{m} + \frac{p_Y(1-p_Y)}{n}}} \sim \mathsf{N}(0,1).$$

A CI for $\hat{p}_X - \hat{p}_Y$ then follows from the large-sample CI considered in Section 4.1.2.

Proposition 5.9. *An approximate* $100(1-\alpha)\%$ *CI for* $p_X - p_Y$ *is given by*

$$\hat{p}_X - \hat{p}_Y \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{m} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n}}\,, \tag{5.6}$$

and, as a rule of thumb, can be reliably used if $m\hat{p}_X$, $m(1-\hat{p}_X)$, $n\hat{p}_Y$, *and* $n(1-\hat{p}_Y)$ *are greater than or equal to* 10.

Proposition 5.9 does not pool the estimators for the population proportions. However, if we are considering a hypothesis test concerning the equality of the population proportions with the null hypothesis

$$H_0 : p_X - p_Y = 0\,,$$

then we assume $p_X = p_Y$ as our default position. Therefore, as a matter of consistency, we should replace the standard error in Equation 5.6 with a pooled estimator for the standard error of the population proportion,

$$\hat{p} = \frac{m}{m+n}\hat{p}_X + \frac{n}{m+n}\hat{p}_Y\,.$$

Proposition 5.10. *Assume that* $m\hat{p}_X$, $m(1-\hat{p}_X)$, $n\hat{p}_Y$, $n(1-\hat{p}_Y)$ *are all greater than* 10.

Consider $H_0 : p_X - p_Y = 0$. *The test statistic is*

$$Z = \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{m}+\frac{1}{n}\right)}}\,.$$

For a hypothesis test at level α, *we use the following procedure:*

If $H_a : p_X - p_Y > 0$, *then* $P = 1 - \Phi(z)$, *i.e., upper-tail* $R = \{z > z_\alpha\}$.

If $H_a : p_X - p_Y < 0$, *then* $P = \Phi(z)$, *i.e., lower-tail* $R = \{z < -z_\alpha\}$.

If $H_a : p_X - p_Y \neq 0$, *then* $P = 2(1 - \Phi(|z|))$, *i.e., two-tailed* $R = \{|z| > z_{\alpha/2}\}$.

5.4 Comparing variances

For a random sample

$$X_1, \ldots, X_m \sim \mathsf{N}(\mu_X, \sigma_X^2)$$

and an independent random sample

$$Y_1, \ldots, Y_n \sim \mathsf{N}(\mu_Y, \sigma_Y^2)\,,$$

the rv

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \quad \sim \mathsf{F}(m-1, n-1)\,, \tag{5.7}$$

that is, F has an F distribution with df $\nu_1 = m-1$ and $\nu_2 = n-1$. The statistic F in Equation 5.7 comprises the *ratio* of variances σ_X^2/σ_Y^2 and not the difference; therefore, the plausibility of $\sigma_X^2 = \sigma_Y^2$ will be based on how much the ratio differs from 1.

Proposition 5.11. *For the null hypothesis $H_0 : \sigma_X^2 = \sigma_Y^2$, the test statistic to consider is:*

$$f = \frac{s_X^2}{s_Y^2}$$

and the P-values are determined by the $\mathsf{F}(m-1, n-1)$ distribution where m and n are the respective sample sizes.

A $100(1-\alpha)\%$ CI for the ratio σ_X^2/σ_Y^2 is based on forming the probability,

$$P(F_{1-\alpha/2,\nu_1,\nu_2} < F < F_{\alpha/2,\nu_1,\nu_2}) = 1-\alpha\,,$$

where $F_{\alpha/2,\nu_1,\nu_2}$ is the $\alpha/2$ critical value from the $\mathsf{F}(\nu_1 = m-1, \nu_2 = n-1)$ distribution. Substituting Equation 5.7 with point estimates for F and manipulating the inequalities it is possible to isolate the ratio σ_X^2/σ_Y^2,

$$P\left(\frac{1}{F_{\alpha/2,\nu_1,\nu_2}}\frac{s_X^2}{s_Y^2} < \frac{\sigma_X^2}{\sigma_Y^2} < \frac{1}{F_{1-\alpha/2,\nu_1,\nu_2}}\frac{s_X^2}{s_Y^2}\right) = 1-\alpha\,.$$

Proposition 5.12. *A $100(1-\alpha)\%$ CI for the ratio of population variances σ_X^2/σ_Y^2 is given by*

$$\left(F^{-1}_{\alpha/2,m-1,n-1}s_X^2/s_Y^2\,,\, F^{-1}_{1-\alpha/2,m-1,n-1}s_X^2/s_Y^2\right)\,.$$

Proposition 5.13. *Assume the population distributions are normal and the random samples are independent of one another.*

Consider $H_0 : \sigma_X^2 = \sigma_Y^2$. The test statistic is

$$F = S_X^2/S_Y^2\,.$$

For a hypothesis test at level α, we use the following procedure:

If $H_a : \sigma_X^2 > \sigma_Y^2$, then P-value is $A_R =$ area under the $\mathsf{F}(m-1, n-1)$ curve to the right of f.

If $H_a : \sigma_X^2 < \sigma_Y^2$, then P-value is $A_L =$ area under the $\mathsf{F}(m-1, n-1)$ curve to the left of f.

If $H_a : \sigma_X^2 \neq \sigma_Y^2$, then P-value is $2 \cdot \min(A_R, A_L)$.

6 Analysis of variance

Analysis of variance, shortened as ANOVA or AOV, is a collection of statistical models and estimation procedures for analysing the variation among different groups. In particular, a single-factor ANOVA provides a hypothesis test regarding the equality of two or more population means, thereby generalising the one-sample and two-sample t tests considered in Section 4.1.3 and Section 5.1.3.

6.1 Single factor ANOVA test

Suppose that we have k normally distributed populations with different means $\mu_1, \dots, \mu_k$ and equal variances σ^2. We denote the rv for the jth measurement taken from the ith population by X_{ij} and the corresponding sample observation by x_{ij}. For samples of size $m_1, \dots, m_k$, we denote the sample means

$$\overline{X}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} X_{ij} \,,$$

and sample variances

$$S_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (X_{ij} - \overline{X}_i)^2 \,,$$

for each $i = 1, \dots, k$; likewise, we denote the associated point estimates for the sample means $\overline{x}_1, \dots, \overline{x}_k$ and the sample variances $s_1^2, \dots, s_k^2$. The average over all observations $m = \sum m_i$, called the grand mean, is denoted by

$$\overline{X} = \frac{1}{m} \sum_{i=1}^{k} \sum_{j=1}^{m_i} X_{ij} \,.$$

The sample variances s_i^2, and hence the sample standard deviations, will generally vary even when the k populations share the same variance; a rule of thumb is that the equality of variances is reasonable if the largest s_i is not much more than two times the smallest.

> Alternative lingo
>
> In the context of ANOVA, these k populations are often referred to as *treatment* distributions.

We wish to test the equality of the population means, given by the null hypothesis,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \,,$$

versus the alternative hypothesis,

$$H_a : \text{at least two } \mu_i \text{ differ} \,.$$

Note that if $k = 3$ then H_0 is true only if all three means are the same, i.e., $\mu_1 = \mu_2 = \mu_3$, but there are a number of ways which the alternative might hold: $\mu_1 \neq \mu_2 = \mu_3$ or $\mu_1 = \mu_2 \neq \mu_3$ or $\mu_1 = \mu_3 \neq \mu_2$ or $\mu_1 \neq \mu_2 \neq \mu_3$.

The test procedure is based on comparing a measure of the difference in variation among the sample means, i.e., the variation between x_i's, to a measure of variation within each sample.

Definition 6.1. The mean square for treatments is

$$\mathsf{MSTr} = \frac{1}{k-1}\sum_{i=1}^{k} m_i(\overline{X}_i - \overline{X})^2 \,,$$

and the mean square error is

$$\mathsf{MSE} = \frac{1}{m-k}\sum_{i=1}^{k}(m_i - 1)S_i^2 \,.$$

The MSTr and MSE are statistics that measure the variation among sample means and the variation within samples. We will also use MSTr and MSE to denote the calculated values of these statistics.

Proposition 6.1. *The test statistic*

$$F = \frac{\mathsf{MSTr}}{\mathsf{MSE}}$$

is the appropriate test statistic for the single-factor ANOVA problem involving k populations (or treatments) with a random sample of size $m_1, \ldots, m_k$ from each. When H_0 is true,

$$F \sim \mathsf{F}(\nu_1 = k-1, \nu_2 = m-k) \,.$$

In the present context, a large test statistic value is more contradictory to H_0 than a smaller value. Therefore the test is upper-tailed, i.e., consider the area F_α to the right of the critical value F_{α,ν_1,ν_2}. We reject H_0 if the value of the test statistic $F > F_\alpha$.

> **i** Note 3: Average Salary Data
>
> The **Average Salary Data** comprises average salaries reported by 20 local councils across the four nations of the United Kingdom (England, N Ireland, Scotland and Wales). The sample means and sample standard deviations are summarised in Table 6.1.

Table 6.1: **Average Salary Data** reported from 20 local councils.

Nation	Average salaries ('000 £)	Sample size	Sample mean	Sample sd
England	17, 12, 18, 13, 15, 12	6	14.5	2.588
N Ireland	11, 7, 9, 13	4	10.0	2.582
Scotland	15, 10, 13, 14, 13	5	13.0	1.871
Wales	10, 12, 8, 7, 9	5	9.2	1.924

Example 6.1. Consider the **Average Salary Data** reported in Note 3. Is the expected average salary in each nation the same at the 5% level?

We begin by exploring the data through the generation and interpretation of some box plots. The box plots in Figure 6.1 indicate that there may be a difference in median average salary by nation.

For $\alpha = 0.05$, we compute the upper-tail area $F_{0.05}$ i.e. to the right of the critical value $F_{0.05,3,16}$ by consulting a statistical table or by using R to find $F_{0.05} = 3.2388715$.

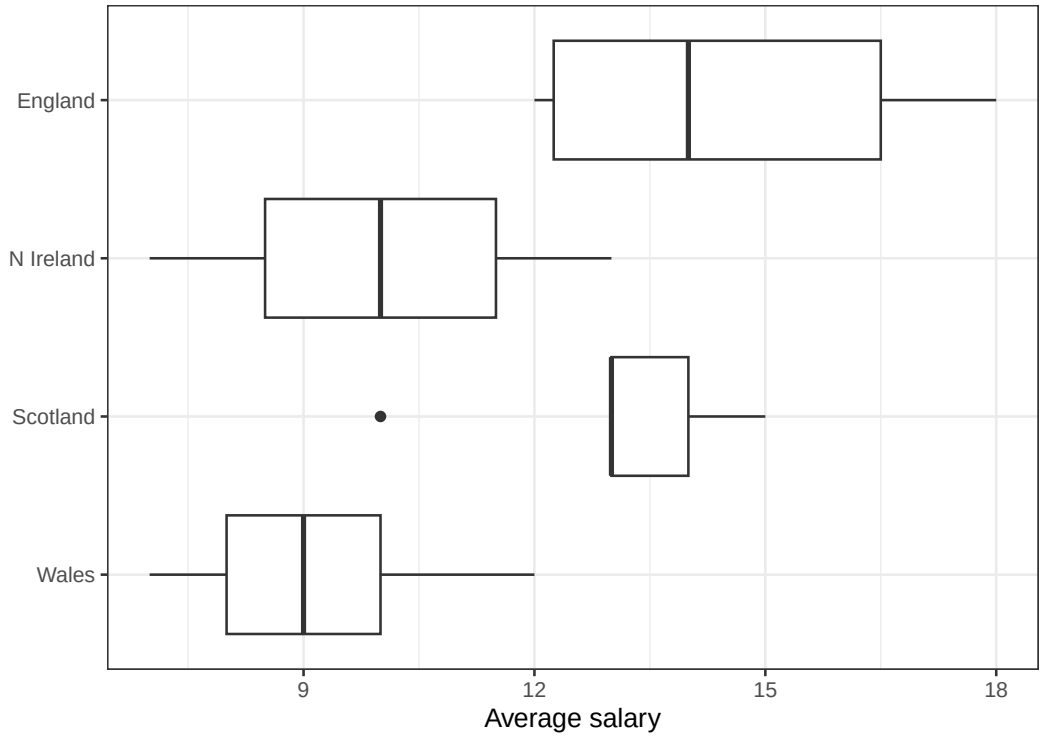


Figure 6.1: Box plots of the average mean salary data in Table Table 6.1 indicate five summary statistics: the median, two hinges (first and third quartiles) and two whiskers (extending from the hinge to the most extreme data point within $1.5 \cdot \mathrm{IQR}$).

```
# alt: qf(.05, df1 = 3, df2 = 16, lower.tail = FALSE)
qf(1-.05, df1 = 4-1, df2 = 20-4)
```

```
[1] 3.238872
```

The grand mean is

$$\overline{x} = \frac{17 + 12 + 18 + \cdots + 8 + 7 + 9}{20} = 11.9\,,$$

and hence the variation among sample means is given by,

$$\begin{aligned}
\mathsf{MSTr} &= \frac{1}{4-1}\left(m_1(\overline{x}_1 - \overline{x})^2 + \cdots + m_4(\overline{x}_4 - \overline{x})^2\right) \\
&= \left(6(14.5 - 11.9)^2 + 4(10.0 - 11.9)^2 + 5(13.0 - 11.9)^2 + 5(9.2 - 11.9)^2\right)/3 \\
&= 32.5\,.
\end{aligned}$$

The mean square error is

$$\begin{aligned}
\mathsf{MSE} &= \frac{1}{20-4}\left((m_1 - 1)s_1^2 + \cdots (m_4 - 1)s_4^2\right) \\
&= \frac{5(2.588)^2 + 3(2.582)^2 + 4(1.871)^2 + 4(1.924)^2}{16} \\
&= 5.14366
\end{aligned}$$

yielding the test statistic value

$$F = \frac{\mathsf{MSTr}}{\mathsf{MSE}} = \frac{32.5}{5.14366} = 6.3184581\,.$$

Since $F > F_\alpha$ we reject H_0. The data do not support the hypothesis that the mean salaries in each nation are identical at the 5% level.

6.2 Confidence intervals

In Section 5.1, we gave a CI for comparing population means involving the difference $\mu_X - \mu_Y$. In some settings, we would like to give CIs for more complicated functions of population means μ_i. Let

$$\theta = \sum_{i=1}^{k} c_i \mu_i \,,$$

for constants c_i. As we assume the $X_i j$ are normally distributed with $\mathbf{E}[X_{ij}] = \mu_i$ and $\mathrm{Var}[X_{ij}] = \sigma^2$, the estimator

$$\hat{\theta} = \sum_{i=1}^{k} c_i \overline{X}_i \,,$$

is normally distributed with

$$\mathrm{Var}[\hat{\theta}] = \sum_{i=1}^{k} c_i^2 \,\mathrm{Var}[\overline{X}_i] = \sigma^2 \sum_{i=1}^{k} \frac{c_i^2}{m_i} \,.$$

We estimate σ^2 by the MSE and standardise the estimator to arrive at a t variable

$$\frac{\hat{\theta} - \theta}{\widehat{\sigma}_{\hat{\theta}}} \,,$$

where $\widehat{\sigma}_{\hat{\theta}}$ is the estimated standard error of the estimator.

Proposition 6.2. *A* $100(1-\alpha)\%$ *CI for* $\sum c_i \mu_i$ *is given by*

$$\sum_{i=1}^{k} c_i \overline{x}_i \pm t_{\alpha/2, m-k} \sqrt{\mathrm{MSE} \sum_{i=1}^{k} \frac{c_i^2}{m_i}} \,.$$

Example 6.2. Determine a 90% CI for the difference in mean average salary for councils in Scotland and England, based on the data available in Table 6.1

For $\alpha = 0.10$, the critical value $t_{0.05,16} = 1.7458837$ is found by looking in a table of t critical values or by using R:

```
# alt: qt(0.1/2, 16, lower.tail = FALSE)
qt(1-0.1/2, df = 20 - 4)
```

```
[1] 1.745884
```

Then for the function $\overline{x}_2 - \overline{x}_1$,

$$\begin{aligned}(\overline{x}_{Eng} - \overline{x}_{Sco}) &\pm t_{0.05,16}\sqrt{\mathsf{MSE}}\sqrt{\frac{1}{m_{Eng}} + \frac{1}{m_{Sco}}} \\ &= (14.5 - 13.0) \pm 1.7458837\sqrt{5.14366}\sqrt{\frac{1}{6} + \frac{1}{5}} \\ &= 1.5 \pm 2.3976575\,.\end{aligned}$$

Thus, a 90% confidence interval for $\mu_{Eng} - \mu_{Sco}$ is $(-0.8977\,, 3.898)$.

> **Consider the following**
>
> How does the result in Example 6.2 compare to the t method in Section 5.1.3?

7 Linear regression

Regression analysis allows us to study the relationship among two or more rvs. Typically, we are interested in the relationship between a response or dependent rv Y and a covariate X. The relationship between X and Y will be explained through a regression function,

$$r(x) = \mathbf{E}[Y \mid X = x] = \int y f(y \mid x) dy \,.$$

In particular, we shall assume that r is linear,

$$r(x) = \beta_0 + \beta_1 x \,, \tag{7.1}$$

and estimate the intercept β_0 and slope β_1 of this linear model from sample data

$$(Y_1, X_1), \ldots, (Y_m, X_m) \sim F_{Y,X} \,.$$

> **Alternative lingo**
>
> The covariates X are also called predictor variables, explanatory variables, independent variables, and/or features depending on who you are talking to.

7.1 Simple linear regression models

The simplest regression is when X_i is one-dimensional and $r(x)$ is linear as in Equation 7.1. A linear regression posits the expected value of Y_i is a linear function of the data X_i, but that Y deviates from its expected value by a random amount for fixed x_i.

Definition 7.1. The simple linear regression model relates a random response Y_i to a set of independent variables X_i,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \,, \tag{7.2}$$

where the intercept β_0 and slope β_1 are unknown parameters and the random deviation or random error ϵ_i is a rv assumed to satisfy:

1. $\mathbf{E}[\epsilon_i \mid X_i = x_i] = 0$,
2. $\mathrm{Var}[\epsilon_i \mid X_i = x_i] = \sigma^2$ does not depend on x_i,
3. ϵ_i and ϵ_j are independent for $i, j = 1, \ldots, m$.

From the assumptions on ϵ_i, the linear model Equation 7.2 implies

$$\mathbf{E}[Y_i \mid X_i = x_i] = \beta_0 + \beta_1 x_i \,.$$

Thus, if $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators of β_0 and β_1, then the fitted line is

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

and the predicted or fitted value $\widehat{Y}_i = \hat{r}(X_i)$ is an estimator for $\mathbf{E}[Y_i \mid X_i = x_i]$. The residuals are defined to be

$$\hat{\epsilon}_i = Y_i - \widehat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right) . \tag{7.3}$$

The residual sums of squares,

$$\mathsf{RSS} = \sum_{i=1}^{m} \hat{\epsilon}_i^2 , \tag{7.4}$$

measures how well the regression line $\hat{r}$ fits the data $(Y_1, X_1), \ldots, (Y_m, X_m)$. The least squares estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are the values that minimize the RSS in Equation 7.4.

Theorem 7.1. *The least squares estimates for $\hat{\beta}_1$ and $\hat{\beta}_0$ are given by, respectively,*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{m}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{m}(X_i - \overline{X})^2} = \frac{S_{xy}}{S_{xx}} , \tag{7.5}$$

and

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} . \tag{7.6}$$

Equation 7.4 is a function of $\hat{\beta}_0$ and $\hat{\beta}_1$ from the definition of the residuals Equation 7.3. Then Equation 7.5 and Equation 7.6 follow by equating the partial derivatives of Equation 7.4 to zero. The $\hat{\beta}_0$ and $\hat{\beta}_1$ are the unique solution to this linear system.

> **Alternative lingo**
>
> The RSS is sometimes referred to as the error sum of squares and abbreviated SSE (no, the order is not a typo).

Example 7.1. In Figure 7.1 and Figure 7.2, we consider the **Cherry Tree Data** (see Table 3.1) and discussion). We fit a least squares regression of timber volume (response variable) to the tree's diameter (independent variable). As you would expect, the timber yield increases with diameter.

The R code below can be used to calculate the least squares regression and residuals.

```
data(trees)
y <- trees$Volume
x <- trees$Girth # NB: this is the diameter; data mislabeled!
fit <- lm(y ~ x)
e <- resid(fit)
yhat <- predict(fit)
```

The `fit` data frame contains the estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$:

```
fit$coefficients
```

```
(Intercept)           x
 -36.943459    5.065856
```

Both Figure 7.1 and Figure 7.2 are scatter plots of the observed values y. In Figure 7.1, the regression line $\hat{y}$ is plotted along with the residuals $\hat{\epsilon}$. In Figure 7.2, the sample mean $\overline{y}$ is plotted together with the deviations $y - \overline{y}$.

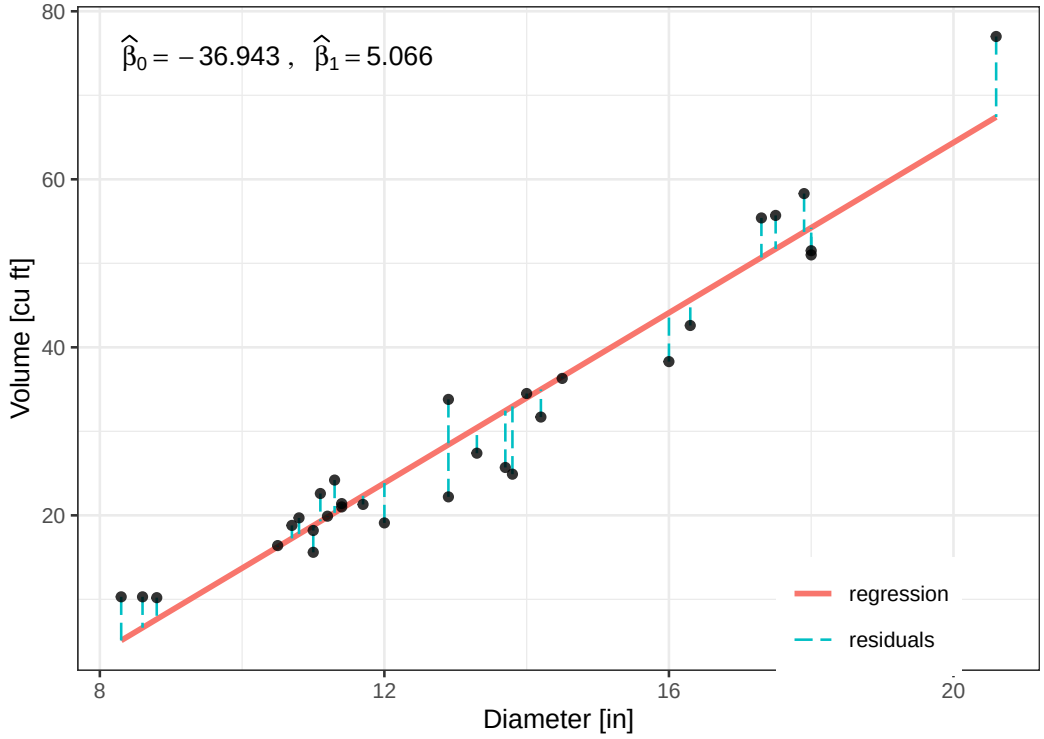


Figure 7.1: Linear regression (or least squares fit) of Volume to Diameter from the **Cherry Tree Data**. The vertical bars between the observed data point and the regression line indicate the error in the fit (the least squares residual). The residuals are squared and summed to yield the RSS (alt: SSE).

7.2 Estimating σ^2 for linear regressions

The parameter σ^2 (the variance of the random deviation) determines the variability in the regression model.

Theorem 7.2. *An unbiased estimate of σ^2 is given by*

$$\widehat{\sigma}^2 = s^2 = \frac{\text{RSS}}{m-2} = \frac{1}{m-2}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \,. \tag{7.7}$$

In Figure 7.3, we present a least squares regression of timber volume on both tree diameter and height (for the **Cherry Tree Data**). As expected, the regressions indicate the volume increases with both covariates. Estimates for the variance of the random deviation Equation 7.7 in both regression models, σ^2_D and σ^2_H, respectively, are computed to be $s^2_D = 18.08$ and $s^2_H = 179.48$. Thus, we see that small variances lead to observations of (x_i, y_i) that sit tightly around the regression line, in contrast to large variances that lead to a large cloud of points.

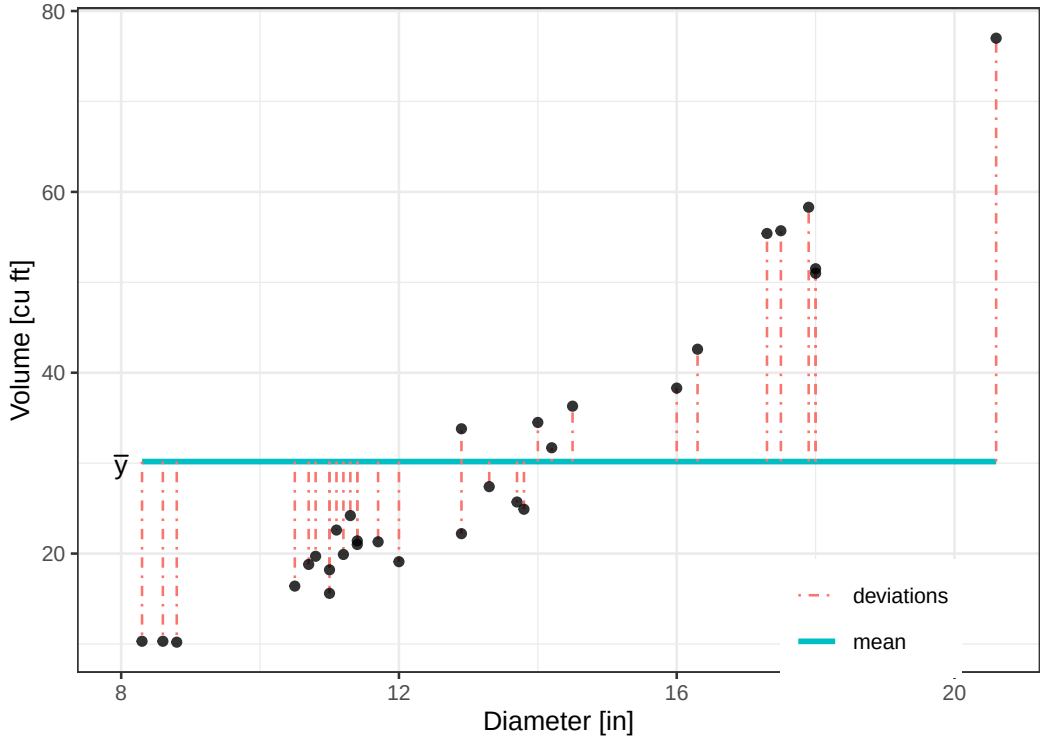


Figure 7.2: The deviations about the sample mean $\bar{y}$. The sum of the squared deviations or SST (total sum of squares) is a measure of the total variation in the observations.

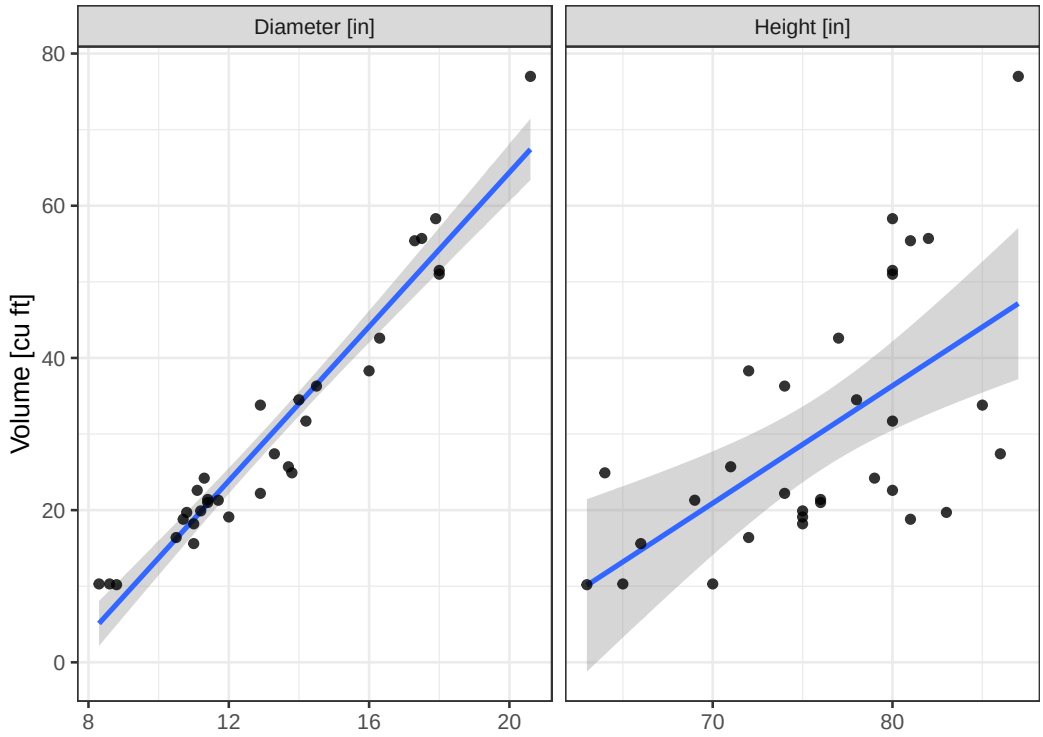


Figure 7.3: For the **Cherry Tree Data**, we estimate the variance to be $s_D^2 = 18.08$ (for Diameter) and $s_H^2 = 179.48$ (for Height); small variances lead to observations of (x_i, y_i) that sit tightly around the regression line, in contrast to large variances that lead to a large cloud of points.

> ⚠ Why do we lose two degrees of freedom?
>
> In Theorem 7.2, the number in the denominator is the df associated with the RSS and s^2. To calculate RSS, you must estimate two parameters β_0 and β_1, which results in the loss of two df. Hence the $m-2$.

We note to make inferences, the statistic

$$S^2 = \frac{\text{RSS}}{m-2}$$

is an unbiased estimator of σ^2 and the random variable

$$\frac{(m-2)S^2}{\sigma^2} \sim \chi^2(m-2)\,.$$

Moreover, the statistic S^2 is independent of both $\hat{\beta}_0$ and $\hat{\beta}_1$.

7.3 Inferences for least-squares parameters

If ϵ_i in Equation 7.2 is assumed to be normally distributed, then we can derive the sampling distributions of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. Hence, we can use these sampling distributions to make inferences about the parameters β_0 and β_1.

Provided iid $\epsilon_i \mid X_i \sim \mathsf{N}(0, \sigma^2)$, the least-squares estimators possess the following properties.

1. Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed.
2. Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased, i.e., $\mathbf{E}[\hat{\beta}_i] = \beta_i$ for $i = 0, 1$.
3. $\text{Var}[\hat{\beta}_0] = c_{00}\sigma^2$ where $c_{00} = \sum_{i=1}^{m} x_i^2/(mS_{xx})$.
4. $\text{Var}[\hat{\beta}_1] = c_{11}\sigma^2$ where $c_{11} = 1/S_{xx}$.
5. $\text{Cov}[\hat{\beta}_0, \hat{\beta}_1] = c_{01}\sigma^2$ where $c_{01} = -\overline{x}/S_{xx}$.

These properties can be determined by working directly from Equation 7.5 and Equation 7.6.

Proposition 7.1. *Consider $H_0 : \beta_i = \beta_{i0}$. The test statistic is*

$$T = \frac{\hat{\beta}_i - \beta_{i0}}{S\sqrt{c_{ii}}}\,. \tag{7.8}$$

For a hypothesis test at level α, we use the following procedure:

If $H_a : \beta_i > \beta_{i0}$, then P-value is the area under $\mathsf{t}(m-2)$ to the right of t.

If $H_a : \beta_i < \beta_{i0}$, tthen P-value is the area under $\mathsf{t}(m-2)$ to the left of t.

If $H_a : \beta_i \neq \beta_{i0}$, then P-value is twice the area under $\mathsf{t}(m-2)$ to the right of $|t|$.

A confidence interval for β_i, based on the statistic Equation 7.8, can be given following the procedures in Chapter 4.

Proposition 7.2. *A $100(1-\alpha)\%$ CI for β_i is given by*

$$\hat{\beta}_i \pm t_{\alpha/2,m-2} S\sqrt{c_{ii}}\,.$$

7.4 Correlation

Let $(X_1, Y_1), \ldots, (X_m, Y_m)$ denote a random sample from a bivariate normal distribution with $\mathbf{E}[X_i] = \mu_X$, $\mathbf{E}[Y_i] = \mu_Y$, $\text{Var}[X_i] = \sigma_X^2$, $\text{Var}[Y_i] = \sigma_Y^2$, and correlation coefficient ρ. The sample correlation coefficient is given by,

$$r = \frac{\sum_{i=1}^{m}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{m}(X_i - \overline{X})^2 \sum_{i=1}^{m}(Y_i - \overline{Y})^2}}, \tag{7.9}$$

which can be rewritten in terms of S_{xx}, S_{xy}, and S_{yy}:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \hat{\beta}_1 \sqrt{\frac{S_{xx}}{S_{yy}}},$$

using Equation 7.5 and we see that r and $\hat{\beta}_1$ have the same sign. A $|r|$ close to 1 means that the regression line is a good fit to the data, and, similarly, an $|r|$ close to 0 means a poor fit to the data. Note that the correlation coefficient (and the least squares regression) are only suitable for describing *linear* relationships; a nonlinear relationship can also yield r near zero (see Figure 7.4).

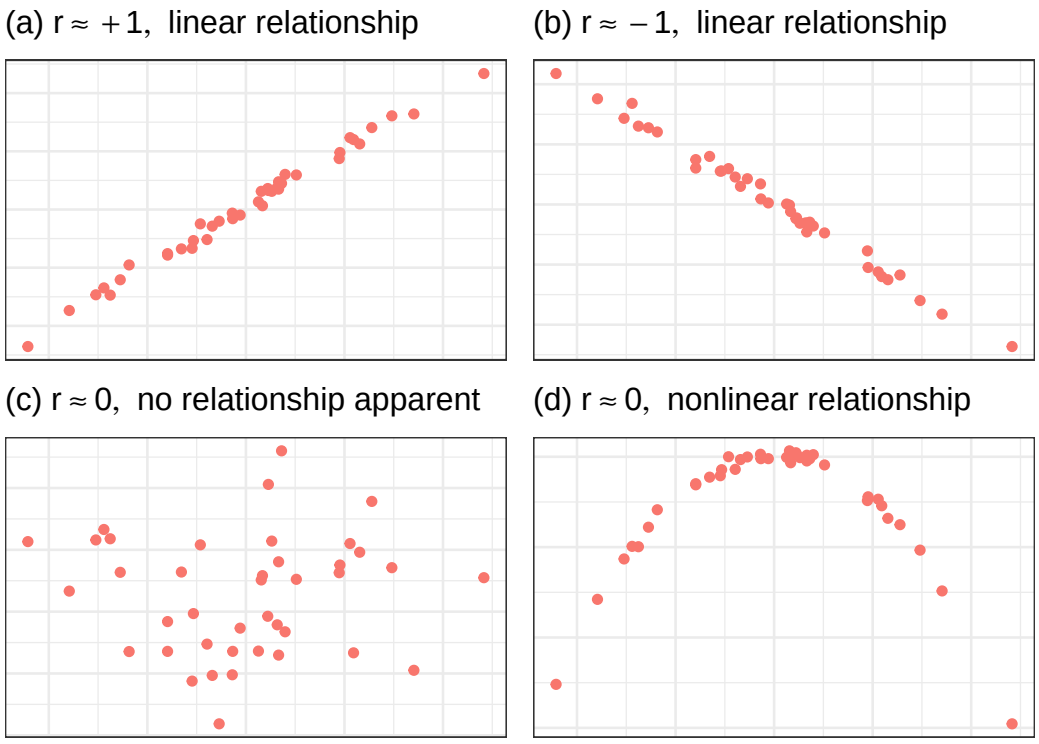


Figure 7.4: Correlations range from −1 to 1 with $|r| = 1$ indicating a strong linear relationship and r near zero indicating the absence of a linear relationship.

7.5 Prediction using linear models

Once a model is fit, it can be used to predict a value of y for a given x. However, the model only gives the most likely value of y; a corresponding prediction interval is usually more appropriate.

Proposition 7.3. *A* $100(1-\alpha)\%$ *prediction interval for an actual value of* Y *when* $x = x^*$ *is given by*

$$(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\alpha/2,m-2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}\,.$$

> **Prediction versus confidence intervals**
>
> The prediction interval is different from the confidence interval for expected Y. Note that the length of the *confidence interval* for $\mathbf{E}[Y]$ when $x = x^*$ is given by
>
> $$2 \cdot t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}$$
>
> whereas the length for the *prediction interval* of Y is
>
> $$2 \cdot t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}\,.$$
>
> Thus the prediction intervals for an actual value of Y are longer than the confidence intervals for $\mathbf{E}[Y]$ if both are determined for the same value x^*.

The linear model

$$\mathbf{E}[Y \mid X = x] = \beta_0 + \beta_1 x\,,$$

assumes that the conditional expectation of Y for a fixed value of X is a linear function of the x value. If we assume that (X, Y) has a bivariate normal distribution, then

$$\beta_1 = \frac{\sigma_Y}{\sigma_X}\rho\,,$$

and thus, for the simple hypothesis tests we have considered (Table 3.2), statistical tests for β_1 and ρ are equivalent.

8 Categorical data

8.1 Multinomial experiments

Suppose we have a population divided into $k > 2$ distinct categories. We consider an experiment where we select m individuals (or objects) from the population and categorise each. We denote the population proportion in the ith category by p_i. If the sample size m is much smaller than the population size M (so that the m trials are independent), this experiment will be approximately *multinomial* with success probability p_i for each category, $i = 1, \dots, k$.

Before the experiment is performed, we denote the number (or count) of the trials resulting in category i by the rv N_i. The expected number of trials that result in category i is given by

$$\mathbf{E}[N_i] = mp_i\,, \quad i = 1, \dots, k\,. \tag{8.1}$$

After the experiment is performed, we denote the corresponding observed value by n_i. Since the trials result in distinct categories,

$$\sum_{i=1}^{k} N_i = \sum_{i=1}^{k} n_i = m\,,$$

which indicates that, for a given m, we only need to observe $k - 1$ of the variables to be able to work out what the kth variable should be.

8.2 Goodness-of-fit for a single factor

We are interested in making inferences about the proportion parameters p_i. Specifically, we will consider the null hypothesis,

$$H_0 : p_1 = p_{10}\,, p_2 = p_{20}\,, \cdots\,, p_k = p_{k0}\,, \tag{8.2}$$

that completely specifies a value p_{i0} for each p_i. The alternative hypothesis H_a will state that H_0 is not true, i.e., that at least one p_i is different from the value p_{i0} claimed under the null H_0.

> **Notation**
>
> Here for $i = 1, \dots, k$ we use the notation p_{i0} to denote the value of p_i claimed under the null hypothesis.

Provided the null hypothesis in Equation 8.2 is true, the expected values Equation 8.1 can be written in terms of the expected frequencies,

$$\mathbf{E}[N_i] = mp_{i0}\,, \quad i = 1, \dots, k\,.$$

Often the n_i, referred to as the observed cell counts, and the corresponding mp_{i0}, referred to as the expected cell counts, are tabulated, for example, as in Table 8.1.

Table 8.1: Observed and expected cell counts.

Category	$i = 1$	$i = 2$	$\cdots$	$i = k$	Row total
Observed	n_1	n_2	$\cdots$	n_k	m
Expected	mp_{10}	mp_{20}	$\cdots$	mp_{k0}	m

The test procedure assesses the discrepancy between the value of the observed and expected cell counts. This discrepancy, or goodness of fit, is measured by the squared deviations divided by the expected count.

> **Why divide by expected cell counts?**
>
> The division by the expected cell counts accounts for possible differences in the relative magnitude of the observed/expected counts.

Theorem 8.1. *For $mp_i \geq 5$ for $i = 1, \ldots, k$, the rv*

$$V = \sum_{i=1}^{k} \frac{(N_i - mp_i)^2}{mp_i} \quad \sim \chi^2(k-1)\,,$$

that is, V has approximately a χ^2 distribution with $\nu = k - 1$ df.

Proposition 8.1. *Consider the null*

$$H_0 : p_1 = p_{10}, p_2 = p_{20}, \cdots, p_k = p_{k0}\,,$$

and the alternative

$$H_a : p_i \neq p_{i0} \text{ for at least one } i\,.$$

The test statistic is

$$V = \sum_{i=1}^{k} \frac{(N_i - mp_{i0})^2}{mp_{i0}}\,.$$

As a rule of thumb, provided $mp_{i0} \geq 5$ for all $i = 1, \ldots, k$, then the P-value is the area under $\chi^2(k-1)$ to the right of v.

If $mp_{i0} < 5$ for some i then it may be possible to combine the categories such that the new categorizations satisfy the assumptions of Proposition 8.1.

> **What about partial information?**
>
> Things are much more complicated if the category probabilities are not entirely specified.

8.3 Test for the independence of factors

In Section 8.2, we considered categorising a population into a single factor. We now consider a single population where each individual is categorised into two factors with I distinct categories for the first factor and J distinct categories for the second factor. Each individual from the population belongs to exactly one of the I categories of the first factor and exactly one of the J categories of the second factor. We want to determine whether or not there is any dependency between the two factors.

For a sample of m individuals, we denote by n_{ij} the count of the m samples that fall both in category i of the first factor and category j of the second factor, for $i = 1, \dots, I$ and $j = 1, \dots, J$. A contingency table with I rows and J columns (i.e., IJ cells) will be used to record the n_{ij} counts (in an obvious way). Let p_{ij} be the proportion of individuals in the population who belong in category i of factor 1 and category j of factor 2. Then, the probability that a randomly selected individual falls in category i of factor 1 is found by summing over all j:

$$p_i = \sum_{j=1}^{J} p_{ij} \,,$$

and likewise, the probability that a randomly selected individual falls in category j of factor 2 is found by summing over all i:

$$p_j = \sum_{i=1}^{I} p_{ij} \,.$$

The null hypothesis that we will be interested in adopting is

$$H_0 : p_{ij} = p_i \cdot p_j \ \forall (i, j) \,, \tag{8.3}$$

that is, an individual's category in factor 1 is independent of the category in factor 2.

Following the same program as for the single category goodness-of-fit test, we note that assuming the null hypothesis Equation 8.3 is true, then the expected count in cell i, j is

$$\mathbf{E}[N_{ij}] = mp_{ij} = mp_i p_j \,;$$

and we estimate p_i and p_j by the appropriate sample proportion:

$$\hat{p}_i = \frac{n_i}{m} \,, \qquad n_i = \sum_j n_{ij} \quad \text{(row totals)} \,,$$

and

$$\hat{p}_j = \frac{n_j}{m} \,, \qquad n_j = \sum_i n_{ij} \quad \text{(column totals)} \,.$$

Thus, the expected cell count is given by

$$\hat{e}_{ij} = m\hat{p}_i\hat{p}_j = \frac{n_i n_j}{m} \,,$$

and we assess the goodness of fit between the observed cell count n_{ij} and the expected cell count $\hat{e}_{ij}$.

Proposition 8.2. *Assume the null hypothesis*

$$H_0 : p_{ij} = p_i p_j \text{ for all } i = 1, \dots, I \,, j = 1, \dots, J \,,$$

against the alternative hypothesis

$$H_a : H_0 \text{ is not true} \,.$$

The test statistic is

$$V = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(N_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} .$$

As a rule of thumb, provided $\hat{e}_{ij} \geq 5$ for all i, j and when H_0 is true, then the test statistic has approximately a $\chi^2(\nu)$ distribution with $\nu = (I-1)(J-1)$ df. For a hypothesis test at level α, the procedure is upper-tailed, and the P-value is the area under $\chi^2(\nu)$ to the right of v.

> **Alternative lingo**
>
> Contingency is just another word for dependency in the context of goodness-of-fit tables.

9 Quality control

Quality control is an area of applied statistics that makes interventions to maintain or improve the outcome of industrial processes. Random variations in output processes might negatively impact the quality of a product. We want to identify the sources of random output-process variations that might have *assignable causes*. Control charts are a tool that helps us to recognise when industrial processes are no longer controlled so that one might then seek to identify assignable causes.

9.1 Control charts

The essential elements of control charting involve specifying a control region and then analysing time-series data. We will specify a baseline value along with an upper and lower control limit and assume that a process is under control unless a test statistic suggests otherwise. To construct a control chart, one collects data about a process at fixed points of time and calculates the running value of a quality statistic. Suppose the quality statistic exceeds the upper or lower control limits. In that case, the process is deemed out of control, and the product quality is assumed to be negatively impacted.

> **!** Default position
>
> The default position adopted for quality control will be reminiscent of hypothesis testing: "assume that a process is under control unless a test statistic suggests otherwise."

The process of creating a control chart is best illustrated through an extended example, like Example 9.1 provided below.

Example 9.1. Here we consider the typical 3σ control charting for a process mean $\overline{X}$ based on estimated parameters. That is, we assume the generating process X is normally distributed with unknown parameters μ and σ^2. We seek to estimate the mean $\overline{X}$. Our control region is specified to be three standard deviations; the process is in control if it remains within three standard deviations of a baseline value.

> **i** Note 4: Beer Production Data
>
> The **Beer Production Data** contains measurements of the features OG, ABV, pH, and IBU for 50 batches of each of three types of product (Premium Lager, IPA, and Light Lager).

```
beer |> glimpse()
```

```
Rows: 150
Columns: 6
$ Batch_Id <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ~
$ OG       <dbl> 5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9, 5.4, 4.8, 4.8, 4.3, 5~
$ ABV      <dbl> 3.5, 3.0, 3.2, 3.1, 3.6, 3.9, 3.4, 3.4, 2.9, 3.1, 3.7, 3.4, 3.0, 3.0, 4~
$ pH       <dbl> 1.4, 1.4, 1.3, 1.5, 1.4, 1.7, 1.4, 1.5, 1.4, 1.5, 1.5, 1.6, 1.4, 1.1, 1~
```

```
$ IBU     <dbl> 9.0, 10.0, 7.0, 9.0, 8.0, 7.7, 7.4, 7.1, 6.8, 6.5, 6.2, 5.9, 5.6, 5.3, ~
$ Beer    <chr> "Premium Lager", "Premium Lager", "Premium Lager", "Premium Lager", "Pr~
```

Let's consider the Beer Production Data in Note 4. We are interested in the IPA's pH value, which influences saccharification. We assume that three batches of IPA are produced per day, and we prepare the data as follows.

```
ipa <- beer |>
 select(Batch_Id, pH, Beer) |>
 filter(Beer == "IPA") |>
 rename(Day = Batch_Id)

m <- 3    # three batches per day
k <- 16   # number of days
ipa$Day[1:(m*k)] <- rep(1:k, each = m)
ipa <- ipa[1:(m*k),]
```

The prepared data, `ipa`, is summarized in the Table 9.1.

```
ipa_stat <- ipa |>
 group_by(Day) |>
 summarise(obs = list(pH), mean = signif(mean(pH), digits = 4),
           sd = signif(sd(pH), digits = 4), range = max(pH) - min(pH))
ipa_stat |>
 kbl(align = "rcccc", booktabs = T, escape = F) |>
 kable_styling(latex_options = c("striped"))
```

We first observe that the pH measurements are (at least approximately) normal, as seen in the quantile-quantile plot in Figure 9.1.

```
ipa |> ggplot(aes(sample = pH)) + stat_qq() + stat_qq_line()
```

Table 9.1: Observations and summary statistics for the **Beer Production Data**.

Day	obs	mean	sd	range
1	4.7, 4.5, 4.9	4.700	0.20000	0.4
2	4.0, 4.6, 4.5	4.367	0.32150	0.6
3	4.7, 3.3, 4.6	4.200	0.78100	1.4
4	3.9, 3.5, 4.2	3.867	0.35120	0.7
5	4.0, 4.7, 3.6	4.100	0.55680	1.1
6	4.4, 4.5, 4.1	4.333	0.20820	0.4
7	4.5, 3.9, 4.8	4.400	0.45830	0.9
8	4.0, 4.9, 4.7	4.533	0.47260	0.9
9	4.3, 4.4, 4.8	4.500	0.26460	0.5
10	5.0, 4.5, 3.5	4.333	0.76380	1.5
11	3.8, 3.7, 3.9	3.800	0.10000	0.2
12	5.1, 4.5, 4.5	4.700	0.34640	0.6
13	4.7, 4.4, 4.1	4.400	0.30000	0.6
14	4.0, 4.4, 4.6	4.333	0.30550	0.6
15	4.0, 3.3, 4.2	3.833	0.47260	0.9
16	4.2, 4.2, 4.3	4.233	0.05774	0.1

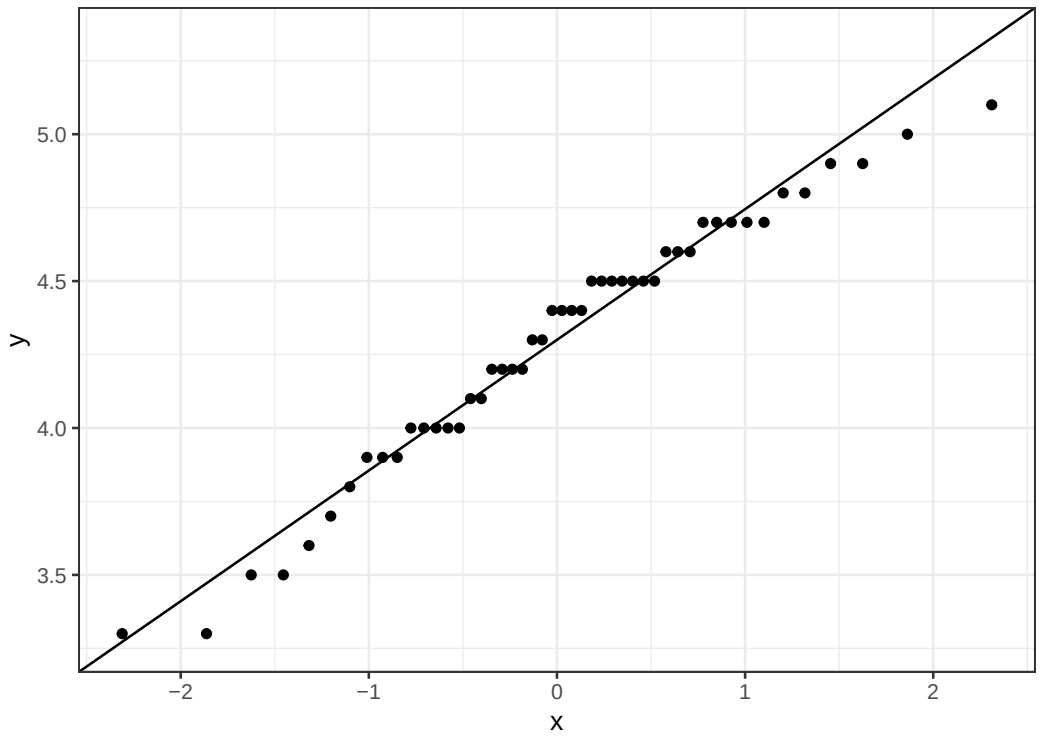


Figure 9.1: Normal quantile-quantile plot of observed pH measurements of the IPA batches.

We consider the data for pH readings from three batches of IPA taken over sixteen days ($k = 16$) presented in Table 9.1. The Table includes the sample mean per day, $\overline{x}$, the sample standard deviation, s, and the range of values per day, $\max x_i - \min x_i$ (each based on $m = 3$ batches).

We estimate the mean

$$\widehat{\mu} = \frac{1}{k}\sum_{i=1}^{k} \overline{x}_i \,,$$

by averaging the means found for the k days and, similarly, estimating the mean of the sample standard deviation,

$$\overline{s} = \frac{1}{k}\sum_{i=1}^{k} s_i \,,$$

by averaging the sample standard deviations for the k days. It can be shown that

$$\widehat{\sigma} = \frac{\overline{S}}{a_m}$$

is an unbiased estimator of σ where

$$a_m = \frac{\sqrt{2}\Gamma(m/2)}{\sqrt{m-1}\Gamma\left((m-1)/2\right)} \,.$$

Thus, we compute the 3σ upper and lower control limits, respectively,

$$\text{UCL} = \widehat{\mu} + 3\frac{\overline{s}}{a_m\sqrt{m}}$$

and

$$\text{LCL} = \widehat{\mu} - 3\frac{\overline{s}}{a_m\sqrt{m}} \,.$$

The computations in R follow, along with the resulting control chart in Figure 9.2.

```
a <- function(m){ sqrt(2) * gamma(m/2) / (sqrt(m-1) * gamma((m-1)/2)) }
muhat = sum(ipa_stat$mean) / k
sbar = sum(ipa_stat$sd) / k
lcl = muhat - 3*sbar / (a(m) * sqrt(m))
ucl = muhat + 3*sbar / (a(m) * sqrt(m))

ggplot(ipa_stat, aes(x = Day)) + geom_point(aes(y = mean)) +
 geom_hline(aes(yintercept = muhat, color = "Mean"), linewidth = lsz) +
 geom_hline(aes(yintercept = lcl, color = "LCL"), linewidth = lsz*1.5) +
 geom_hline(aes(yintercept = ucl, color = "UCL"), linewidth = lsz*1.5) + ylab("pH") +
   theme(legend.justification = c(1,1), legend.position = c(0.9,0.9),
         legend.title = element_blank(),
         legend.box.margin = margin(c(4, 4, 4, 4), unit = "pt"))
```

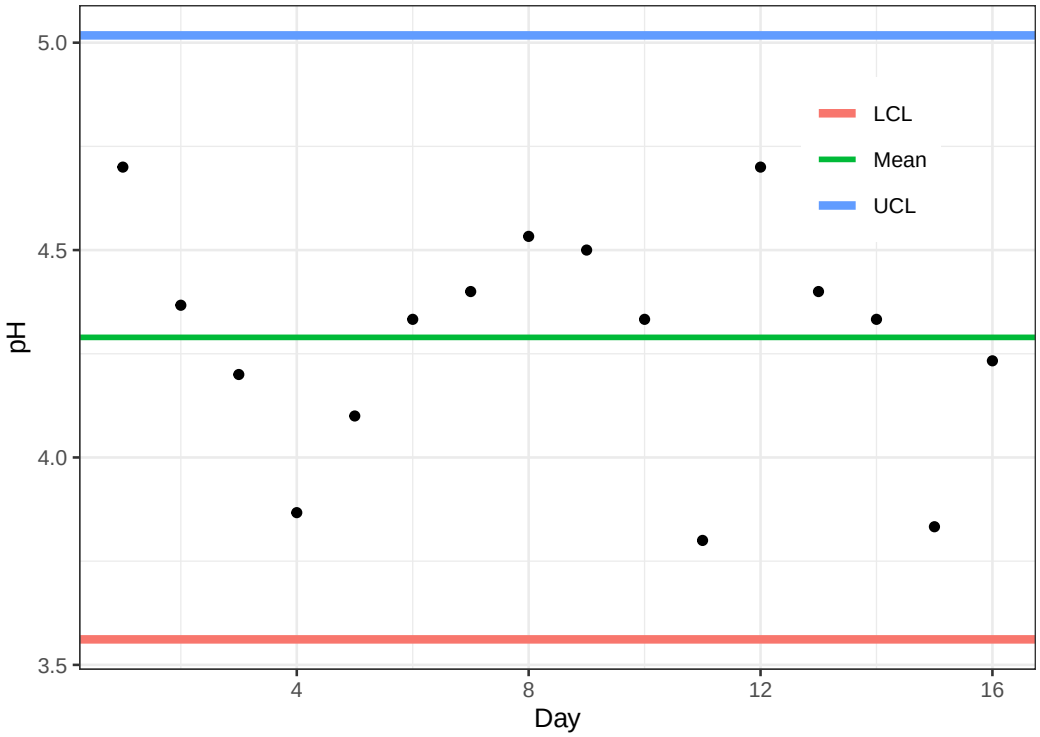


Figure 9.2: The 3σ control chart illustrates that with respect to pH the brewing process is in-control over the selected timeframe as the observations fall within the $(\mathsf{LCL}, \mathsf{UCL})$ control interval.

From Figure 9.2, we observe for each day the process is in-control as the observed mean pH values fall within the control limits $(\mathsf{LCL}, \mathsf{UCL})$. If this were not the case, our initial assumption that the process is in control would be violated. The violation of the assumption would require that we seek to identify an assignable cause for the variation. If a cause could be identified, we would need to recompute our control limits with the observations that were out of control removed.

References

Belle, Gerald van. 2008. *Statistical Rules of Thumb*. Second. Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons, Inc.

Spiegelhalter, David J. 2020. *The Art of Statistics: Learning from Data*. London: Pelican Books.

Wasserman, Larry. 2004. *All of Statistics*. New York: Springer-Verlag.

Curated Content

Below we provide links to supplementary online material. Hopefully, some of the items will inspire you to view the module material in a broader context and lead to further investigations.

Investigation 1

What is Statistics?

- **Cambridge Ideas - Professor Risk**
  https://www.youtube.com/watch?v=a1PtQ67urG4
  Prof David Spiegelhalter (Cambridge University) discusses public understanding of risk. You may also be interested in reading (Spiegelhalter 2020).
- **The Joy of Statistics**
  https://www.youtube.com/watch?v=jbkSRLYSojo
  Prof Hans Rosling (Karolinska Institute and Gapminder Foundation) analyses data from 200 Countries over 200 Years in 4 Minutes - The Joy of Stats - BBC Four.
- **Teach statistics before calculus!**
  https://www.ted.com/talks/arthur_benjamin_teach_statistics_before_calculus
  Prof Arthur Benjamin (Harvey Mudd College) argues that the pinnacle of math education is probability and statistics — not calculus.
- **Kaggle**
  https://www.kaggle.com/
  Towards data science.
  https://www.youtube.com/watch?v=TNzDMOg_zsw
  What's Kaggle?

Investigation 2

Defence against the dark arts.

- **Three ways to spot bad statistics**
  https://www.ted.com/talks/mona_chalabi_3_ways_to_spot_a_bad_statistic
  Mona Chalabi (Data Journalist) discusses three ways to spot bad statistics.
- **Statistics Done Wrong**
  https://www.statisticsdonewrong.com/
  A book by Dr Alex Reinhart (Carnegie Mellon University).

- **How to defend yourself against misleading statistics in the news**
  https://www.youtube.com/watch?v=mJ63-bQc9Xg
  Sanne Blauw (Journalist) discusses how the presentation of statistics can mislead.

Investigation 3

Data analysis and visualisation.

- **The Grammar of Graphics**
  https://www.youtube.com/watch?v=h-62NwWUI5c
  What Makes A Good Visualisation? Rhys Jackson from RocketMill, a UK Digital Marketing Agency, gives a perspective on visualising data from a marketing perspective.
  https://www.youtube.com/watch?v=kepKM7Z2O54
  David Keyes (RStudio) discusses how the grammar of graphics underpins the `ggplot2` data visualization package in `R`.
- **Same Stats, Different Graphs**
  https://www.autodeskresearch.com/publications/samestats
  Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing (ACM SIGCHI Conference on Human Factors in Computing Systems) by Justin Matejka, George Fitzmaurice.
- **Why do we so often use 0.05 for hypothesis testing?**
  https://www.openintro.org/book/stat/why05/
  In this online exercise, you will gain an improved understanding of what a significance level is, and why a value in the neighbourhood of 0.05 is reasonable as a default.
- **Data visualisations**
  https://flowingdata.com/
  FlowingData blog by Nathan Yau.
  https://fivethirtyeight.com/
  FiveThirtyEight blog by Nate Silver.
- **Storytelling with data**
  http://www.storytellingwithdata.com/blog
  Blog with nice hints and tips for how to present data in tables, graphics, and visualisations.
  https://community.storytellingwithdata.com/challenges
  Monthly challenge.

Investigation 4

Statistical paradoxes.

- **How statistics can be misleading (TED-Ed)**
  https://www.ted.com/talks/mark_liddell_how_statistics_can_be_misleading
  Mark Liddell (Educator) discusses Simpson's Paradox in this TED-Ed animation.
- **Low birth-weight paradox**
  https://www.wikiwand.com/en/Low_birth-weight_paradox

- **Gambler's Fallacy**
  https://www.youtube.com/watch?v=4eVluL-idkM
  Prof Kelly Shue (Chicago Booth) discusses the gambler's fallacy.

Investigation 5

The law and interpreting statistics.

- **How stats fool juries.**
  https://youtu.be/kLmzxmRcUTo
  Prof Peter Donnelly (Oxford University) discusses common mistakes in interpreting statistics.

- **Measurement Uncertainty Calculator (MUCalc)**
  https://discovery.dundee.ac.uk/en/publications/measurement-uncertainty-calculator-mucalc
  The Leverhulme Research Centre for Forensic Science Measurement Uncertainty Calculator (MUCalc) is an application for calculating measurement uncertainty in accordance with the standards of International Organization for Standardization ISO/IEC 17025.

- **Prosecutor's fallacy**
  https://www.wikiwand.com/en/Prosecutor%27s_fallacy
  A fallacy of statistical reasoning, typically used by a prosecutor to exaggerate the likelihood of guilt: because P(hypothesis | evidence) $\neq$ P(evidence | hypothesis)!

Investigation 6

Data-driven decision making in epidemiology.

- **Project Tycho**
  https://www.tycho.pitt.edu/
  Digitized archival epidemiological data for the United States and the world.
  https://www.youtube.com/watch?v=Kn9OJy1BPDo
  An overview of the origins of project Tycho.

- **Our World in Data**
  https://ourworldindata.org/
  A project of the Oxford Martin School to make public health data, including progress in UN Sustainable Development Goals, available and accessible.

- **Demographic Party Trick**
  https://www.youtube.com/watch?v=2nDh8MQuS-Y
  Prof Hans Rosling (Karolinska Institute and Gapminder Foundation) and Bill Gates seek to shed light on the true statistics of childhood vaccinations.

Investigation 7

Spurious correlations!

- **The danger of mixing up causality and correlation**
  https://www.youtube.com/watch?v=8B271L3NtAw
  Prov Ionica Smeets (University of Leiden) discusses causality and correlation.
- **Spurious correlations**
  https://tylervigen.com/spurious-correlations
  Tyler Vigen's site dedicated to spurious correlations.
- **Cause & Effect**
  https://www.youtube.com/watch?v=lbODqslc4Tg
  Correlation vs. causality from the Clip from the 2010 documentary "Freakonomics: The Movie".

Investigation 8

Data and Society: can data-driven and predictive modelling lead to a better world? What are the ethics of mass data collection?

- **Science behind the news: Predictive Policing**
  https://www.youtube.com/watch?v=74_jreara3w
  The Los Angeles Police Department is using a new tactic in their fight against crime called "predictive policing." It's a computer program originally developed by a team at UCLA, including mathematician Andrea Bertozzi and anthropologist Jeff Brantingham. "Science Behind the News" is produced in partnership with NBC Learn. (Provided by the National Science Foundation & NBC Learn)
- **You should get paid for your data**
  https://www.nytimes.com/video/opinion/100000006678020/data-privacy-jaron-lanier-2.html
  Jaron Lanier (Computer Scientist and Author) discusses a compensation plan and data dignity.
  https://www.ted.com/talks/jennifer_zhu_scott_why_you_should_get_paid_for_your_data
  Jennifer Zhu Scott (Computer Scientist) also thinks you should get paid for your data.
- **How tech companies deceive you into giving up your data and privacy**
  https://www.ted.com/talks/finn_lutzow_holm_myrstad_how_tech_companies_deceive_you_into_giving_up_your_data_and_privacy
  Finn Lützow-Holm Myrstad (Norwegian Consumer Council) discusses consumer protections and data collection.
- **Your company's data could help end world hunger**
  https://www.ted.com/talks/mallory_freeman_your_company_s_data_could_help_end_world_hunger
  Mallory Freeman (Data Scientist) discusses how to do the most good with data.

Investigation 9

Machine learning / big data.

- **What is Machine Learning?**
  https://www.youtube.com/watch?v=f_uwKZIAeM0
  OxfordSparks discusses the topic of supervised learning algorithms and how machine learning is used all around us.
- **Big Data (TED-Ed)**
  https://www.youtube.com/watch?v=j-0cUmUyb-Y
  Tim Smith (educator) discusses the historical arc of big data in this TED-Ed animation.
- **The human insights missing from big data**
  https://www.ted.com/talks/tricia_wang_the_human_insights_missing_from_big_data
  Tricia Wang (Ethnographer) discusses the human insights missing from big data.
- **How we can find ourselves in data**
  https://www.ted.com/talks/giorgia_lupi_how_we_can_find_ourselves_in_data
  Giorgia Lupi (Designer) discusses a humanistic approach to data and data visualization.