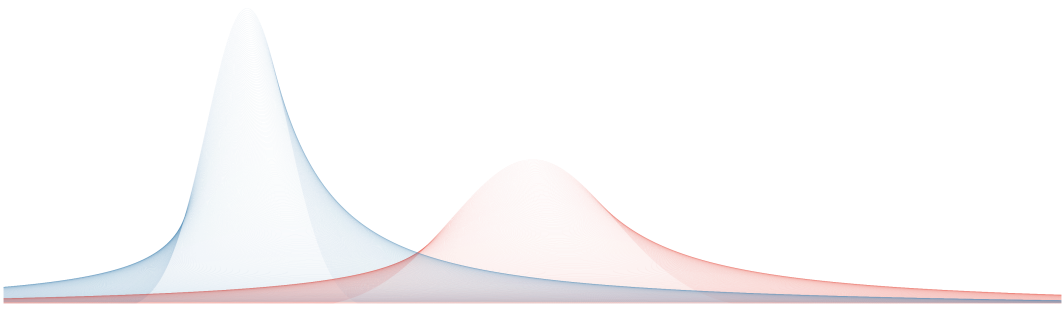
Frans Rodenburg

Introduction to Biostatistics



Contents

Preface

This book aims to guide the reader through the basics of drawing conclusions from observations. Biology is an empirical science, which means that evidence comes from measuring, recording, or otherwise observing phenomena. Evidence from observations comes with uncertainty and the field of statistics is concerned with expressing that uncertainty.

Statistics is a crucial part of your education as a scientist. Besides learning to analyze data, it enables you to make informed decisions on experimental design and be more critical of scientific literature. Even if you will not perform your own analyses, a good understanding of statistics will translate to a better understanding of empirical science as a whole.

Biostatistics is not fundamentally different from mathematical statistics, machine learning, AI, psychometrics, or econometrics. The main difference lies in the techniques that you will be using the most. For biology and other life sciences, many research questions can be answered through some form of regression analysis. This is therefore the main subject of the second book: Elements of Biostatistics (hereafter referred to as ***Elements***). Before delving into regression analysis, this book covers the necessary background knowledge, along some basic statistical tests and models. In the final book, Advanced Biostatistics (***Advanced***), the previous models are further expanded upon, and predictive modeling is introduced.

- **Some paragraphs are denoted by an asterisk (*). These serve for completeness, correctness, or clarification and are non-essential.**
- **A list of recommended literature is included in appendix *B*.**
- **Feel free to print a copy of this file for personal use.**
- **Please address comments to: biostatistics.bookseries@gmail.com**

Use of Software

This book is accompanied by the syllabus *Statistics in R and RStudio*. You may be familiar with Excel or other software for basic calculations and plotting. However, as analyses becomes more complex, the limitations of software like Excel become increasingly prohibitive. On the contrary, R can be used for any analysis in your education and beyond. In this book, R code is displayed as follows:

```
1 + 1  # This is a comment, the output is shown below.
```

```
[1] 2
```

Those interested in neural networks are probably aware of the R/Python debate. If you are among those people, don't worry: (1) R's syntax is very similar to Python. Your skills in one language will transfer well to the other; (2) R has interfaces to Keras (`keras`), TensorFlow (`tensorflow`) and even to Python directly (`reticulate`).

Structure

The only techniques you need to learn in this book are in chapters *8* and *9*. However, not everyone has the same background in mathematics. Some readers may already be familiar with the basics of hypothesis testing, others may never have formulated a null-hypothesis before. Hence, the structure of this book is as follows:

1. *What is Statistics?* — General introduction
2. *Probability Theory* — Mathematical background 1
3. *Study Design* — Considerations *before* conducting experiments
4. *Exploring Data* — First steps *after* conducting experiments
5. *Estimation* — Mathematical background 2
6. *Distributions* — Mathematical background 3
7. *Hypothesis Testing* — Ideology behind a hypothesis test
8. *Statistical Tests* — Techniques 1: T-test, χ^2-test and F-test
9. *Statistical Models* — Techniques 2: ANOVA and linear regression

Motivation

This book was originally intended to be a syllabus for statistics course(s) for biologists at Leiden University, coordinated by Dr. H.G.J. van Mil. To avoid multiple literature resources, it quickly grew to encompass the background information on statistics as well.

1 What is Statistics?

The word statistics refers to abstractions of data (e.g. a mean, or a percentage), but also to the field of science concerned with data collection, summarization and interpretion. The goal of statistical analysis is to understand and express the strength of evidence for scientific arguments.

Science is the collection of knowledge and testable hypotheses. Some hypotheses can be proven with existing rules and knowledge, using a *deductive* argument. This is any argument where if its premises are true, the conclusion must also be true. For example (fig. 1.1):

- All birds reproduce through eggs (*premise*);
- All canaries are birds (*premise*);
- Therefore, all canaries reproduce through eggs (*conclusion*).

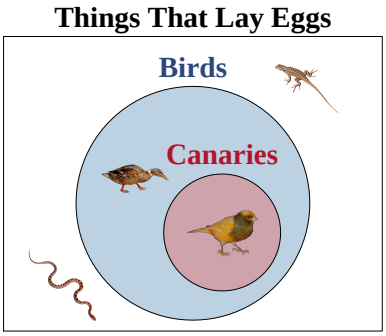


Figure 1.1: By being a member of birds, canaries automatically belong to things that lay eggs. If all birds lay eggs and all canaries are birds, then we need not check whether canaries lay eggs. This is a basic example of deductive reasoning.

Unfortunately, this playing around with definitions is not likely to lead to profound new insight in biology. Instead, most scientific arguments are *inductive*, such as the following:

- Lung cancer is relatively rare in non-smokers (*premise*);
- Lung cancer rates are relatively high in smokers (*premise*);
- Smoking increases the chance of lung cancer (*conclusion*).

Even if the premises are true, they do not guarantee the conclusion. Bizarre as it might be, all smokers with lung cancer might simply have gotten unlucky. Decades of **evidence** suggest otherwise, but no amount of evidence consitutes **proof**.

Experiments and observations are called empirical evidence. The goal of statistics is to express the strength of this evidence. How much evidence is enough to believe a conclusion, is entirely up to you. However, many fields of research have guidelines for what amounts to sufficient evidence to make a scientific claim.[1]

Moreover, when choosing between hypotheses, by believing that which is demonstrated to be the most likely, you are most likely to be correct. Scientific journals will reject conclusions that are not supported by the data. Speculations, assumptions and prior beliefs are allowed, but must be clearly described as such.

[1] In biology, measures of uncertainty of 5%, 1%, or less are commonly used.

1.1 Probability Theory (Ch. 2 & 6)

Choosing between hypotheses without irrefutable proof for one or the other, is done by finding the most *probable*. Doing so requires some understanding of what probability is. In particular, how chances propagate, how conditional probability works and what a probability distribution is, are the most essential pieces of background knowledge.

1.2 Study Design (Ch. 3)

Part of the reason why statistics is important for empirical science is to guide in setting up experiments. After all, why would you choose to perform something in duplo, in triplo, or even 10×, if not for statistical reasons? Understanding the sources of uncertainty, assigning treatments to replicates, and choosing an appropriate sample size are all statistical parts of study design.

1.3 Exploring Data (Ch. 4)

As long as no statistical tests or models with an outcome variable are used, the analysis can be considered purely exploratory. The goal of exploration is to gain insight in the data and is generally a first step in analysis: Incorrectly read data, outliers and missing values can be easily revealed through some simple plots and summaries. Moreover, almost all scientific reports contain summarizing tables and figures.

Exploratory research can also help form new ideas and hypotheses, by searching for patterns or potential associations. However, these cannot be supported by the same data used to form them, as this would be biased towards positive results. This is one reason why e.g. clinical trials involve multiple stages of independent research.

Examples of questions primarily concerned with exploration are:

- What is the effect of this new compound on the immune system?
- How do different species correlate in an ecosystem?
- What are the demographics of a certain condition?

When exploration goes beyond tables and figures, it is also known as *unsupervised learning*, meaning that no outcome variable is involved. Examples include principal component analysis (***Elements***), k-means and nearest neighbour clustering.

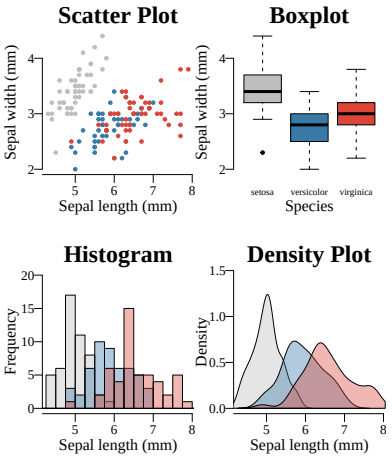


Figure 1.2: The bread and butter of any analysis. Try and see which plots are related. These plotting techniques are explained in detail in chapter ***4***.

1.4 Estimation (Ch. 5 & 9)

Estimation is calculation with uncertainty. Expressing the uncertainty of an estimate (e.g. through standard errors or confidence intervals) is paramount to conveying its precision.

Most statistical methods are *parametric*, meaning that a model can be defined in terms of parameters, like the mean and the variance[2]. Estimating those parameters is a first step for almost all statistical methods, whether used for statistical inference, confirmation, or prediction. However, if statistical inference is of primary interest, the quality of estimation is more important than statistical significance or prediction accuracy.

[2] Variance is a measure of dispersion: the spread of the data around the mean.

Examples of questions primarily concernced with estimation are:

- How large is the increased risk of cancer due to smoking?
- How tall do mice grow on each of six diets?
- What is the typical length and width of an *Iris* flower?

For statistical inference, careful attention must be paid to omitted variable bias, potential confounders, or mediating effects. These concepts will be covered in Elements of Biostatistics.

1.5 Confirmation (Ch. 7 & 8)

Confirmatory statistics attempt to prove a hypothesis through a test. Since biological evidence generally comes from observations rather than deductive proof, we have to agree on how much evidence is enough: *When do we stop believing in coincidence?*

For example, throwing two sixes in a board game is nothing noteworthy, but throwing two sixes for 10 turns in a row suggests the dice are loaded. To express the chance of observing certain results, some basic knowledge of probability theory is required.

Examples of questions primarily concerned with confirmation are:

- Is there an increased risk of cancer from smoking?
- On which diets do guinea pig teeth differ in length?
- Which segments of *Iris* flowers differ between species?

Since empirical evidence cannot prove anything with certainty, confirmational statistics is about controlling the chance of a false positive or false negative. If these are below some agreed upon level, the result can be considered statistically significant.

False Positive

It's a bird!

False Negative

It's not a bird.

1.6 Prediction (Advanced Statistics)

The goal of prediction is to predict the outcome of new observations. The data are usually split into a train and test set. Statistical models are fitted on the train set, and the outcome of the test set is predicted. These prediction are then compared to the actual outcome of the test set, and model with the lowest prediction error is chosen.

Examples of questions primarily concerned with prediction are:

- Which patients should be regularly screened for cancer?
- How long will guinea pig teeth be, given dietary vitamin C levels?
- Can we classify new *Iris* flowers based on sepal/petal dimensions?

More general predictive problems include:

- Forecasting the weather;
- Anticipating the stock market;
- Predicting which patients will develop a disorder.

Predictive modelling generally involves the same models used for parameter estimation, although the best model is determined by prediction accuracy, rather than quality of the parameter estimates. These need not coincide, as explained in the following paragraph.

1.7 Summary

The best statistical approach depends on the research question. Two scientists may use different methods on the same data, but neither is necessarily better unless their goals are the same.

Predictive models are not fundamentally different from models used solely for inference. However, if the objective is different, the best model may also be different.

As an example, neural networks are becoming increasingly popular. They are frequently used in facial recognition, because they excel at learning complex relationships between input variables, classifying faces with very high accuracy. However, their interpretability is very low. Questions like: *Which features of a face are recognized?* or *What are the defining differences between human faces?* cannot be answered easily with a neural network, even though they may vastly outperform a simpler model in terms of prediction accuracy. Hence, the objective—and not the type of data—determines the best statistical approach.

Overview of Statistical Objectives

This book series distinguishes between **exploration**, **estimation**, **confirmation** and **prediction**. While a typical analysis draws on several of these, the best statistical method depends on which of these research goals is of primary interest. Below is a summary highlighting the differences. In computer science, the listed aliases are more common.

Objective	Description	Primary Outcome	Examples	Alias
Exploration	Gain insight in data	Summaries, figures	Plots, tables, PCA	Unsupervised Learning
Estimation	Estimate parameters	Estimates, confidence intervals	Regression analyis	Supervised Learning
Confirmation	Test hypothesis	Confidence intervals, p-values	T-test, F-test, χ^2-test	Hypothesis testing
Prediction	Predict new outcome	Predictions, classifications	Regression analyis	Supervised Learning

Table 1.1: Summary of research objectives.

2 Probability Theory

Emprical sciences like biology rely on evidence rather than proof. In order to express the strength of evidence, some understanding of probability is required. Try to read through this chapter quickly and focus on the exercises as at the end of each paragraph.

2.1 Sample Spaces

In statistics, *space* refers to the values that something can take on. The possible outcomes of an experiment are its *sample space*. For example, the sample space (Ω) of a single roll of a 6-sided die is:

$$\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$$

This sample space Ω is a *set*, and has several proporties:

- The sum of all probabilities is always 1;
- The size (or cardinality) is $|\Omega| = 6$;
- If the die is fair, each outcome has the same probability: $\frac{1}{|\Omega|} = \frac{1}{6}$.

From here, some basic probability can already be calculated:

1. What is the chance of rolling $\boxdot$?
2. What is the chance of *not* rolling $\boxdot$?
3. What is the chance of rolling either $\boxdot$ or $\boxdot$?

Of course you know the answer to these questions, but let's go over their mathematical notations:

1. $P(\boxdot) = \frac{1}{6}$;
2. $P(\boxdot^c) = 1 - P(\boxdot) = \frac{5}{6}$;
3. $P(\boxdot) + P(\boxdot) = \frac{1}{3}$.

If an event is called A, then everything in the sample space that is not A is called A^c: The *complement* of A (fig. 2.1).

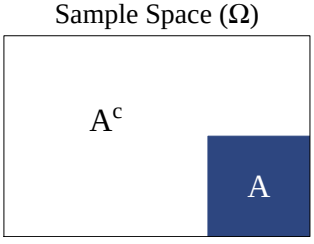


Figure 2.1: A and its complement A^c.

Now consider **two consecutive die rolls**: If the first and second roll each have 6 possibilities, the sample space comprises all 6×6 possible outcomes:

$$\Omega = \left\{ \begin{matrix}\boxdot\\\boxdot\end{matrix}, \begin{matrix}\boxdot\\\boxdot\end{matrix}, \begin{matrix}\boxdot\\\boxdot\end{matrix}, \begin{matrix}\boxdot\\\boxdot\end{matrix}, \begin{matrix}\boxdot\\\boxdot\end{matrix}, \begin{matrix}\boxdot\\\boxdot\end{matrix}, \begin{matrix}\boxdot\\\boxdot\end{matrix}, \begin{matrix}\boxdot\\\boxdot\end{matrix}, \begin{matrix}\boxdot\\\boxdot\end{matrix}, \dots, \begin{matrix}\boxdot\\\boxdot\end{matrix}, \begin{matrix}\boxdot\\\boxdot\end{matrix}, \begin{matrix}\boxdot\\\boxdot\end{matrix} \right\}$$

Every sequence of rolls has the same chance of $\frac{1}{|\Omega|} = \frac{1}{36}$. However, if the order of the rolls does not matter, then the number of combinations is a lot smaller, since then: $\begin{matrix}\boxdot\\\boxdot\end{matrix} = \begin{matrix}\boxdot\\\boxdot\end{matrix}$, and $\begin{matrix}\boxdot\\\boxdot\end{matrix} = \begin{matrix}\boxdot\\\boxdot\end{matrix}$, and so on. The sample space of **the sum of two consecutive rolls** is even smaller:

$$\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

Certain outcomes occur in more different ways than others (table 2.1). Hence, not every outcome is as likely as the rest. **The chance that the sum of two rolls** (X) **results in a particular outcome** (x) is:

$$P(X = x) = \frac{\text{number of ways to get } x}{\text{number of ways to roll the dice}} \tag{2.1}$$

	⚀	⚁	⚂	⚃	⚄	⚅
⚀	2	3	4	5	6	7
⚁	3	4	5	6	7	8
⚂	4	5	6	7	8	9
⚃	5	6	7	8	9	10
⚄	6	7	8	9	10	11
⚅	7	8	9	10	11	12

Table 2.1: The sum of each combination.

Some new questions can now be answered, such as:

1. What is chance of rolling a total of 10?
2. What is the chance of rolling more than 10?
3. What is the chance of rolling an even total?

Using equation 2.1, we obtain:

1. $P(X = 10) = \frac{3}{36} = \frac{1}{12}$;
2. $P(X > 10) = P(X = 11) + P(X = 12) = \frac{2}{36} + \frac{1}{36} = \frac{3}{36} = \frac{1}{12}$;
3. $P(\text{even}) = \frac{1}{36} + \frac{3}{36} + \frac{5}{36} + \frac{7}{36} + \frac{5}{36} + \frac{3}{36} + \frac{1}{36} = \frac{18}{36} = \frac{1}{2}$.

When considering multiple events, it is sometimes necessary to use the complement of an event (fig. 2.1). For example, **the chance of rolling ⚅ three times in a row** is:

$$P(⚅) \cdot P(⚅) \cdot P(⚅) = P(⚅)^3 = \left(\frac{1}{6}\right)^3 = \frac{1}{216}$$

But what is **the chance of rolling at least one ⚅**? The answer is clearly not $\frac{1}{6} + \frac{1}{6} + \frac{1}{6}$, since by that logic, 6 rolls would yield 100%. Instead, using the complement (the chance of *not* rolling ⚅):

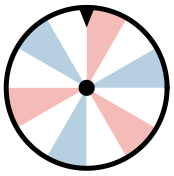
$$P(\text{at least one } ⚅) = 1 - P(⚅^c)^3 = 1 - \left(1 - \frac{1}{6}\right)^3 = \frac{91}{216}$$

If this equation is not immediately clear, read it as follows: When rolling three times, the chance of at least one ⚅ is 1 minus the chance of rolling something else three times in a row.[1]

[1] This is also how the chance of a false positive among multiple independent tests is calculated (chapter ***X***).

Questions

Below are some example questions and answers. Try to solve them yourself, then check your answers using the solutions below. You can use R to perform calculations if you want.



1) If you spin the wheel on the right and all outcomes are equally likely...
   (a) ...what is the chance of landing on white?
   (b) ...what is the chance of not landing on red?
2) Suppose there are 8 specific mutations required for developing a certain type of cancer, and the chance of acquiring any one of them in a lifetime is 70%...
   (a) ...what is the chance of developing this type of cancer?
   (b) ...what is the chance of not developing this type of cancer?
3) If there are three significant results reported in a scientific paper, each with a 5% chance of being a false positive...
   (a) ...what is the chance all 3 are false positives?
   (b) ...what is the chance of at least one false positive?

!!! ADD: Memoryless property

	Analytical Solution	**Solution in R**
1 a	$P(\text{white}) = \frac{\text{number of ways to get white}}{\text{number of ways to spin the wheel}} = \frac{6}{12} = \frac{1}{2}$	`6 / 12`
1 b	$P(\text{red})^c = 1 - P(\text{red}) = 1 - \frac{3}{12} = \frac{3}{4}$	`1 - 3 / 12`
2 a	$0.70^8 \approx 5.8\%$	`0.70^8`
2 b	$1 - 0.70^8 \approx 94.2\%$	`1 - 0.70^8`
3 a	$P(\text{false positive})^3 = 0.05^3 = 0.0125\%$	`0.05^3`
3 b	$P(\text{at least one false positive}) = 1 - (1 - 0.05)^3 \approx 14.3\%$	`1 - (1 - 0.05)^3`

(To understand the solution to 3 b*, consider that "at least once out of three" is the complement of "three times not".)*

2.2 Combinations & Permutations

If a restaurant serves 7 main courses and 5 deserts, then there are $7 \times 5 = 35$ combinations to choose from. This is the **multiplication principle**. In general, the number of combinations between p variables, each with $k_1, k_2, \ldots, k_p$ categories, is equal to the product of the number of categories:

$$\prod_{i=1}^{p} k_i = k_1 \times k_2 \times \cdots \times k_p \tag{2.2}$$

But what if for some reason you wanted more than one desert, how many combinations of different deserts could you choose? This can be calculated using the **binomial coefficient**:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{2.3}$$

In R, you can evaluate the binomial coefficient using the function `choose()`. For example: `choose(5, 2)` returns `10`. If you want to return the combinations, try `combn(5, 2)` instead.

Where n is the number of categories to choose from and k the number of categories in each combination. For example, if you were to have 2 different deserts out of 5 possible choices, the number of possible combinations would be:

$$\binom{5}{2} = \frac{5!}{2! \times (5-2)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 3 \times 2 \times 1} = 10$$

Combinations matter to experimental design and sample size. If there are fewer experimental units than combinations of factors, then those factors cannot all be compared.

Permutations are the possible rearrangements of a sequence (fig. 2.2). How many ways a sequence can be rearranged depends on whether *replacement* is allowed: Using a category more than once. The number of permutations **with replacement** can be calculated as:

$$n^t = n \times n \times \cdots \times n \tag{2.4}$$

Where n is the number of categories and t the length of the sequence. Permutations **without replacement** can be calculated as:

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 1 \tag{2.5}$$

In the life sciences, permutations are particularly relevant in the study of DNA, RNA and proteins, where one might consider the number of ways their constituent parts can be arranged.

Permutations

with replacement			without
1, 1, 1	2, 1, 1	3, 1, 1	1, 2, 3
1, 1, 2	2, 1, 2	3, 1, 2	1, 3, 2
1, 1, 3	2, 1, 3	3, 1, 3	2, 1, 3
1, 2, 1	2, 2, 1	3, 2, 1	2, 3, 1
1, 2, 2	2, 2, 2	3, 2, 2	3, 1, 2
1, 2, 3	2, 2, 3	3, 2, 3	3, 2, 1
1, 3, 1	2, 3, 1	3, 3, 1	
1, 3, 2	2, 3, 2	3, 3, 2	
1, 3, 3	2, 3, 3	3, 3, 3	

Figure 2.2: All permutations of $\{1, 2, 3\}$ with and without reusing the numbers.

Questions

Below are some example questions about combinations and permutations in biology. Try and solve these using equations 2.2 - 2.5.

HINT: In R, multiply using `a * b`, calculate the binomial coefficient with `choose(n, k)`, and calculate the number of permutations using `n^t` (with replacement) or `factorial(n)` (without replacement).

1) Common complications of obesity are type 2 diabetes and high blood pressure. Suppose there are 3 available kinds of medicine for diabetes and 4 for high blood pressure. How many combinations are possible?

2) Some antibiotics display synergy: Administering two simultaneously increases the effectivenes beyond the sum of their individual effectiveness. If there are 10 available kinds of antibiotics:

   (a) How many combinations of 2 antibiotics are possible?

   (b) How many combinations of 3 antibiotics are possible?

   (c) How many combinations of 2 or more antibiotics are possible?

3) Codons are sequences of 3 nucleotides and there are 4 kinds of nucleotides (A, C, G, T). How many different codons exist?

4) Many codons are redundant and code for the same amino acid. If there are 20 kinds of amino acids:

   (a) How many different amino acid chains of length 5 exist?

   (b) How many different amino acid chains of length <5 exist?

	Analytical Solution	**Solution in R**
1	$k_1 \cdot k_2 = 3 \cdot 4 = 12$	`3 * 4`
2 a	$\binom{n}{k} = \binom{10}{2} = \frac{10!}{2!(10-2)!} = \frac{10\times9\times8\times7\times\ldots\times1}{2\times1\times8\times7\times\ldots\times1} = \frac{10\times9}{2} = 45$	`choose(10, 2)`
2 b	$\binom{n}{k} = \binom{10}{3} = \frac{10!}{3!(10-3)!} = \frac{10\times9\times8\times7\times6\times\ldots\times1}{3\times2\times1\times7\times6\times\ldots\times1} = \frac{10\times9\times8}{3\times2} = 120$	`choose(10, 3)`
2 c	$\binom{10}{2} + \binom{10}{3} + \binom{10}{4} + \binom{10}{5} + \binom{10}{6} + \binom{10}{7} + \binom{10}{8} + \binom{10}{9} + \binom{10}{10} = 1013$	`sum(choose(10, 2:10))`
3	$n^t = 4^3 = 64$	`4^3`
4 a	$n^t = 20^5 = 3.2 \cdot 10^6$	`20^5`
4 b	$20^4 + 20^3 + 20^2 + 20^1 = 168420$	`sum(20^(4:1))`

2.3 Conditional Probability

Conditional probability is the chance of an event A, given the outcome of another event B, denoted as: $P(A|B)$.

If A and B are not independent, then $P(A|B) \neq P(A)$, as shown in figure 2.3: While the chance of any random individual getting struck by lightning on a given day is extremely low: $P(A) = \frac{1}{7{,}000{,}000}$, the chance of getting struck by lightning *given* that you are currently outside in a thunder storm is considerably higher.

A quintessential use of conditional probability in statistics is the *likelihood*, which is the probability of the observed sample given a particular model. One way to find the best fitting model for a sample is by maximizing the likelihood (to be included).

Figure 2.3: Comic about conditional probability. `https://xkcd.com/795/`

2.3.1 A Worked Out Example

Consider the possible outcomes of a test for viral infection (table 2.4). Patients are either infected or not and the virus is either detected or not. This means that there is still a chance of infection, even if the test is negative, and vice versa.

	Infected	**Not infected**
Virus detected	true positive	**false positive**
No virus detected	**false negative**	true negative

Table 2.4: Summary of false positives and negatives, which are also referred to as type I and II errors, respectively.

Of course, a perfect test only has true positives and true negatives. However, it is easy to imagine the potential for a **false negative**:

- There may be viral particles, but none bind to the testing agent;
- If the test binds antibodies, some patients might not produce any;
- The virus may not have proliferated yet shortly after exposure.

Every test, no matter how good, can result in false negatives. But how could a virus be detected if there is none? Below are some of the many reasons a **false positive** could occur:

- The testing agent could accidentally bind something else;
- Antibodies could be detected after the virus is already cleared;
- Antibodies (but not the virus) could be transferred by breastmilk.

What does this mean for a patient going to the doctor? Consider the example shown in table 2.5. This test detects most viral infections, with a sensitivity of $\frac{144}{150} = 96\%$. It also classifies most uninfected patients correctly, with a specificity of $\frac{4559}{4850} = 94\%$.

	Infected	**Not infected**	*Total*
Virus detected	144	291	435
No virus detected	6	4559	4565
Total	150	4850	5000

Table 2.5: Toy example of viral test.

Then **what is the chance of infection given a positive test result?** To answer this question, look at the row for positive test results and observe that $\frac{144}{435} \approx 33\%$ is actually infected. In other words, even if a patient tests positive, there is still about a $\frac{2}{3}$ chance that it is merely a false positive. How could that be with such high sensitivity and specificity? The answer is simply: Because the infection is rare, occurring only in $\frac{150}{5000} = 3\%$ of those tested for it.

Medical tests are but one example of conditional probability. In real life, the marginal distribution of events is rarely interesting. If you were to ask a doctor the chance of having any particular disease or condition, the doctor would first *condition* this probability on your age, gender, familial history, etc. In addition, the overall prevalence of a disease (and thus the marginal probability) is meaningless to a patient, who is likely already at the doctor presenting with symptoms.

Statistical models generally aim to estimate some outcome *given* one or multiple explanatory variables (chapter **9**). One model in particular estimates conditional *probability* directly: Logistic regression, also known as a binomial GLM. This is a special kind of linear regression model introduced in *Elements of Biostatistics*.

2.3.2 Bayes' Theorem

Unlike the example from table 2.5, there is usually no way to know for certain which patients are false positives and which are actually infected. However, if certain other information is known from historical data, the chance of infection given a positive result can still be obtained through Bayes theorem:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{2.6}$$

Suppose 3% of the population has the infection and the 96% of infections are detected. Out of 5000 patients 435 test positive:

1. $P(B|A)$ is the chance of detection given infection (sensitivity);
2. $P(A)$ is the chance of infection (prevalence);
3. $P(B)$ is the chance of detection which is $\frac{435}{5000} = 8.7\%$ in the sample.

Using equation 2.6, the answer is $\frac{0.96 \cdot 0.03}{0.087} \approx 33\%$, which is the same as the answer obtained previously, using table 2.5.

Questions

1) A common fallacy is to assume that $P(A|B) \approx P(B|A)$. While it may coincidentally hold in some cases, it is generally false. Using table 2.6 estimate the chance that...

   (a) ...someone has cancer, given that they are a smoker.

   (b) ...someone is a smoker, given that they have cancer.

	Smoking	**Non smoking**	*Total*
Cancer	122	450	572
No cancer	78	525	603
Total	200	975	1175

Table 2.6: Data from a case-control study of esophageal cancer in Ille-et-Vilaine, France.

2) Estimate the following conditional probabilities from table 2.4:

   (a) $P(\text{infected}|\text{not detected})$

   (b) $P(\text{not infected}|\text{detected})$

   (c) Which errors do (a) and (b) represent?

3) Consider a test for viral infection to screen potential donors:

   (a) Which error weighs more heavily for blood donors and why?

   (b) Kidney transplants are considerably harder to find, how does this change the severity of the errors?

	Analytical Solution	**Solution in R**
1 a	$P(\text{cancer}|\text{smoking}) = \frac{122}{200} = 61\%$	`122 / 200`
1 b	$P(\text{smoking}|\text{cancer}) = \frac{122}{572} \approx 21.3\%$	`122 / 572`
2 a	$P(\text{infected}|\text{not detected}) = \frac{6}{4565} \approx 0.13\%$	`6 / 4565`
2 b	$P(\text{not infected}|\text{detected}) = \frac{291}{435} \approx 66.9\%$	`291 / 435`
2 c	(a) false negative (b) false positive	
3 a	Due to the risk of infecting the recipient, a false negative is more dangerous when screening potential donors.	
3 b	Since donors are scarce, a false positives now has a more detrimental effect, wasting a healthy donor. *(Of course, a false negative is still as dangerous as before.)*	

2.4 *Central Limit Theorem*

(. . .) to be included.

3 Study Design

Before any experiments take place, the first statistical considerations should already be made: Think of sample size, treatment assignment and deciding on which variables to are of interest.

Study design is crucial to the simplicity and power of analysis. To aid in proper study design, this chapter introduces the concepts of uncertainty, representative sampling and treatment assignment.

3.1 Sources of Uncertainty

It is almost always impractical, unaffordable or even impossible to measure an entire population. Instead, a random subset of the population, called a *sample*, is used to draw conclusions. This results in uncertainty, comprised of three major parts: **measurement error**, **variance** and **bias**. How many samples amount to sufficient evidence largely depends on the amount of uncertainty.

When thinking of uncertainty, always try to imagine a sample of multiple observations. Variance and bias describe the behavior of a group, not of a single observation.

3.1.1 Measurement Error

Mislabeling, rounding errors, limited precision of tools (e.g. a pipet) and other small errors during data collection accumulate. By working very carefully and using very precise tools, the measurement error can be kept minimal, but it is never zero.

Examples of sources of measurement error include:

- A sore erroneously identified as oesophageal cancer;
- Minor differences in the exact supplied dose of a drug;
- Rounding errors from measuring physical dimensions with a ruler.

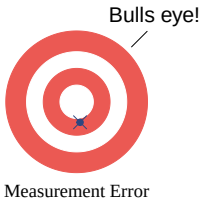


Figure 3.1: Mislabeling is one example of measurement error.

To understand the second example, bear in mind that quantities cannot be exactly determined: A 200 μg tablet might as well be 199 μg or 201 μg. Hence, any effect of dosage will introduce *some* error.

3.1.2 Variance

Variance is a measure of the natural variability, or random nature of a process. It is the extent to which things are unpredictable. We can predict the weather tomorrow, but the exact temperature is always subject to small unpredictable effects. The sum of all small things that cannot be accounted for accumulate into the variance.

Colloquially, variance is often used to simply mean dispersion. In statistics, the word variance is used specifically as the squared differences from the mean (paragraph **5.2**).

Examples of sources of variance include:

- Randomly occurring mutations required for cancer development;
- Differences in tooth length due to genetic variability;
- Differences in sepal length due to small differences in soil.

Other examples of effects that cannot be determined but *can* to some small extent influence biological processes include: background radiation, failure of genetic repair mechanisms and even the original position and momentum of elementary particles.

In practice, variance is influenced by *much* larger effects than states of elementary particles. It is impossible to exactly determine every contributing effect using only observations. At best, statistics can extract the deterministic part of a process to a satisfactory extent. The larger the variance, the more observations are needed to do so.

Sample variance is not only affected by population variance, but also measurement error, sampling error and bias. When you draw a sample, it is impossible to tell what part of the spread in the data is due to which of these components of uncertainty. The sample variance is therefore generally larger than the population variance.

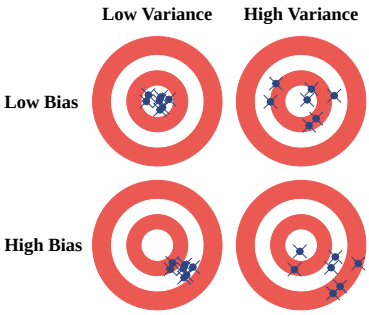


Figure 3.2: Illustration of variance and bias in a sample. The center of the circles represent the population average.

3.1.3 Bias

Whereas variance is a *random* deviation from the expectation, bias is a systematic deviation (fig. 3.2). High bias usually occurs due to poor sampling schemes or poor treatment assignment and can be largely avoided by randomization (paragraphs **3.4** & **3.5**). However, even random sampling and treatment assignment cannot guarantee unbiasedness: Almost all samples differ slightly from their population.

Examples of sources of high bias include:

- Collecting different groups at different locations;
- Assigning consistently older subjects to one supplement;
- Using multiple rulers (differing ever so slightly in length).

3.2 Observational or Interventional

Studies where any intentional change is brought about are interventional. The intervention could be any kind of treatment, such as:

- A stressor, like bacteria grown with suboptimal temperature, pH, or osmotic pressure;
- A treatment for a disease or condition, including medicine, surgery, dietary changes, or a change in physical activity;
- Intentionally inflicted conditions, such as injecting cancer cells into laboratory mice.

Studies without intervention, are observational. Examples include:

- A comparison of the amount of antibiotic production of two kinds of bacteria, which thrive under different conditions;
- A comparison of long-term survival between patients which have undergone surgery and those who have not, using historical data;
- A comparison of the diet of mice who develop cancer within 2 years and mice who live cancer-free for longer.

Only interventional studies can demonstrate causal relationships, and only if proper controls are included in the study. To see why, consider what might cause differences in the previous examples:

- Efficiency of antibiotic production *could* be caused by the conditions under which the bacteria thrive, or by differences in their production mechanisms, differences in their growth curve, combinations of the former, or a myriad of other reasons;
- Long-term survival *could* be affected by the surgery, or other differences between patients, combinations of the two, or even more convoluting: Patients which did not undergo surgery may be underrepresented in the data set;
- Diet *may* have caused/contributed to cancer development, or developing cancer may have changed the dietary habits of the mice, or other completely unrelated factors may have a dominating effect on cancer development.

Whether *you* are directly involved in the data collection process or not, always carefully consider whether the study is observational or interventional. Only the latter can demonstrate causality. Additional illustrations of why observational evidence cannot be used to support causal claims are included in paragraph **5.5**.

3.3 Representative Sampling

A sample that is representative of its population has minimal bias. Although in most cases, this simply implies *random* sampling, sometimes the random sampling should be stratified according to some other variable. For example, if gender has an influence on the effectiveness of treatments, then each treatment should ideally have an approximately equal ratio of males and females assigned to it. See paragraph **3.7** for an example of how to do this in R.

3.3.1 Drawing a Sample: The Right Way

Whether a sample is representative of its population cannot be assessed directly—other than comparing it to the entire population. However, a *randomly* drawn sample greatly reduces the *chance* of high bias. The example in figure 3.3 has low bias, because:

1. The ratio of males and females reflects that of the population;
2. The variance in height roughly reflects that of the population;
3. The sample was drawn randomly and not hand-picked, avoiding bias by other, unknown variables.

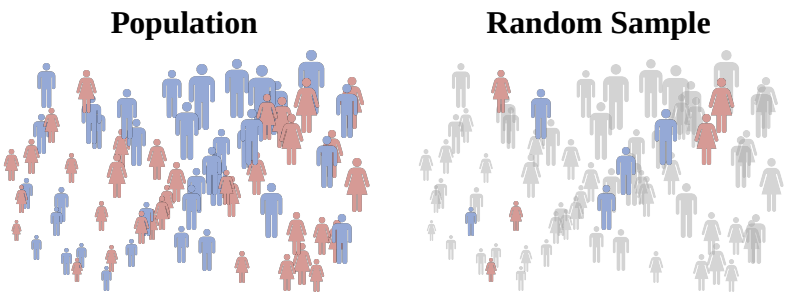


Figure 3.3: A population of males and females with varying heights (left) and a random sample (right).

3.3.2 Drawing a Sample: The Wrong Way

Non-random sampling is almost always guarenteed to yield a biased sample. Two examples are illustrated in figure 3.4.

Neither represents the population well: On the left, a sample was purposely drawn with relatively tall women and short men, which is of course fraudulous. On the right, a sample was drawn 'lazily' from the top-right corner. The latter tends to occur when samples are drawn 'closest' to the researcher, such as:

- A survey conducted on family and friends;
- Ecological research in a park nearby;
- Testing a new drug in a single hospital.

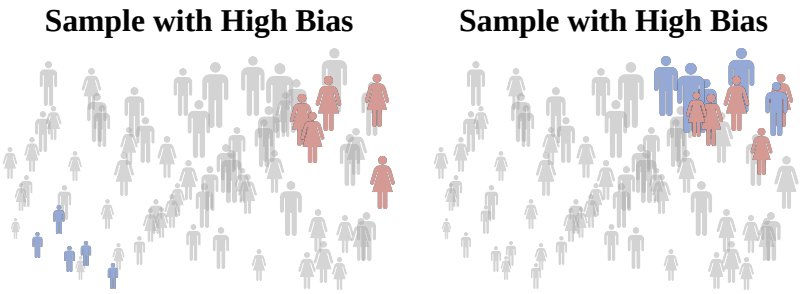


Figure 3.4: Two examples of systematic (i.e. wrong) sampling. On the left, a sample with relatively short men and tall women was drawn. On the right, samples were drawn locally and thus only tall individuals were observed.

The examples given above may not always be avoidable for practical reasons. When this is the case, be extra careful with the conclusions and clearly state the limitations of the study in the discussion.

3·4 *Random Assignment*

Like random sampling, random assignment avoids bias. This can refer to actual treatments to be assigned to samples, but also the order in which experiments or observations take place (fig. 3.5). Reasons for randomizing treatment assignment include:

1. If more than one researcher does the experiments/observations, there might be a difference in their measurement error;
2. If an experiment is repeated multiple times, a learning effect may occur, and later experiments may have lower measurement error;
3. Precision and accuracy of equipment may degrade over time;
4. If there is time in between different observations, anything from short periodic effects (e.g. day and night) to seasonal effects and trends may affect the results.

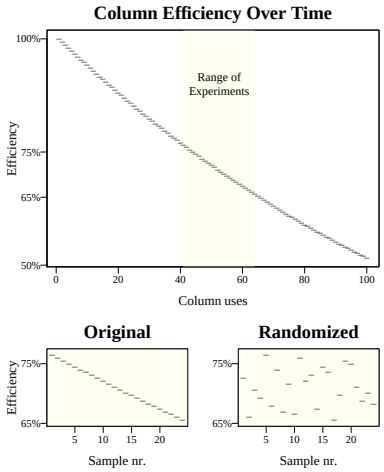


Figure 3.5: Example of the effect of randomization on a spin column that is repeatedly used. If in this case you were to analyze one group first and then the other, the observed difference in groups would be biased by the degrading efficiency of the column.

Random treatment assignment does not remove these effects, but prevents them from having a systematic effect (fig. 3.5). Suppose for example that a clinical trial *first* measures all cases and *then* all controls. This is problematic, since you cannot determine whether observed differences are due to status (case/control), or bias.

Randomization should be performed by a random number generator and not by scientists themselves, because we tend to think certain patterns are 'more random' than others. Every variable or experiment requires a new random assignment. A simple example procedure is included at the end of this chapter.

Of course, a random assignment could follow a pattern by pure *coincidence*. However, the *chance* of a biased design is minimal when using randomization. Collecting more data, or performing a replication study further decreases the chance of biased results.

3.5 Choosing a Sample Size

In terms of strength of evidence, **more observations is always better**. The only reasons to limit sample size are financial, ethical, privacy, or time constraints. Hence, only recommendations for a *minimal* sample size will be made here.

However, a minimal sample size is typically not sufficient to give a satisfactory answer the research question. Try to collect as many samples as possible, while the time and monetary investment remain reasonable with respect to the importance of the conclusion.

That said, the steps for determining a minimal sample size are:

1. Formulate a research question;
2. Determine the nature of this question (chapter ***1***);
3. Choose a suitable test or model;
4. Count the number of parameters in this test or model;
5. Add 1.

Two observations can be made here: (1) To determine a minimal sample size, you must already be familiar with the test or model you intend to use;[1] (2) In general, the sample size should simply be larger than the number of parameters. How many parameters a test or model has, will be explained as tests and models are introduced.

A bare minimum number of observations will usually not yield interesting or reliable conclusions. Many textbooks on regression analysis, recommend at least 3, or even 5 observations per parameter. If you cannot collect enough samples, you should reconsider the practicality of the research question and perhaps simplify the problem first.

It is technically possible to use *regularization* to allow for more parameters than observations. However, this requires a reasonable understanding of statistical concepts and will not be introduced until the third book in this series: Advanced Biostatistics.

[1] This is particularly hard, since you cannot be reasonably expected to know many models as a beginner in statistics. However, it is increasingly easier to find expert help on websites like CrossValidated. Hence, focus on learning the concepts, such that you can get better at asking questions on how to model the research question you have. As we progress through the books, and as you progress through your education, the range of models you are familiar with will expand naturally.

3.6 Examples in R

If you are not sure how to use R, read *Statistics in R and RStudio* first. Try adjusting these examples to match your own experiments.

If your design is more complex than that covered here, try asking a question on **CrossValidated** (statistics) or stack**overflow** (programming), or consult a statistician.

3.6.1 Random Sampling

```
patients <- 1:1000  # Suppose there are 1000 patients to choose from
n        <- 30      # Suppose we want to draw 30 random patients
sample(patients, n) # A random selection of size n of those patients
```

```
 [1] 375 565 595  11 622  71  13 205 480 384 749 276 599 717  60 978 354 345 784
[20] 105 115 628 671 696 965 832 114 999 360 474
```

3.6.2 Stratified Random Sampling

The easiest way to stratify is by sampling separately for each category, according to the ratio with which you want to sample them. If the category is suspected to influence the results of an unrelated comparison, try to draw categories with the same probability they are represented in the population. If on the other hand your goal is to compare these categories, try to draw an equal number of each category.

```
patients <- 1:1000  # Suppose there are 1000 patients to choose from
gender   <- rep(c("male", "female"), times = c(100, 900)) # Suppose 10% is male
n        <- 100     # Suppose we want to draw 100 random patients

# In case of 1:9 sampling
sample(patients[gender == "male"], n * 0.1) # Sample males
```

```
 [1]  5 23 87  6 99 78  7 19 47  3
```

```
sample(patients[gender == "male"], n * 0.9) # Sample females
```

```
 [1]  37  67  72  99  88  78   7  80  57  30  42  38  73  20  97  49  18   8  24
[20]  87   4  13  14  25  11  81  76  68  12  95  92  90  86  98  29  21  47  66
[39]  43  64  35  85  52  65  93  10  69  28  77  50  83   5  82  53  41  19  75
[58]  63  51  26  15  39  58  45  40   9  16  60  74  56   3  44  54  62  34  46
[77]  71 100  91  27  22   6  55  79  94  70  59  61  96  33
```

```
# In case of equal sampling
sample(patients[gender == "male"], n * 0.5) # Sample males
```

```
 [1] 68 81 42 85 73 49 94 20 60 87 59 78 50 74 61 48 31 36 17 88 53 29 98 55 16
[26] 30 71 67 91 39 63 27  8 43 54 93 47 65  3  4 92 97 40 41 75 45  2 34 25 56
```

```
sample(patients[gender == "male"], n * 0.5) # Sample females
```

```
 [1]  96  98  85  61   2  35   3  43  99  77  36  46 100  31  84  16  91  76  11
[20]  68  70  25   6  42  47  59  20  10  66  87  90  51  45  19   9  75  94  49
[39]  54  30  28   8  79  89  93  17  27  74  55  34
```

If you want to stratify for continuous variables, or multiple variables, try the function `stratified` from the R package `splitstackshape`.

3.6.3 *Random Treatment Assignment*

If possible, you should use a balanced design (the same number of observations for each treatment):

```
n <- 3                                      # The sample size per treatment
treatments <- rep(c("A", "B", "C"), each = n) # The treatments to be applied
sample(treatments)                          # A random assignment
```

```
[1] "B" "B" "B" "A" "A" "C" "A" "C" "C"
```

If a balanced design is not possible, because there is a limited sample size among which to distribute treatments, use the following instead:

```
n    <- 9                              # Suppose there are 9 plants available
pH   <- c(7, 9)                        # The treatments (neutral and high pH)
k    <- length(pH)                     # The number of treatments
full <- rep(pH, each = ceiling(n / k)) # Start from a balanced design
sample(full, n)                        # Reduce to a random sample of size n
```

```
[1] 9 7 9 9 9 7 7 7 7
```

3.6.4 *Random Assignment of Multiple Treatments (*)*

Balanced version, 2 explanatory variables:

```
n <- 3                 # Sample size per treatment
temp <- c(37, 48)      # Treatment 1: Temperature
conc <- c(10, 20, 100) # Treatment 2: Concentration of something
combinations <- expand.grid(temp, conc) # Combinations of treatments
treatments   <- do.call("rbind", rep(list(combinations), n)) # Magic
treatments[sample(1:nrow(treatments)), ] # A random assignment
```

```
   Var1 Var2
10   48   20
5    37  100
11   37  100
13   37   10
8    48   10
3    37   20
9    37   20
7    37   10
14   48   10
4    48   20
17   37  100
18   48  100
2    48   10
12   48  100
1    37   10
15   37   20
16   48   20
6    48  100
```

Unbalanced version:

```
# Start from the balanced design shown above
max_n <- 15                                    # Specify available samples
treatments[sample(1:nrow(treatments), max_n), ] # Sample 15 random combinations
```

```
   Var1 Var2
4    48   20
1    37   10
7    37   10
17   37  100
10   48   20
11   37  100
18   48  100
12   48  100
15   37   20
14   48   10
9    37   20
8    48   10
16   48   20
6    48  100
3    37   20
```

3.6.5 *Random Spatial Assignment (*)*

This is useful when measurements are performed in a physical space, like a plot of land, a 96-well plate, or an arrangement of plants in a laboratory. Measurements that are physically closer to each other can be logically expected to correlate more strongly, so you should randomize which assignment ends up where.

```
k <- 4                                          # The number of treatments
m <- 4                                          # The number of rows
n <- 5                                          # The number of columns
treatments <- rep(LETTERS[1:k], (m * n) / k) # The treatments
M <- matrix(nrow = m, ncol = n)                 # Empty matrix
M[] <- sample(treatments)                       # Fill with random treatments
print(M)                                        # Random assignment
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,] "C"  "B"  "B"  "C"  "D"
[2,] "A"  "A"  "C"  "B"  "D"
[3,] "D"  "D"  "C"  "B"  "C"
[4,] "A"  "B"  "A"  "D"  "A"
```

The example below is small, for readability. For example, to adjust this for a 96-well plate, set $m = 8$ and $k = 12$.

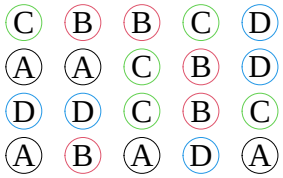


Figure 3.6: The random spatial assignment in this example, with colours to indicate the spread of treatments over the area.

3.6.6 *Latin Square Design (*)*

A special case of the former, a latin square design is the best choice when there are k treatments to divide over a $k \times k$ square.

(. . .) to be included.

4 Exploring Data

Prior to any type of analysis, it is wise to inspect data numerically and visually. Exploration can not only condense large datasets to figures and tables that are easy to digest, but can also reveal potential relationships or outliers.

While the data in table 4.1 consists of only 30 observations, it is still hard to tell which group has higher weight, or whether any observations are outliers. Even such a small dataset is impractical to inspect as a whole. Using summary statistics and visualization, you can easily gain insight in data, regardless of the number of observations.

	Weight (g)	Group
1	4.17	control
2	5.58	control
3	5.18	control
4	6.11	control
5	4.50	control
6	4.61	control
7	5.17	control
8	4.53	control
9	5.33	control
10	5.14	control
11	4.81	treat. 1
12	4.17	treat. 1
13	4.41	treat. 1
14	3.59	treat. 1
15	5.87	treat. 1
16	3.83	treat. 1
17	6.03	treat. 1
18	4.89	treat. 1
19	4.32	treat. 1
20	4.69	treat. 1
21	6.31	treat. 2
22	5.12	treat. 2
23	5.54	treat. 2
24	5.50	treat. 2
25	5.37	treat. 2
26	5.29	treat. 2
27	4.92	treat. 2
28	6.15	treat. 2
29	5.80	treat. 2
30	5.26	treat. 2

Table 4.1: Plant growth data.

4.1 Summary Statistics

An example of a **summary table** produced by R is shown in table 4.2, using the data from table 4.1. Since this is a standard data set in R, you can reproduce this table by running: `summary(PlantGrowth)`.

Displayed here are the:

- Minimum (the smallest observation);
- First quartile (the first 25% of observations after sorting);
- Median (the midpoint, or first 50% of observations after sorting);
- Mean (the average);
- Third quartile (the first 75% of observations after sorting);
- Maximum (the largest observation).

Weight (g)	Group
Min. : 3.590	control : 10
1st Qu.: 4.550	treatment 1: 10
Median : 5.155	treatment 2: 10
Mean : 5.073	
3rd Qu.: 5.530	
Max. : 6.310	

Table 4.2: A summary of table 4.1.

`Group` is a categorical variable, so the number of observations per category are displayed instead. When summarizing categorical variables, values like the mean, median or quartiles are meaningless. For example, there is no 'average' group.

Instead, the frequencies of (combinations of) categories are often reported in a **contingency table** (table 4.3). This can be useful to assess whether any groups are under- or overrepresented.

	Sex	
Supplementation	male	female
Orange juice	15	15
Vitamin C	15	15

Table 4.3: Contingency table of a toy data set. Frequencies of combinations of categorical variables are shown.

4.2 Visualisation

Figures are arguably the best way to present your scientific findings. The best way to visualize data depends on what you want to show the audience. This section covers the basic methods.

4.2.1 Boxplot

Boxplots show quartiles[1] of the data and are useful to compare groups visually. A boxplot displays the following:

- The **location** of the median (fig. 4.1), useful in judging the magnitude of a difference between groups (fig. 4.3);
- The **dispersion** of the data (width of the box/tails in fig. 4.1), which gives an indication of the spread of the observations;
- The **shape**, giving a rough idea of whether the data are distributed symmetrically, like in the normal distribution (chapter ***6***);
- Potential **outliers** (fig. 4.2).

[1] Quartiles are partitions of the data into 0-25%, 25-50%, 50-75%, and 75-100%, after sorting. Each quartile contains $\frac{1}{4}$ of the observations.

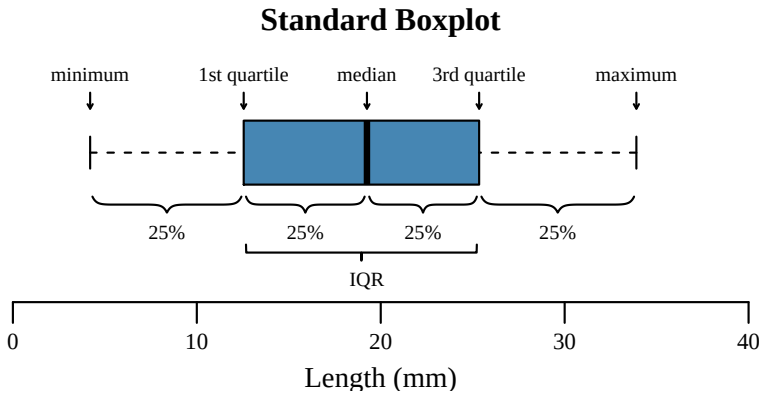


Figure 4.1: A boxplot of tooth length from the `ToothGrowth` dataset. Every quartile contains 25% of the data, so the box contains 50%. The distance between the first and third quartile is called the inter-quartile range (IQR). It gives us an idea of the spread of the data (dispersion) and is used to determine potential outliers (figure 4.2).

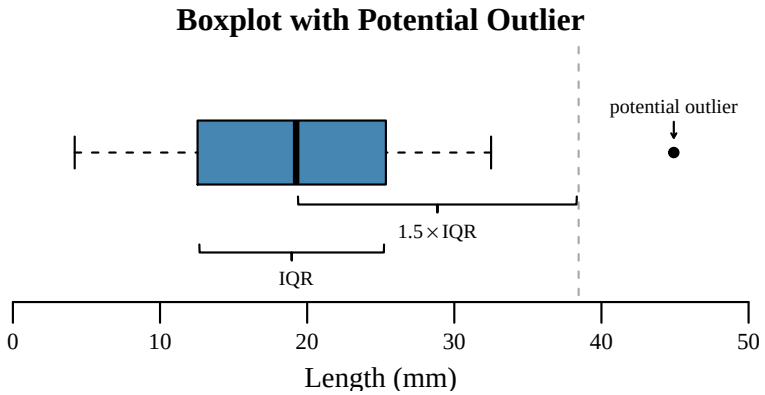


Figure 4.2: A boxplot with a potential outlier. When an observation exceeds 1.5×IQR, it is marked as a potential outlier. This boundary might seem somewhat arbitrary and more well-defined rules for outlyingness will be introduced in later chapters. Nevertheless, boxplots provide a quick way to check the quality of your data.

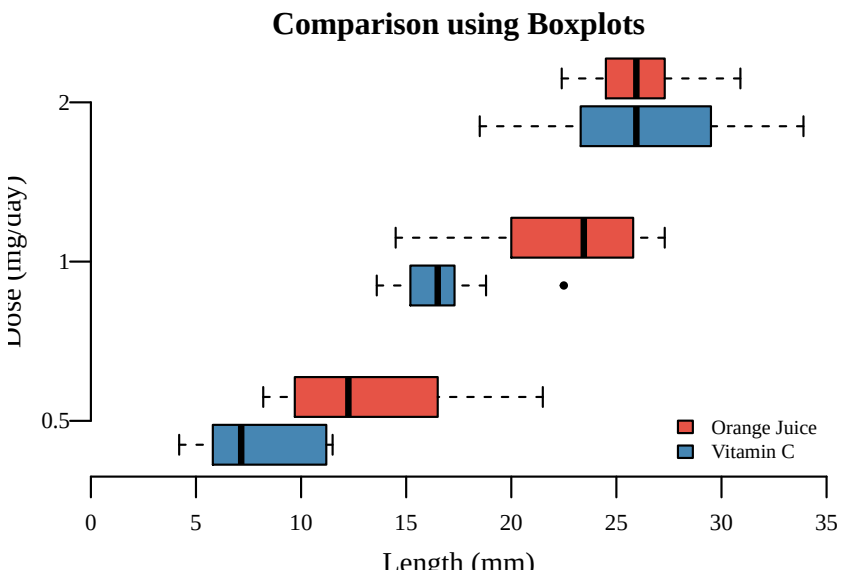


Figure 4.3: A series of boxplots to visualize the effect of vitamin C on tooth growth of guinea pigs. This plot shows several things: (1) There is a clear trend towards longer teeth with greater dose. (2) Orange juice and purified vitamin C appear equally effective at high dose, but at a lower dose, orange juice appears to have a greater effect. (3) There is a possible outlier in the 1 mg vitamin C group. (4) Most groups have approximately symmetric boxplots, so assuming an approximate normal distribution (chapter ***6.2***) is not unreasonable.

4.2.2 *Histogram*

Whereas a boxplot vizualizes quartiles, a histogram plots *frequencies*. Consider the length of guinea pig teeth in the `ToothGrowth` data set in R: A histogram first splits the variable into bins. Let's say we choose bins of 5 mm:

You can view the data set yourself with `ViewToothGrowth`.

- 0-5 mm (1 guinea pig)
- 5-10 mm (11 guinea pigs)
- 10-15 mm (7 guinea pigs)
- 15-20 mm (13 guinea pigs)
- 20-25 mm (12 guinea pigs)
- 25-30 mm (13 guinea pigs)
- 30-35 mm (3 guinea pigs)

A histogram displays the frequencies of each of these bins (fig. 4.4). Unless specified, R will attempt to determine the 'best' number of bins. You can also create a histogram with smaller or larger bins than 5 mm.

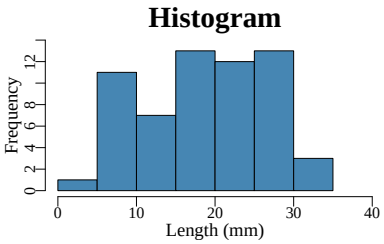


Figure 4.4: A histogram of the length of guinea pig teeth. The 7 bins each have a height corresponding to the number of observations within its range.

Although you can guess the value of the mean or median by looking at a histogram and get a fairly good idea of the dispersion, histograms are mainly useful in determining the **shape** of the data. Assessing the shape of the data allows one to determine what type of probability distribution the data may come from, which is covered in chapter ***6***.

Histograms are especially useful when there is only a limited number of possible values (e.g. the 3 doses of vitamin C), since the frequencies of all values can be visualized in a single figure. Outliers, however, are generally difficult to detect in a histogram.

4.2.3 *Density Plot*

A density plot is a smooth estimate of the probability distribution (chapter ***6***). Think of it as a histogram with infinite bins of infinitesimal width (fig. 4.5). Instead of frequencies, probability density is shown on the y-axis.[2] Density plots are preferred over histograms for continuous variables with a sufficient number of observations.[3]

For example, in figure 4.4, information is essentially discarded in the binning process, where 5 mm was chosen somewhat arbitrarily. Length is a continuous variable, and we have many observations ($n = 60$), so a density plot might be preferable.

If a variable is discrete, or the number of observations limited, a histogram with an appropriate number of bins can reflect the data more accurately.

[2] Probability density is the relative amount of probability in that range. The area under the curve within a certain range is equal to the probability of of values within that range.

[3] A continuous variable can take on any value within a certain range, whereas a discrete variable only takes on certain values (e.g. integers).

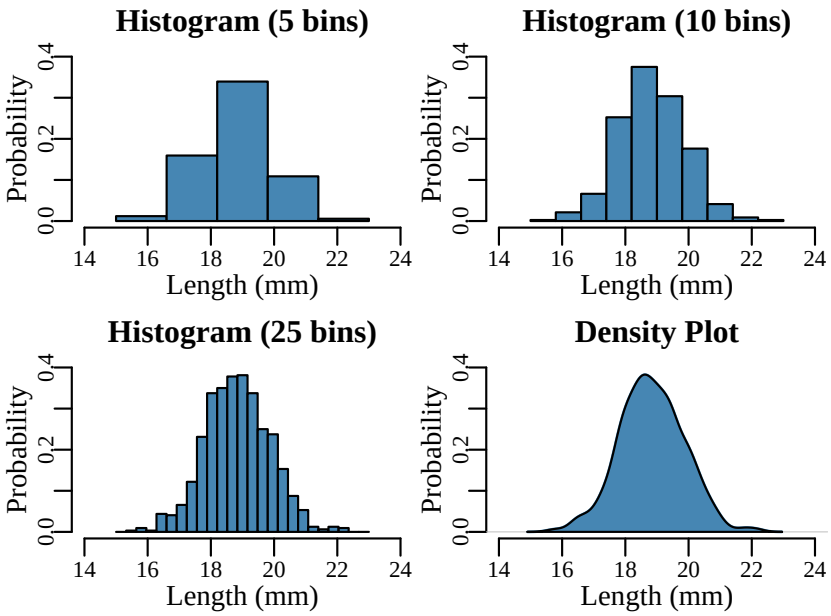


Figure 4.5: As the number of bins increases, a histrogram approaches a density plot. For ease of comparison, these histograms also have probability density on their y-axis.

4.2.4 *Scatter Plot*

A scatter plot shows two different variables on the x and y-axis. It is mainly useful for visualizing potential **relationships** between continuous variables. In some cases, it can also help detecting outliers.

Figure 4.6 shows a scatter plot, using the `iris` dataset: Data on three species of *Iris* flowers' sepal and petal dimensions. A positive, linear relationship can be observed between the length and width. An extreme observation with very small sepal width is also shown in red. However, this observation more or less follows the same trend as the rest of the data, so is unlikely to be an outlier.

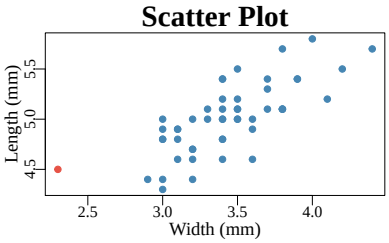


Figure 4.6: A scatter plot of the length and width of *Iris setosa* sepals. The red dot depicts a potential outlier.

4.2.5 Summary of Techniques

Below is a summary of the four plot types discussed so far. It is never wrong to use any of these techniques to gain some quick insight into the data. However, whether any observed anomalies matter depends on what you are plotting, and whether your test or model makes any assumptions about that. Also keep in mind that at small sample sizes, strange shapes need not be indicative of *anything*, other than the small sample size itself.

Technique	Primary use	Outlier detection	Limitations
Boxplot	Visual comparison(s) between groups or categories	Yes	Comparison(s) only meaningful for unimodal distributions.†
Histogram	Visualize frequencies of a discrete variable	No	Binning can obscure interpretation of the data, especially when the bins do not coincide with the original values of a discrete variable.
Density plot	Visualize distribution of a continuous variable	No	Smoothing can obscure interpretation of the data, especially when the number of observations is low.
Scatter plot	Visualize relationship between (continuous) variables	Yes	Cannot visualize relationships between more than 2 variables well.

Table 4.4: An overview of commonly used plotting techniques.
†: Unimodal implies a single peak (e.g. normal distribution).

4.2.6 Questions

(. . .) to be included.

4.3 Examples in R

If a line of code in unclear, try running it from the inside out.[4] Googling "r . . . " will return helpful results from **CrossValidated** (statistics) and stack**overflow** (programming). For example: "r boxplot switch axes", "r how to read data file", "r Error: . . . ", etc.

[4] In other words, first run only what is inside the brackets.

Summary Table

```
data(iris)    # Load the iris dataset
summary(iris) # Produce a summary table
```

Contingency Table

```
data(esoph) # Load the esoph (oesophageal cancer) dataset
str(esoph)  # View the structure: first 3 variables are factors, last 2 numeric
table(esoph[, c(1, 2)]) # Produce a contingency table of the first two factors
table(esoph[, c(1, 3)]) # Produce one for the first and third factor
```

Boxplot

```
data(ToothGrowth)                       # Load the ToothGrowth dataset
boxplot(len, data = ToothGrowth)        # Make a boxplot of the length
boxplot(len ~ dose, data = ToothGrowth) # Boxplot of the length per dose
```

Histogram

```
data(ToothGrowth)                # Load the ToothGrowth dataset
hist(ToothGrowth$len)            # Make a histogram of the length
isOJ <- ToothGrowth$supp == "OJ" # Store which observations are Orange Juice
hist(ToothGrowth$len[isOJ])      # Make a histogram using only those
```

Density Plot

```
data(iris)                                # Load the iris dataset
plot(density(iris$Sepal.Length))          # Make a density plot of sepal length
setosa <- iris$Species == "setosa"        # Store which observations are setosa
plot(density(iris$Sepal.Length[setosa])) # Make a density plot using only those
```

Scatter Plot

```
data(iris) # Load the iris dataset
plot(Sepal.Length ~ Sepal.Width, data = iris) # Scatter plot of sepal length vs width
plot(Sepal.Length ~ Sepal.Width, data = iris, col = Species) # Colour based on species
```

5 Estimation

A representative sample can be used to estimate population parameters. The larger a sample, the closer its estimates will be to the population parameters.

This chapter covers basic parameter estimation. The *parameter* is what that you actually want to know, such as the population mean (μ), the population variance (σ^2), or the true correlation between two variables (ρ). The sample estimates of these parameters have uncertainty, which can be expressed through standard errors (paragraph **5.4**) or confidence intervals (paragraph **5.5**).

Formulas are included to show how the actual estimation works, but most of the time, you will be using statistical software to perform these calculations, examples of which are included in paragraph **5.9**.

5.1 Mean

- The population mean is denoted by μ;
- The sample estimate of the mean is denoted by $\bar{y}$.

The **population mean** is the average value of a variable. It is denoted by μ. Usually, it cannot be calculated, but if somehow all possible observations in a population are available, it is calculated by summing all observations and dividing by the population size (eq. 5.1). Here, y_i means the i^{th} observation and N is the population size.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{5.1}$$

The **sample mean** is an approximation of the population mean. It is denoted by $\bar{y}$ (read: y-bar). It is calculated as the sum of all observations divided by the sample size (eq. 5.2). Here, y_i means the i^{th} observation and n is the sample size.

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{5.2}$$

As with any estimate: As the sample size n approaches the population size N, the estimate will be closer to the true mean. How much the sample mean deviates from the population can only be expressed if μ is somehow known. However, the precision of the estimate can be expressed through a confidence interval (paragraph **5.5**).

5.2 Variance

- The population variance is denoted by σ^2;
- The sample estimate of the variance is denoted by s^2.

Variance is a measure of dispersion: The extent to which observations differ from one another. It is calculated as the average squared differences from the mean (eq. 5.3).

The **population variance** (eq. 5.3) is the sum of squared distances from the population mean (μ) divided by the population size (N). The differences are squared because positive and negative differences from the mean would otherwise average to zero.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2 \tag{5.3}$$

When drawing a sample, its mean *and* variance might differ from that of the population. To take this uncertainty into account, the denominator for the **sample variance** is $n-1$ instead of n. By dividing by a smaller number, the variance is slightly increased (eq. 5.4). Put differently, the data were already used to estimate the sample mean, so 1 degree of freedom is used and $n-1$ remain.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{5.4}$$

If the variance is low, obervations are expected to be close to each other, while if variance is high, observations are further away from each other. What can be considered low or high variance depends on the *scale*: If you were to convert height in meters to centimeters, the variance would increase by 100^2 (because differences are squared).

5.3 Standard Deviation

- The population standard deviation is denoted by σ;
- The sample standard deviation is denoted by s.

The standard deviation is simply the square root of the variance (eq. 5.5 & 5.6 for the population or sample, respectively). It is useful because it has the same unit of measurement as the mean.

$$\sigma = \sqrt{\sigma^2} \tag{5.5}$$

$$s = \sqrt{s^2} \tag{5.6}$$

For example, suppose the average male adult height is 1.80 m and the variance is 0.010. Since the variance takes *squared* differences from the mean, it could be expressed as 0.010 m², but the squared-meter deviation from the height in meters has no meaningful interpretation. Instead, by taking the square root, we have a measure of dispersion that is again expressed in meters ($s = \sqrt{0.010} = 0.10$ m).

The next chapter introduces the *normal distribution*, for which the standard deviation is especially useful. Namely, if adult male height follows a normal distribution, then 95% of adult males have a height between about 2 standard deviations from the mean (1.60 to 2.00 m).

5.4 Standard Error

- The standard error of an estimate is denoted by SE;
- Population values are calculated, not estimated. They have no SE.

The standard error (SE) is a measure of the precision with which a sample estimate approximates its population parameter.The standard error of the mean is calculated as the standard deviation divided by the square root of the sample size (eq. 5.7), such that increasing the sample size decreases the standard error.

$$\text{SE} = \frac{s}{\sqrt{n}} \tag{5.7}$$

In this book, only the standard error of the mean will be covered. However, standard errors can be estimated for most estimates. A general method for standard error estimation is by bootstrapping (*Advanced Biostatistics*).

5.5 Confidence Interval

- The confidence interval of an estimate is denoted CI;
- Population values are calculated, not estimated. They have no CI.

A confidence interval consists of two numbers: a lower bound and an upper bound. Together, they express the precision of an estimate as a *range* (fig. 5.1). If the range is narrow, the estimate is precise and vice versa. The correct interpretation is as follows:

> If samples were drawn repeatedly from this population and x% confidence intervals were constructed each time, then the true value would be contained in x% of these intervals.

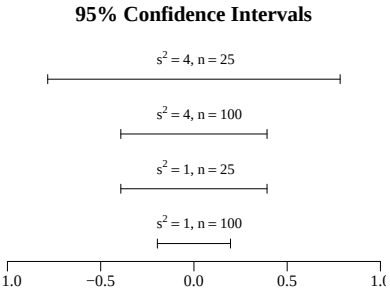


Figure 5.1: Four confidence intervals with the same mean ($\bar{y} = 0$), but different variances and sample sizes.

'Confidence' is a bit misleading, as there is no guarantee that any particular confidence interval contains the true value with x% chance. Despite this, CI are popular because they are easy to calculate, reproduce, and to be used for hypothesis testing (paragraph **7.3**).

It is unfortunately not uncommon to see papers report an estimate ± 1 standard error. However, this has a confidence level of only about 68%. While the calculation for this is beyond the scope of this chapter, you can easily check it in R:

```
k <- 1                                # Nr of standard errors
diff(1 - (1 - pnorm(k * c(-1, 1)))) # Confidence level
```

```
[1] 0.6826895
```

Try and change the `k` variable, to see how many standard errors yield a coverage of approximately 95%. What about 99%?

Most would consider a confidence level of 68% to be too low. To understand why, consider the interpretation of a 68% CI: Almost $\frac{1}{3}$ such intervals do not contain the population value.

More commonly, 95% CIs are reported. The next chapter introduces the normal distribution. This distribution is frequently used in the construction of CIs. For normally distributed variables (such as the sample mean), CIs are calculated as the estimate (e.g. $\bar{y}$) plus-minus approximately 1.96 times the standard error of that estimate (eq. 5.8). The exact calculation for this is beyond the scope of this chapter, but is included in the examples in R at the end.

$$\text{CI}_{95\%}(\bar{y}) \approx \bar{y} \pm 1.96 \cdot \text{SE} \tag{5.8}$$

5.6 Covariance

- The population covariance is denoted by $\text{Cov}[X, Y]$;
- The sample covariance is denoted by $q_{X,Y}$.

Recall that variance describes the extent to which observations from a single random variable differ from one another (paragraph **5.2**). **Co**variance is a property of **two** random variables: It is the extent to which one variable changes when the other does.

If covariance is positive, an increase in one variable corresponds to an increase in the other. In other words, the variables "move in the same direction" together (fig. 5.2). If covariance is negative, an increase in one variable corresponds to a decrease in the other. Finally, a covariance of 0 means the two variables are not linearly related.

Like all population parameters, the covariance is not something that we usually know or can calculate. Due to the uncertainty of a random sample, the *estimated* covariance is rarely exactly 0, even when variables are completely unrelated.

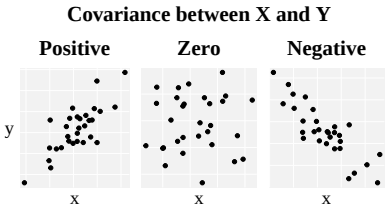


Figure 5.2: Covariance is a measure of linear association.

$$\text{Cov}[X, Y] = \frac{1}{N} \sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y) \tag{5.9}$$

$$q_{X,Y} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{5.10}$$

5.7 Correlation

- The population correlation is denoted by $\rho_{X,Y}$;
- The sample correlation is denoted by $r_{X,Y}$.

Like variance, covariance is scale-dependent. This means conversions like meters to millimeters will change its value. This is undesirable when discussing the extent to which one variable affects another. Covariance is therefore usually standardized, such that it is always between -1 and 1. This standardized covariance is called **correlation** (eq. 5.12). In this equation, s_X and s_Y are the standard deviation of X and Y, respectively (eq. 5.6).

$$\rho_{X,Y} = \frac{\text{Cov}[X, Y]}{\sigma_X \cdot \sigma_Y} \tag{5.11}$$

$$r_{X,Y} = \frac{q_{X,Y}}{s_X \cdot s_Y} \tag{5.12}$$

Perfect correlation ($\rho = 1$) means that a change in one variable is precisely proportional to a change in the other variable. Likewise, perfect negative correlation ($\rho = -1$) means that an increase in one variable always corresponds to the same proportional decrease in the other variable. A correlation of (almost) zero means the covariance is (almost) zero. In other words, X and Y are not linearly related.

Correlation 'looks' the same as covariance. However, since correlation is standardized, its actual value is meaningful in terms of strength of association.

Note that neither covariance nor correlation implies a *causal* relationship, no matter how strong or significant the correlation is. Figure 5.3 shows six different directions of causality, all of which could be the reason for observing a (strong) correlation.

A correlation between X and Y could mean that X causes Y, or the other way around, or a third variable Z could influence the observed correlation. Finally, always bear in mind that a random sample of X could correlate coincidentally with Y, even though the population correlation between X and Y is zero.

The only way to demonstrate causality is by means of intervention: Intentionally changing X to observe its effects on Y (paragraph **3.2**).

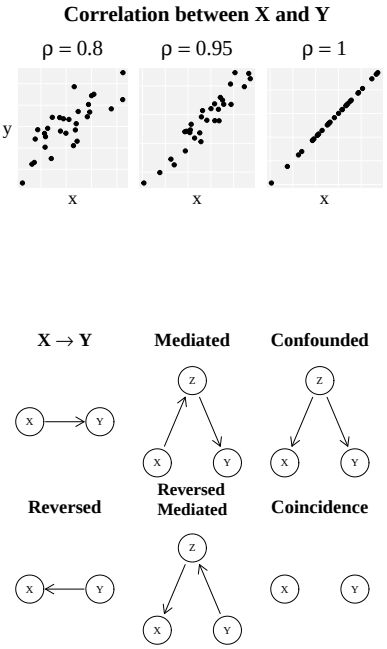


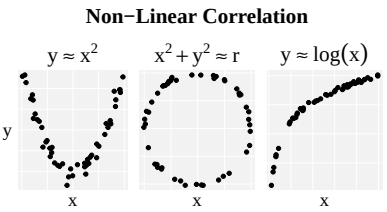
Figure 5.3: All of the causal directions shown above can result in a (strong) correlation.

5.8 *Non-linear correlation (*)*

See the examples on the right. Each of these has a strong correlation, but none of these have a strong *linear* correlation (eq. 5.12).

Non-linear relationships can be endlessly varied, so there is no general way to determine how change in one variable affects the other. However, if the relationship is monotonic (strictly increasing or strictly decreasing), then you can use *rank correlation* instead. Two such measures are commonly used: Spearman's ρ and Kendall's τ.

Since ranks are being calculated, there is always the possibility of ties (two values with the same rank). In the absence of ties, confidence intervals for Kendall's τ provide slightly better coverage. When ties are present, confidence intervals for Spearman's ρ provide slightly better coverage.



5.9 Examples in R

First, a fictive population is created and a sample drawn from it, then the estimates are made and compared to the population. Note that this only serves to illustrate concepts like precision. The population is generally not known in research.

As you will see, most of these estimates have a standard function in R for their calculation.

Simulating Data (*)

```
set.seed(1234)     # Ensures you and I see the same 'random' numbers
N          <- 10^6 # Population size (here: N = 1,000,000)
mu         <- 2.5  # Population mean
sigma      <- 2    # Population standard deviation
Population <- rnorm(1000, mean = mu, sd = sigma) # Generates the population
```

Drawing a Sample

```
n <- 30                          # Sample size
y <- sample(Population, size = n) # Draw a random sample
```

Sample Mean

```
mean(y)
```

```
[1] 2.676472
```

This value is fairly close to the true value of $\mu = 2.5$ we simulated.

Sample Variance

```
var(y) # R assumes you want the sample variance and thus divides by n - 1
```

```
[1] 5.884081
```

Standard Deviation

```
sd(y)
```

```
[1] 2.425712
```

This value is fairly close to the true value of $\sigma = 2$ we simulated.

Standard Error

```
(SE <- sd(y) / sqrt(n)) # Brackets around this line print the outcome
```

```
[1] 0.4428725
```

Confidence Interval

```
coverage <- c(0.025, 0.975)    # This range contains 95%
mean(y) - qnorm(coverage) * SE # Try running qnorm(coverage) separately
```

```
[1] 3.544487 1.808458
```

The interval ranges all the way from 1.81 to 3.54. Try increasing n and rerunning the whole simulation to see its effect on the confidence interval.

Better Confidence Interval (*)

The precise calculation of a confidence interval for this type of data uses a t-distribution (paragraph ***6.4***) with $n - 1$ degrees of freedom.

$n - 1$ because 1 degree of freedom was already used to calculate the mean.

```
coverage <- c(0.025, 0.975)        # 95% is within these values
mean(y) + qt(coverage, n - 1) * SE # Lower bound; upper bound
```

```
[1] 1.770697 3.582248
```

Simulating a Second Population, Correlated to the First (*)

```
set.seed(5678)         # Ensures consistent 'random' numbers
N2          <- 10^6    # Population size (here: N = 1,000,000)
mu2         <- 1       # Population mean
sigma2      <- 3       # Population standard deviation
Population2 <- rnorm( # Generates population 2 (correlated)
  1000, mean = mu + 0.5 * Population, sd = sigma
)
```

Drawing a Second Sample

```
n <- 30                            # Sample size
x <- sample(Population2, size = n) # Draw a random sample
```

Covariance

```
cov(x, y)
```

```
[1] -1.902282
```

Correlation

```
cor(x, y)
```

```
[1] -0.3709102
```

Rank Correlation (*)

```
cor(x, y, method = "spearman") # Spearman's rho
```

```
[1] -0.3704116
```

```
cor(x, y, method = "kendall")  # Kendall's tau
```

```
[1] -0.2413793
```

The example above shows the inefficiency of using non-linear correlation to observe a linear relationship. In this case, Spearman's ρ performed well, but Kendall's τ is quite far off the actual correlation. Below is an example where non-linear correlation is appropriate:

```
x <- 1:1000 # The numbers 1 to 1000
y <- log(x) # A perfect logarithmic relationship. Try: plot(y ~ x)
cor(x, y)   # Linear correlation
```

```
[1] 0.8733047
```

```
cor(x, y, method = "spearman") # Spearman's rho
```

```
[1] 1
```

```
cor(x, y, method = "kendall")  # Kendall's tau
```

```
[1] 1
```

6 Distributions

Statistical tests & models assume a distribution of the underlying process. All values may be equally likely, or some may be more likely than others.

A probability distributions describes the chance of different outcomes of a random event. There are two types of distributions: Discrete can only take on certain values, while continuous can take on any value within a certain range.

Discrete

Imagine you flip a coin, roll a six-sided die or count the number of birds outside. The coin has only two possible outcomes (heads or tails), the die has 6 $(1, 2, 3, 4, 5, 6)$ and the number of birds could be any non-negative integer (0, 1, 2, ...). Even though there are theoretically infinite possible outcomes for the number of birds, we will never count 0.99 bird or $\frac{1}{2}$ bird, or any other non-integer number. Such fixed outcomes are called discrete.

Figure 6.1: A 6-sided die has 6 possible outcomes and is thus discrete.

Continuous

Imagine you measure the height of a person in meters. It could be 1.80, 1.79, or 1.793, or anything in between. There is no limited number of values it can have. The only discreteness in height comes from our inability to measure it with infinite precision. A variable that can take on any value within a certain range is continuous. Examples include time, weight and concentration.

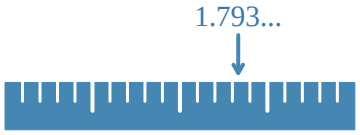


Figure 6.2: Height can take on any value within a certain range and is thus continuous.

Probability & Infinity (*)

Since continuous variables have uncountably infinite different outcomes, the probability of any particular value is: $\lim_{x \to \infty} \frac{1}{x} = 0$. We *know* it will take on a value, we just don't know which. This leads to the counter-intuitive conclusion that events with zero probability (0% chance) *can* actually happen. Likewise, events with probability 1 (100% chance) need not necessarily occur. Hence, mathematicians say that events with zero probability happen *almost never* and events with probability 1 happen *almost surely*.

6.1 Uniform Distribution

The uniform distribution is the simplest distribution, implying all outcomes within a certain range have an equal probability of occurring. Its shape is therefore rectangular (fig 6.3, 6.4). It has two parameters, which are its minimum and maximum.

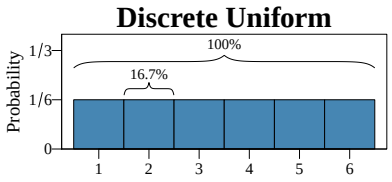


Figure 6.3: The chance of an outcome is one divided by the number of outcomes (here: one of six sides of a die).

Discrete

Suppose you roll a six-sided die: each outcome has an equal chance of $\frac{1}{6}$. This follows a discrete uniform distribution (fig. 6.3).

Continuous

Suppose you spin a perfectly round wheel and mark the position that is facing is up after spinning. Since the surface is round, there are infinite unique positions that can come up. If the result is measured in degrees, then this process is uniformly distributed between 0° and 360° (fig. 6.4).

The chance of any individual outcome is zero, since there are infinite numbers between 0 and 360. However, the chance of an outcome within a certain range is non-zero. For example, a number between 90° and 180° has a chance of $\frac{180-90}{360} = \frac{1}{4}$, or 25% of ending up.

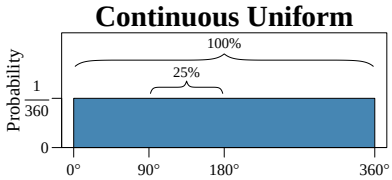


Figure 6.4: The chance of an outcome between two positions is the fraction of the total range it occupies (here: $\frac{1}{4}$ of a 360° circle).

6.2 Normal Distribution

The normal distribution is a continuous distribution that implies central tendency. In other words: Observations are expected to be close to the mean, with greater deviations in either direction being exceedingly rare.

While many distributions satisfy this condition, the normal distribution in particular is used very frequently in statistical models. It has two parameters: mean (μ) and variance (σ^2), which are its location and dispersion. The square root of the variance is called the standard deviation (σ) and is useful because it has the same unit of measurement as the mean (see paragraph ***4.3***).

Scientific articles often report the mean and standard deviation of variables to give the reader an idea of what values to expect. Namely, approximately 68% of observations from a normally distributed variable are within one standard deviation from the mean and approximately 95% are within two standard deviations (fig. 6.5).

- **Reporting means and standard deviations is only meaningful for (approximately) normally distributed variables.**

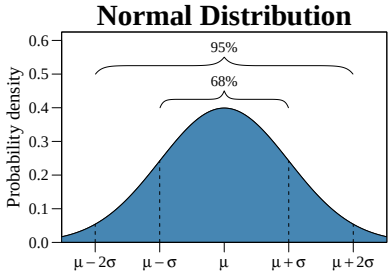


Figure 6.5: The normal distribution with mean μ and variance σ^2.

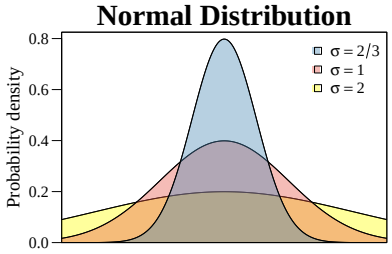


Figure 6.6: Normal distributions with the same mean and different variances.

The name 'normal' distribution is a bit misleading: Data are not 'ordinarily' generated by a process that follows a normal distribution. However, the normal distribution is ubiquitous in statistics. Both for its mathematical convenience, and its relationship to other probability distributions. Some of these reasons are elaborated on in the (optional) paragraph ***6.3***.

6.2.1 *Chance of Values within a Certain Range (*)*

The chance of values within a particular range for the normal distribution is calculated in the same manner as any other continuous distribution: Calculate the area under the curve within that range (fig. 6.7). Specifically, to calculate the chance of observing a value between a and b, take the integral of the probability density function (eq. 6.1) from a to b. The integral of the probability density function is called the *cumulative* density function, and is shown in equation 6.2. For all intents and purposes, it is much easier to just ask R:

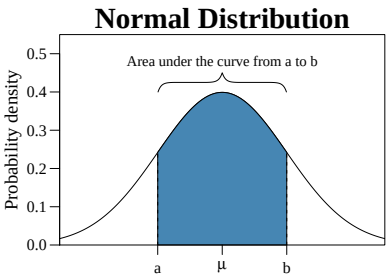


Figure 6.7: The normal distribution with mean μ and variance σ^2.

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6.1}$$

$$\frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} \mathrm{d}x \tag{6.2}$$

```
mu    <-  0 # The mean of the distribution
sigma <-  1 # R takes std. dev. as an argument, not variance
a     <- -1 # From where...
b     <-  1 # ...to where
pnorm(b, mu, sigma) - pnorm(a, mu, sigma)
```

```
[1] 0.6826895
```

As you can see, for the standard normal distribution ($\mu = 0$, $\sigma^2 = 1$), there is about a 68% chance of values between -1 and 1 (which equals $\mu \pm 1 \cdot \text{SD}$ here). Try changing a variable to see what happens.

6.3 *Is my variable normally distributed?*

Probably not. The normal distribution is continuous from $-\infty$ to ∞, is your variable too? Do observations of your variable form a symmetric bell-shaped curve that can be matched by a theoretical normal distribution? Do you even have enough observations to demonstrate this? Even if your variable follows an exact normal distribution, there is no real way to prove this. However, before resorting to non-normal tests or models, first consider if this is a problem:

- Most methods require *approximate* normality, not exact normality. This can be reasonably assessed with a QQ-plot (section ***6.3.2***);
- Normality of the *conditional* distribution is usually assumed. Not normality of the *data*. The conditional distribution is the remainder after accounting for effects like groups.[1]

[1] For example, a t-test assumes that the observations minus the mean of the group they belong to, are approximately normally distributed.

Approximate normality is hard to judge from a histogram or density plot. For example, see figure 6.8 with samples of different sizes (n), all drawn from a standard normal distribution ($\mu = 0$, $\sigma^2 = 1$):

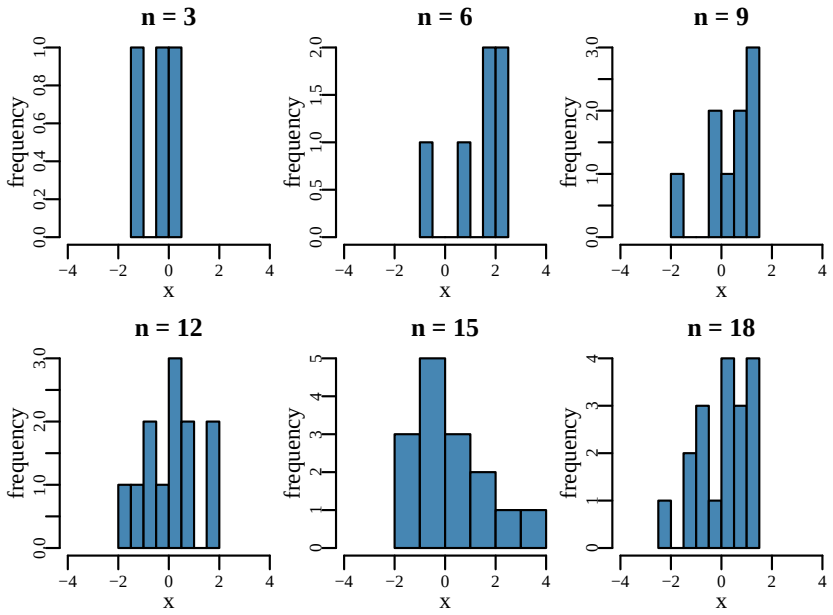


Figure 6.8: Histograms of 6 samples from the normal distribution with different sample size (n).

All these data come from a normal distribution, and yet several do not *look* normal. So can normality be demonstrated some other way?

In short, the answer is no. Empirical evidence cannot *prove* a hypothesis. Just because a sample does not deviate from normality, does not guarentee the underlying process is normal. By analogy: Just because the height of a person does not deviate from the average male height, does not guarentee this person is male.

The only thing that can be done is to *assume* normality and attempt to find evidence against this assumption. If no significant evidence is found, then the assumption might be reasonable. Otherwise, the assumption is rejected and non-normality is concluded.

6.3.1 *Shapiro-Wilk test*

The Shapiro-Wilk test is a hypothesis test for normality. This test generates a p-value[2] for deviation from a theoretical normal distribution. The table below shows p-values from the Shapiro-Wilk test using the 6 examples from figure 6.8.

[2] See chapter 7. For now, consider $p < 0.05$ as significant.

n	3	6	9	12	15	18
p-value	0.597	0.040	0.454	0.824	0.400	0.048

Table 6.1: Shapiro-Wilk tests on the 6 examples in figure 6.8.

Table 6.1 illustrates the problem with hypothesis testing for normality. These observations were explicitly drawn from a normal distribution and yet two p-values are below 0.05 and thus significantly non-normal according to the test. Moreover, p-values give no indication of the severity of a deviation, only whether it is significant.

6.3.2 QQ-plot

A quantile-quantile plot (QQ-plot) shows a theoretical normal distribution as a straight line.[3] Deviations from this line represent deviations from normality. Confidence intervals can also be included, to give and indication of the severity of the deviation.

Each of the 6 examples in figure 6.8 approximately follows the line representing a theoretical normal distribution (fig. 6.9). Moreover, the severity of any deviations can be observed using the plot. Deviations near the center of the plot are considered more severe than near the corners (bottom left and top right). The justification for this is that it is hard to tell whether an extreme observation actually comes from the extremes (tails) of the distribution, or is an outlier. The previously observed deviation at $n = 6$ can be attributed to a single deviation in the left tail of the distribution, which can be ignored. The deviation at $n = 18$ is hardly visible in the QQ-plot.

[3] It is difficult to judge whether a variable follows a bell-shaped curve, so the *quantiles* of the normal distribution are shown instead, which form a straight line. Hence the name.

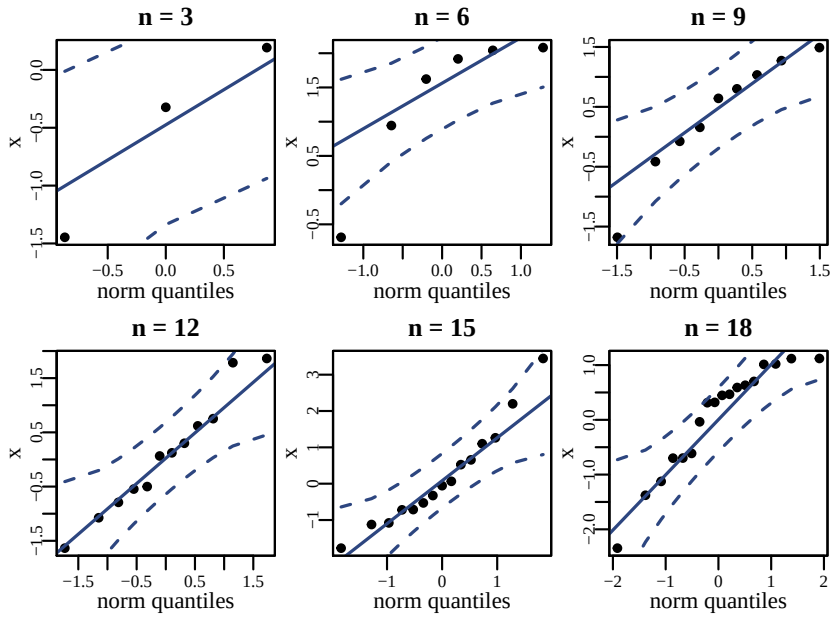


Figure 6.9: QQ-plots for normality of the 6 samples from figure 6.8. The solid blue line represents the quantiles of a theoretical normal distribution and the dashed lines indicate the confidence interval. The black dots represent the observations. Observations outside the confidence intervals are significant deviations from normality.

- **QQ-plots are more informative than tests for normality.**
- **If no deviation from normality is found, a variable is still merely *assumed* to be normally distributed. It cannot be proven.**

6.3.3 What if a Variable is not Normally Distributed? (*)

Consider whether this is important: Does the test or model require the variable or residuals (chapter **9**) to follow a normal distribution?

Next, assess the severity of the deviation: Does the deviation occur in the tail (less worrisome) or near the center (more worrisome); Does it involve one or multiple deviations; Could the deviation be caused by an outlier? You could also consider a transformation.[4]

Finally, if a variable is truly not normally distributed but the test or model requires it to be, then a different test or model should be used. Refer to chapters **8** and **9** for specific cases.

[4] Covered in Elements of Biostatistics.

6.4 T-distribution

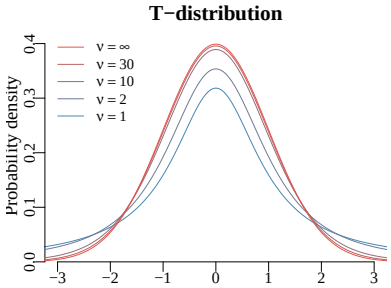


Figure 6.10: *T*-distributions with various degrees of freedom.

Recall that the normal distribution has only two parameters: mean (μ, its location) and variance (σ^2, a measure of dispersion). The *t*-distribution has an extra parameter: degrees of freedom (ν). It is a normal distribution with heavier tails, to take into account the uncertainty when the sample size (n) is limited. Hence, a *t*-distribution with $\nu = \infty$ is a normal distribution (fig. 6.10) and a *t*-distribution with more than 30 degrees of freedom is nearly indistinguishable from a normal distribution. The degrees of freedom are generally the sample size minus one ($\nu = n - 1$)

6.5 Other Distributions (*)

If you have a question that seems to fit with any of these probability distributions, have a look at the ***GLM*** chapter in the next book.

There are many other useful probability distributions in statistics. However, the models that use these distributions are beyond the scope of this introductory book. Nevertheless, to give you an idea of common probability distributions and to show why the normal distribution is the first you learn about, see the following table:

Distribution	Interpretation	Example	Relationship to normal distribution
Uniform	All chances are equal.	Rolling a 6-sided die.	The sum of ∞ uniformly distributed effects follows an exact normal distribution.
Normal	Central tendency, quadratically decreasing chance of further deviations.	The sum of many independent effects, like blood pressure, IQ, or height.	

Continued on next page

Distribution	Interpretation	Example	Relationship to normal distribution
Binomial	Successes out of a certain number of trials.	The number of blue colonies out of all colonies on an X-gal plate.	As the number of trials goes to ∞ and the chance of success tends to $\frac{1}{2}$, a binomial distribution approximates a normal distribution.
Poisson	Number of independent events within a certain time or space.	Counting the number of plants within a certain area.	As the mean count goes to ∞, the Poisson distribution approximates a normal distribution.

Table 6.2: An overview of common probability distributions.

As you can see in the table above, many distributions can be reasonably approximated by the normal distribution under some cirumstances. The figure below illustrates this.

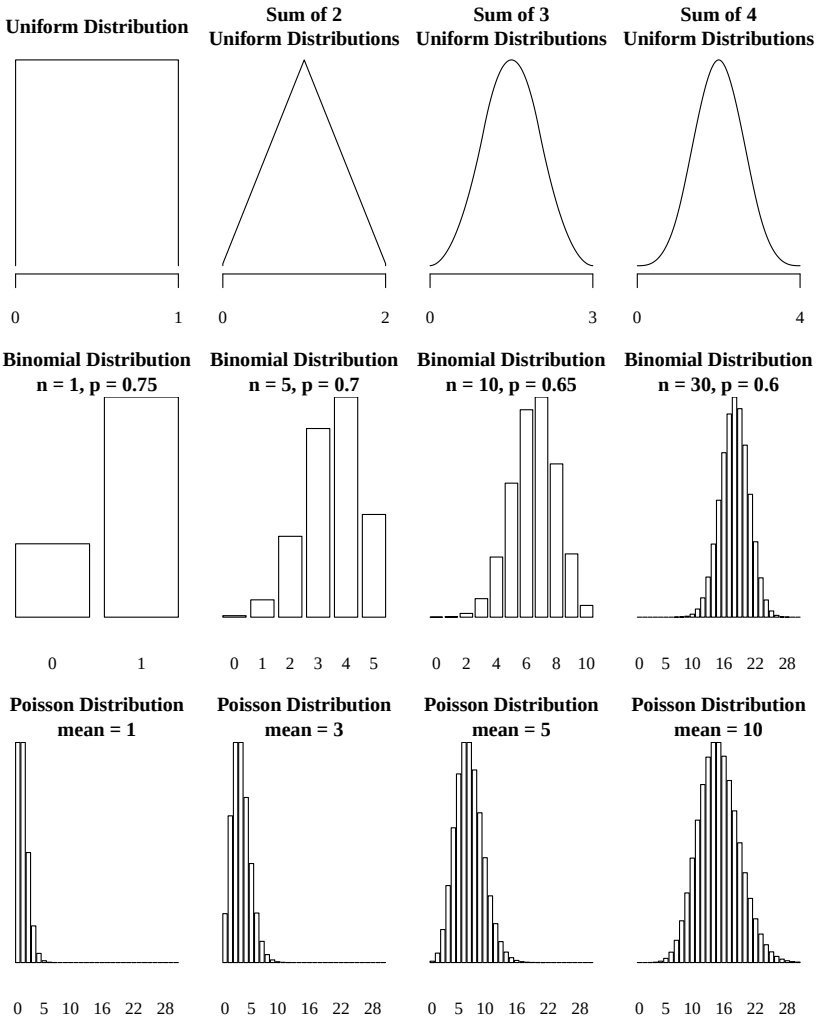


Figure 6.11: Visualization of the relationship between the normal distribution and other common distributions. The plots further to the right begin to resemble a normal distribution.

6.6 *Examples in R*

Shapiro-Wilk Test (a QQ-plot is usually better!)

```
x <- rnorm(30) # 30 random samples from the standard normal distribution (example)
shapiro.test(x)
```

QQ-Plot

```
library('car') # Required. Install if needed: install.packages('car')
x <- rnorm(30) # 30 random samples from the standard normal distribution (example)
qqPlot(x)
```

7 Hypothesis Testing

A fundamental problem in empirical science is deciding on truth. For example: Does smoking increase the risk of cancer? Decades of supporting evidence certainly suggest so, but in theory all smokers who developed cancer could have done so by extreme coincidence. Deciding when to stop believing in such coincidence is the essence of hypothesis testing.

Hypothesis testing addresses this issue by expressing the likeliness of observing something 'by coincidence'. The statement implying coincidence is the *null-hypothesis* (H_0) and if it is unlikely to be true based on observations, the result is considered *significant*. For example, if the null-hypothesis is that there is no difference in cancer prevalence between smokers and non-smokers, then the difference under the null-hypothesis is zero. The further we move away from zero, the less likely it becomes that the real difference is zero.

By convention, significance is achieved if there is a chance smaller than 5% of observing something at least this extreme under the null-hypothesis. However, 5% is arbitrary and some fields use 1%, or even lower, such as the 5σ-rule in particle physics, which corresponds to about 0.000035% (fig. 7.1). Ultimately the choice is up to the scientist: *When are you convinced?* Since this is subjective, replication studies are often performed, which reproduce the entire study to see if the same conclusions are reached.

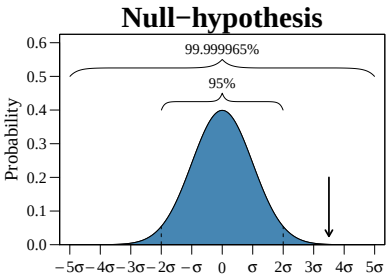


Figure 7.1: In biology, a result about 2σ, or two standard deviations away from the null-hypothesis is considered significant (less than 5% chance of 'coincidence'), while in particle physics 5σ is the norm (less than 0.000035%). The difference from zero indicated by the black arrow is significant at the 5% level, but not at the 0.000035% level.

Lastly, *power* is the ability to achieve significance if the null-hypothesis is in fact not true. For example, low power due to low sample size might not yield significant results even though the null-hypothesis is not true. Power should be sufficiently high, such that a non-significant result convinces the researcher there is need to repeat the experiment with a larger sample size. This is another subjective choice, often taken to be at least 80-90%.

- **Statistical significance does not necessarily imply biological relevance. Always take the size of estimates into consideration.**

7.1 From Research Question to Null-Hypothesis

Forming a null-hypothesis starts with a scientific curiosity and ends as a formal statement that can be disproven. For example:

Scientific problems

- *What is the relationship between smoking and cancer?*
- *What is the effect of vitamin C supplements on cellular growth?*
- *How well can we discern between plant species using morphological attributes?*

To address a scientific problem, a research question must be restricted to something that can be answered in a single study. This implies shifting from an open question to a closed question:

Research questions

- *Does smoking contribute to oesophageal cancer development?*
- *Does supplemental vitamin C affect odontoblast*[1] *growth in guinea pigs?*
- *Is there a difference in sepal length between* I. setosa, I. versicolor *and* I. virginica*?*

[1] Cells responsible for tooth growth.

Although these are valid research questions, they still cannot be expressed in numbers or a simple yes or no. Hence, the next step is to define a formal statement, which reflects our belief about the research question as close as possible. Whether it is true or false is the essence of hypothesis testing.

Hypotheses

- *Oesophageal cancer is more prevalent in smokers.*
- *Supplemental vitamin C enhances odontoblast growth in guinea pigs.*
- I. setosa *has shorter sepals than* I. versicolor*, which has shorter sepals than* I. virginica*.*

One final piece of ambiguity remains. Namely, these hypotheses could be true or false, depending on what we consider to be significantly more, enhanced or shorter. To solve this problem, we instead define a statement that can be falsified: The null-hypothesis. This is an unambiguous statement of no difference.

Null-hypotheses

- *Oesophageal cancer prevalence is* ***equal*** *in smokers and non-smokers.*
- *There is* ***no difference*** *in mean tooth length between supplements.*
- I. setosa*,* versicolor*, and* virginica *have* ***equal*** *sepal length.*

Contrary to having to define what we consider to be *more*, *enhanced* or *shorter* for every single study, we just have to agree on what is significantly different from **zero**.

7.2 Level of Significance

The level of significance is the threshold for concluding a result is significant. In most research, the accepted chance of a type I error (false positive) is at most 5%, or $\alpha = 0.05$.[2] As mentioned before, this choice is somewhat arbitrary and other levels of significance are also possible (e.g. $\alpha = 0.01$, $\alpha = 0.005$), depending on the severity of making a type I error.

It might be tempting to choose a very small value of α to reduce the chance of a false positive. However, the stricter the level of significance, the harder it will be to detect a real difference. On the other hand, a very lenient value (e.g. $\alpha = 0.1$) might yield more significant results, but is also prone to result in a false positive.

- **The level of significance (α) is decided prior to analysis.**

[2] A type I error (false positive) is to reject the null-hypothesis, when in reality it is true. For example, a test reveals a patient has an illness, while in reality the patient is healthy. This occurs at a rate of α.

7.3 Confidence Intervals

Confidence intervals are not only useful for determining the quality of an estimate (chapter 4), but also to determine significance. To determine significance at the $\alpha = 0.05$ level of significance, a 95% confidence interval (eq. 5.8) is constructed. If the value under the null-hypothesis is not included in the interval, the estimate differs significantly from it (fig. 7.2).

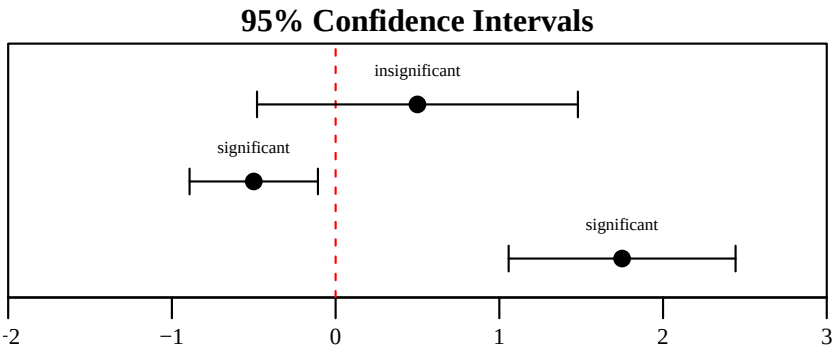


Figure 7.2: A 95% confidence interval for an insignificant (top) and two significant (middle, bottom) differences from the value under the null hypothesis (here: zero). The black dots are the estimated parameters. Note that although the middle estimate is significant, the estimate is quite small.

Similarly, a 99% confidence interval can be used to test at the $\alpha = 0.01$ level. In general $100 \cdot (1 - \alpha)\%$ confidence intervals can be constructed to test at any level of α. Refer to the examples in R at the end of this chapter.

7.4 One-Sided Testing

In figure 7.2, both positive and negative differences could be significant, which is called two-sided testing and is the default method for hypothesis testing. When there is only interest in a single direction, one-sided testing is more powerful.

For example, if a scientist wants to demonstrate that an expensive medicine A works better than a cheaper medicine B, interest only lies in whether A $>$ B. If the expensive medicine does not perform significantly better (i.e. A $\leq$ B), there is no reason to use it.

To test one-sided at $\alpha = 0.05$, either a lower or upper bound is estimated, but not both. Figure 7.3 illustrates testing for positive differences and figure 7.4 illustrates testing for negative differences.

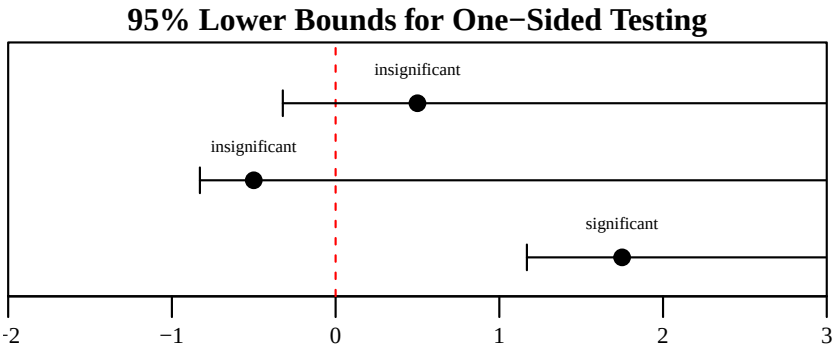


Figure 7.3: If interest only lies in positive differences, only lower bounds of 95% confidence intervals are calculated. The middle estimate was significantly different from zero in fig. 7.2, but not significantly *greater* than zero. Only the bottom estimate is significantly greater than zero.

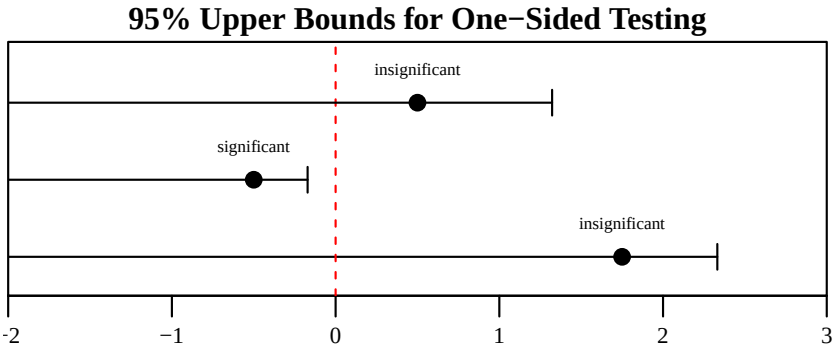


Figure 7.4: If interest only lies in negative differences, only upper bounds of 95% confidence intervals are calculated. The bottom estimate was significantly different from zero in fig. 7.2, but not significantly *less* than zero. Only the middle estimate is significantly less than zero.

A close inspection of figures 7.3 & 7.4 reveals that one-sided bounds are closer to the estimate than in figure 7.2. The reason for this will be explained in the following paragraph.

- **Use one-sided testing if and only if there is no interest in differences in the other direction. If you are unsure, test two-sided. Choosing to test one-sided must be done *before* any analysis.**

7.5 P-Values

Despite being inferior to confidence intervals, p-values are still widely used in empirical science. It is therefore important to develop a good understanding of the meaning of this value.

The p-value is the chance of obtaining results this far from the null-hypothesis, if the null-hypothesis were in fact true. It is area under the curve of the null-hypothesis (fig. 7.5). If the p-value is smaller than the accepted chance of a false positive (i.e. $p < \alpha$), the estimate is significant.

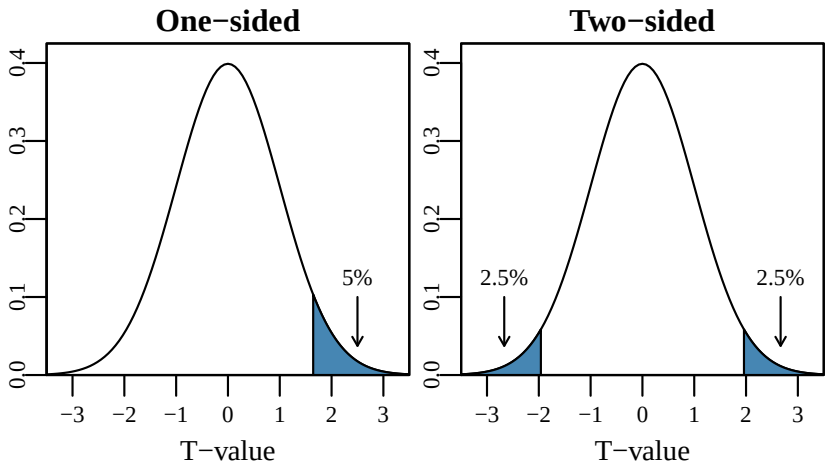


Figure 7.5: A p-value is the chance of finding a test value at least as extreme as the one calculated from the sample, under the assumption the null hypothesis is correct. The distribution of the null hypothesis of a T-test is the T-distribution and the chance at either side is the blue surface.

The p-value is the size of the blue surface(s) in figure 7.5 and can be calculated using an integral. R generally includes p-values in the output of statistical tests.

Limitations of P-Values (*)

The use of p-values has been heavily criticized for several reasons. Most importantly, they do not convey anything about the size of the estimate. An estimate may be smaller than biologically relevant, but still significant because the sample size was large enough, the variance small enough or just by coincidence. Remember that $\alpha = 0.05$ implies that a false positive *will* be made in about 5% of hypothesis tests where there is no true effect.

A more fundamental problem with p-values is that they are based on the null-hypothesis, which is 'made up' data. The probability under the curve is based on the assumption that this distribution for the null-hypothesis is correct. Confidence intervals on the other hand, can be constructed using the actual data.[3]

- **Use confidence intervals instead of p-values whenever possible.**

[3] For example, by bootstrapping, which is covered in Advanced Statistics.

7.6 Power

What if a difference is not significant? Perhaps there is no difference after all, but it is also possible that a type II error (false negative) is made.[4] The ability of a test to detect a real difference is called power ($1 - \beta$) and determines the chance of a false negative (β).

Commonly, $\beta \leq 0.10$ or $\beta \leq 0.20$ are considered acceptable in life sciences, implying a maximum chance of 10% or 20% of a false negative, respectively. These are arbitrary numbers — the sample size should be chosen such that if there is no significant effect, you are convinced there is no real effect.

[4] A type II error (false negative) is failure to reject the null-hypothesis, even though it is false. For example, a test does not detect an illness, even though a patient is ill. This occurs at a rate of β.

- **The power of a test is not chosen, but depends (among others) on the sample size. Choose a sample size such that the power is at least the desired amount.**

7.6.1 Power & Sample Size

An estimate of the minimum required sample size is often required prior to experimenting, be it due to ethical considerations, time or other resources. Online calculators are available, which are fine when the goal is clear and the model is simple (e.g. t-test). However, there is no 'general' calculation of power, as it depends on the type of model used. Power calculations will be included in relevant chapters. As a rule of thumb, the relationship between power and the sample size, effect size,[5] variance and level of significance is as follows:

[5] Effect size is the magnitude of the estimate divided by its variance. For example, the difference in male height and female height in centimeters is different than that in meters, but the effect size is the same.

- Greater sample size increases power
- Greater effect size increases power
- Greater variance in the outcome decreases power
- Stricter level of significance (i.e. lower α) decreases power

Stopping Rule

In some research, more data is collected until significance is achieved. Although justifiable in certain cases, this is generally not recommended, as it is biased towards significance. The p-value calculation can be adjusted for the stopping rule, but this is beyond the scope of this book.

If there is doubt about the required sample size, the correct way to determine it is by a pilot study or by simulation. Data from a pilot study is not included when assessing significance in the actual study.

- **The sample size should be determined prior to experimenting. Stopping rules should be avoided.**

7.6.2 Relationship Type I & II Error

There is an inverse relationship between the type I and type II error: A stricter level of significance (e.g. $\alpha = 0.01$) will make it harder to conclude a significant effect (after all, we are more stringent about significance). Hence, the chance of *not* being able to detect an effect (β) has increased. The only way to decrease the chance of a type II error without affecting the chance of a type I error is by increasing the sample size (n).

	H_0 true	H_0 not true
H_0 rejected	α (type I error)	$1 - \beta$ (correct)
H_0 not rejected	$1 - \alpha$ (correct)	β (type II error)

Table 7.1: Relation between chosen level of significance (α), power ($1 - \beta$) and the type I & II errors.

7.7 Cautionary Note

Although statistics is often associated with significance, keep in mind that significance is *only* relevant for confirmation. Overuse of significance tests—both by confidence intervals and by p-values—leads to a battery of problems, including:

- Reporting 'significant' findings when estimates are in fact smaller than might be considered relevant
- Omitting insignificant estimates
- Multiple testing,[6] leading to inflated false positive rates
- Significance as a requisite for publishing (publication bias)

[6] Covered in Elements of Biostatistics

Furthermore, keep in mind that hypothesis testing is only meaningful if the null-hypothesis reflects the opposite of our beliefs as closely as possible. The null-hypothesis is supposed to imply coincidence, but, its true interpretation is as follows:

> The chance of observing a result at least as extreme as the one observed in the experiment, if the null-hypothesis is in fact true.

Lastly, if the null-hypothesis is not rejected, this does not prove the null-hypothesis is true. In fact, it cannot be proven that the null-hypothesis is true. Therefore, when a result is insignificant, the conclusions that can be drawn are limited. This leads some statisticians to recommend doing away with hypothesis testing altogether and instead focus on parameter estimation and expressing the quality of the estimates.

- **Only use hypothesis testing if the objective is confirmation.**

8 Statistical Tests

This chapter covers the bread and butter of hypothesis testing: The t-test, which assesses the significance of a difference in means and the χ^2-test, which compares observed and expected frequencies. For example, male and female height can simply be expressed as a difference in means (t-test), but prevalence of smoking in males and females is better expressed in frequencies (χ^2-test).

Lastly, the F-test is introduced, which compares variances. For example, the omnibus test of an ANOVA (section ***9.1.4***) compares the residual variance to the total variance of the variable of interest.

8.1 T-Test

A t-test is used to assess the significance of a difference in means. The t-value (eq. 8.1 for one-sample and eq. 8.2 for two-sample) can be used to construct a confidence interval, or calculate a p-value. R includes both in the output. The larger the t-value gets, the more significant a result (i.e. confidence interval further from zero, and a lower p-value). There are several types of t-tests for different comparisons.

$$t = \frac{\bar{y} - \mu_0}{\text{SE}}$$
$$\text{SE} = \frac{s}{\sqrt{n}} \tag{8.1}$$

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\text{SE}_\text{p}}$$
$$\text{SE}_\text{p} = \text{s}_\text{p} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{8.2}$$

s_p : See two-sample t-test.

8.1.1 One-Sample T-Test

A one-sample t-test compares a sample mean ($\bar{y}$) with a specific value (μ_0). For example: Does average male height differ significantly from 1.70m? Figure 8.1 is a visual representation of such a comparison: The mean of a group of males is compared to a specific value.

The value of t grows as the difference ($\bar{y} - \mu_0$) gets larger, or the standard error (SE) gets smaller. The latter implies that the more precise we know the value of the mean, the smaller a difference will be significant. Since this precision depends on sample size (n), a larger sample size is more likely to yield significant results.

- **Examples of each t-test are included at the end.**

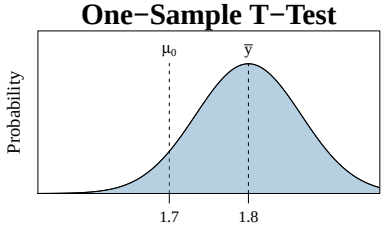


Figure 8.1: Visual representation of a one-sample t-test.

8.1.2 Two-Sample T-Test

A two-sample t-test compares the mean of one group ($\bar{y}_1$) with that of another group ($\bar{y}_2$). For example: Does average male and female height differ significantly? Figure 8.2 illustrates this type of comparison: The mean of a group of males is compared to the mean of a group of females.

Since there are two groups, there are also two variances. Because of this, another choice must be made in a two-sample t-test: Is the variance of groups equal or unequal?

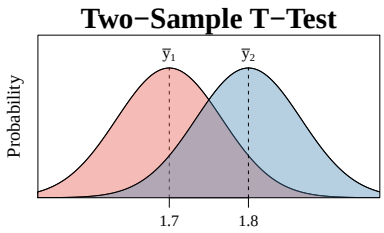


Figure 8.2: Visual representation of a two-sample t-test. The variances are equal in this example

Equal or Unequal Variance

In two-sided t-tests (eq. 8.2), the t-value depends on the pooled variance (s_p^2), which is automatically estimated in R. However, if groups have equal variance, the estimation can be simplified, rendering the test more powerful.

An F-test (paragraph **7.3**) can determine whether variances differ significantly. If no significant deviation from equal variance is found, an equal variance t-test may be used. A t-test for unequal variance is called a Welch t-test. By default, R assumes variances unequal.

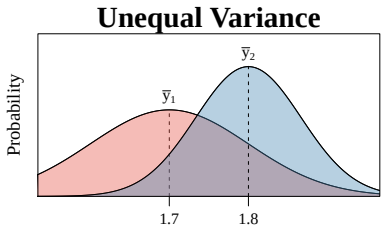


Figure 8.3: Example of unequal variances. The variance of the red group is larger than that of the blue group.

Independent or Paired

Observations from the same experimental units are paired.[1] For example, a clinical trial with a placebo group can be unethical if both groups of patients require immediate treatment. An alternative way to study the effectiveness of a treatment is by measuring some outcome variable *before* and *after* the treatment. Since these measurements are performed on the same individuals, the data is paired.

Another example is measuring two outcome variables on the same units. In the `iris` dataset, the length and width of sepal and petal leaves were measured on each flower. If a comparison were made between *I. setosa* sepal length and petal length, the data is paired. However, if a comparison were made between *I. setosa* and *I. virginica* sepal length, the data is unpaired, because these measurements were performed on different experimental units.

[1] An experimental unit is the smallest division to which different treatments can be applied. For example, individual patients, plants, animals, or subdivisions of land in ecological research. The experimental unit depends on the research question and is covered in depth in Elements of Biostatistics.

8.1.3 Non-Parametric Alternative: Wilcoxon Test

Non-parametric means *without assuming a distribution*. Some data is not approximately normally distributed, or not even symmetric. In such cases, the mean is not informative (fig. 8.4). Since a t-test compares means, it should not be used for non-normal variables.

The easiest non-parametric alternative for a comparison of two groups is the Wilcoxon test. This test does not compare means, but

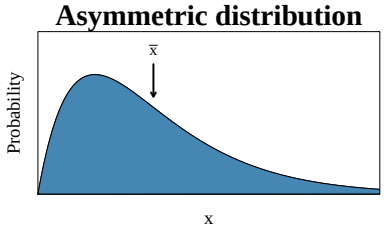


Figure 8.4: Example of non-normal data. The mean is uninformative for comparisons, because it is not the point of the highest probability (the mode).

rather the rank of observations in both groups by ordering them from high to low. The interpretation of the test is similar, except that a difference in *location* or *rank* can be concluded, rather than a difference in means.

Note on Non-Parametric Statistics

Generally speaking, non-parametric tests are less powerful and more complex. Many biological processes are on a logarithmic or exponential scale, meaning a simple transformation can yield a normal distribution. Moreover, if some other distribution than normal can be assumed, there are better alternatives, including generalized linear models (GLM) and generalized additive models (GAM).

- **Transformation & GLMs are covered in Elements of Biostatistics.**

8.2 χ^2-Test

The χ^2-test (*chi-squared*) compares observed to expected frequencies. This paragraph covers a number of tests based on the χ^2-statistic (eq. 8.3). It is the most basic tool for analysing count data. The only requirement for a χ^2-test is that the expected counts are 5 or greater.

$$\chi^2 = \sum_{i=1}^{k} \frac{(\mathbf{O}_i - \mathbf{E}_i)^2}{\mathbf{E}_i} \tag{8.3}$$

Where k is the number of categories, $\mathbf{O}_i$ the observed frequency of category i, $\mathbf{E}_i$ the expected frequency of category i.

- **A statistical model for count data is included in Elements of Biostatistics (Poisson GLM).**

8.2.1 Goodness-of-Fit

A simple example of a χ^2-test is whether a set of numbers differ significantly from equal probabilities. For example, to assess whether a 6-sided die is fair, it could be thrown 60 times (table 8.1).

Side	⚀	⚁	⚂	⚃	⚄	⚅
Rolls	5	13	9	10	12	11

Table 8.1: Outcome of 60 die rolls.

The χ^2 test considers all deviations from the expected frequencies (each side should come up $10\times$) and calculates a p-value to decide whether these deviations are significant.

- **See the example in R at the end of the chapter.**

Explanation of the Name (*)

Testing whether a discrete set of outcomes differ from equal probabilities is equivalent to testing whether this sample deviates significantly from a discrete uniform distribution (fig. 6.3). Hence, the test essentially considers how well these data fit a certain distribution.

The goodness-of-fit test is not limited to the discrete uniform distribution and can be used to compare the observed frequencies to expected frequencies from any discrete probability distribution.

However, as demonstrated in section ***6.3.1***, statistical tests might not be the best tool to determine the distribution a sample comes from.

- **Guidelines for deciding the most probable distribution are included in Elements of Biostatistics.**

8.2.2 *Contingency Tables*

Contingency tables, introduced in chapter **3** (table 4.3), can also be analyzed with a χ^2-test to determine whether there is significant over- or underrepresentation of a certain category.

For example, a study might be conducted on smoking and a critic might argue that there are relatively less smoking men than women in the sample (table 8.2). The expected values are calculated using the totals of the rows and columns (R does this automatically). If the resulting χ^2-value is high enough, the corresponding p-value will indicate significant evidence for deviation from an equal ratio of smoking and non-smoking men and women.

	men	women
smoking	68	82
non-smoking	41	33

Table 8.2: A 2×2 contingency table.

- **See the example in R at the end of the chapter.**

8.3 *F-Test*

An F-test is a comparison of variances (eq. 8.4). Its null hypothesis is that the variances follow a $1 : 1$ ratio (i.e. variances are equal). There are three uses of the F-test throughout this book:

- Test whether variances can be considered equal for a t-test;
- The omnibus test in an ANOVA;
- Test whether a significant amount of the variance in the response variable is explained by the model (simple linear regression).

The latter two are different interpretations of the same issue.

Keep in mind that the F-test is just a test. It has a chance of a false positive and gives no indication of whether the deviation from equal variance is problematic.

$$F = \frac{s_a^2}{s_b^2} \tag{8.4}$$

Where a and b are normally distributed.

Cautionary Note

F-tests are very sensitive to non-normality, which means the outcome is only meaningful if the variances come from populations that are normally distributed. Contrary to the t-test, the F-test should not be used to compare *approximately* normal populations.

Since it is not possible to prove exact normality, this means that careful consideration must be taken whether the assumption of normality is reasonable from a biological perspective.

8.4 Examples in R

Run the code in order to see the output.

Simulate Data (*)

```
set.seed(1234) # Ensures you and I see the same 'random' numbers
n     <- 30    # The sample size
mu_1  <- 1.80  # The population average of men
sigma <- 0.15  # The population standard deviation
men   <- rnorm(n, mu_1, sigma) # Sample of n = 30 from this population
mu_2  <- 1.70  # The population average of women
women <- rnorm(n, mu_2, sigma) # Sample of n = 30 from this population
```

One-Sample *T*-Test

```
t.test(men, mu = 1.75) # Do men differ significantly from 1.75m?
```

Two-Sample *T*-Test

```
t.test(men, women)                           # Two-Sided
t.test(men, women, alternative = "greater") # One-Sided (men > women)
t.test(men, women, alternative = "less")    # One-Sided (men < women)
```

F-Test for Equality of Variance

```
var.test(men, women)
```

T-Test for Equal Variance

```
t.test(men, women, var.equal = TRUE) # May be used if F-test is insignificant
```

Wilcoxon Test (Non-Parametric *T*-Test)

```
wilcox.test(men, women) # Two-sided, unpaired. Accepts same arguments as t.test()
```

Simulate Paired Data (*)

```
men2 <- men - rnorm(n, 0.04, 0.1) # 0.04 is the shift in mean and 0.1 adds noise
```

Paired T-Test

```
t.test(men, men2, paired = TRUE)
```

Power of a T-Test

This function allows one to estimate one of the following, provided the others are known:

- Sample size (n)
- Effect size ($\hat{y}_1 - \hat{y}_2$)
- Power (1 - β)
- Standard deviation (s)
- Level of significance (α)

In this example, the required sample size is calculated:

```
Effect <- mean(men) - mean(women) # Effect size
s      <- max(sd(men), sd(women)) # Using the greatest of the SDs is conservative
Power  <- 0.90 # Desired power: chance of correctly rejecting H0
alpha  <- 0.05 # Level of significance: chance of false positive

power.t.test(delta = Effect, sd = s, power = Power, sig.level = alpha,
             alternative = "two.sided")
```

To calculate the power given a sample size, omit `power` and enter `n`:

```
power.t.test(n = 30, delta = Effect, sd = s, sig.level = alpha,
             alternative = "two.sided")
```

χ-Squared Test (Goodness-of-Fit)

The 6 outcomes from section ***7.2.1*** (table 8.1) can be analyzed as follows:

```
Die <- c(5, 13, 9, 10, 12, 11) # Frequencies
chisq.test(Die) # Chi-squared test for goodness-of-fit
```

χ-Squared Test (Contingency Table)

The contingency table from section ***7.2.2*** (table 8.2) can be analyzed as follows:

```
C <- data.frame('men' = c(68, 41), 'women' = c(82, 33)) # Contingency table
chisq.test(C) # Chi-squared test for independence of smoking and gender
```

Since $p > \alpha$, there is no significant evidence for over- or underrepresentation of smoking or non-smoking men or women.

9 Statistical Models

Statistical models seek to describe the systematic part of a process, such that only the stochastic part (i.e. random noise) remains. For example, cancer incidence may be random for an individual,[1] but across the population it is largely influenced by contributing factors:

$$\underset{\text{e.g. cancer incidence}}{\textbf{biological process}} = \underset{\substack{\text{e.g. smoking behaviour,}\\ \text{genetic predisposition}}}{\textbf{systematic part}} + \underset{\substack{\text{e.g. random mutations,}\\ \text{environmental effects}}}{\textbf{stochastic part}} \tag{9.1}$$

[1] Small, unknown or unpredictable factors can influence whether an individual develops cancer. Think of UV-light, background radiation, accumulation of carcinogenic substances, random failure of DNA repair mechanisms, etc.

The systematic part is generally of interest, as it can help understand a biological process or predict the values of new observations. The stochastic part is also important, because it can help justify the model used through diagnostics. Namely, if the model captures the systematic part well, then the remainder should be random noise (i.e. only the stochastic part). However, if there is still structure left in the remainder, then apperently the model does not capture the systematic part well.

This chapter covers two basic models which form the basis for many more complex models: ANOVA & simple linear regression. For any parametric model,[2] the basic steps involved are as follows:

[2] All models covered in this book make assumptions on the underlying distribution. Such models are said to be parametric. Non-parametric regression is beyond the scope of this book series.

1. Define the research question;
2. Determine the model required to answer the research question. Can its assumptions be justified from a biological point of view?
3. Perform experiments, collect data;
4. Estimate the parameters of the model;[3]
5. Assess whether any assumptions are violated with diagnostics. In case of serious violations, consider using a different model or try transforming the data;[4]
6. (optional) Determine the significance of the parameters;
7. Interpret the outcome.

[3] See the examples in R in *8.3*.

[4] Covered in Elements of Biostatistics.

- **Statistical software often automatically includes p-values in model output, but keep in mind that hypothesis tests may not always be of interest.**

9.1 ANOVA

Analysis of variance (ANOVA) is a model for the means of categories of a factor. While *t*-tests can only be used to compare two categories, ANOVA can be used to compare any number of categories.

Suppose one wants to know whether three *Iris* species differ in mean sepal length (fig. 9.1). The ANOVA model would be as follows:

$$\underset{\text{response variable}}{y_{ij}} = \underset{\text{mean per group}}{\beta_j} + \underset{\text{residual}}{\epsilon_{ij}} \tag{9.2}$$

Read eq. 9.2 as follows: The sepal length of an observation i of species j is equal to the mean of its species (β_j) plus its particular deviation from that mean (ϵ_{ij}).

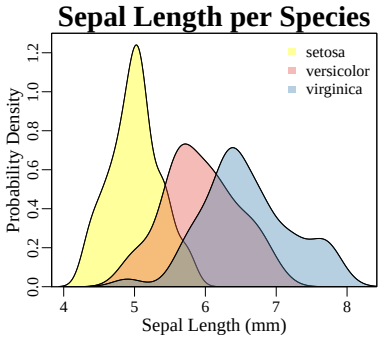


Figure 9.1: Estimated distribution of the sepal length of three *Iris* species.

9.1.1 Assumptions

ANOVA implicitly makes the following assumptions:

1. **Normality of the residuals**;[5]
2. **Equal variance of groups** (also called *homoscedasticity*);
3. **Independence of observations**;
4. **Balanced data**.[6]

Assumptions 1-3 are analogous to the choices made when choosing a *t*-test. Unlike the *t*-test, ANOVA is not as easily adapted for violations of these assumptions. Hence, a list of alternatives is presented in ***8.1.2***. The fourth assumption is less stringent and can often be relaxed as long as the variance of groups is still approximately equal.

[5] If all systematic components of the process have been accounted for by the model, the remainder (i.e. residuals) should be normally distributed noise.

[6] Balanced means that every category has an equal number of observations.

9.1.2 Estimation

The mean per species is estimated with the ANOVA method, or equivalently by *least squares*, which minimizes the sum of the squared residuals: ($\sum \epsilon_{ij}^2$). This ensures that as much systematic information as possible is contained in the model. By squaring the differences from the mean, two things are achieved:

- Positive and negative values would otherwise sum to zero;
- Squared values are more sensitive to large differences from the mean than e.g. absolute values.

Squared differences are a property of the normal distribution (eq. 9.3), where greater deviations from the mean are exceedingly rare. It also means that this method is not robust to outliers.

- **Performing ANOVA is simple in R. See the examples at the end.**

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{9.3}$$

The probability density function of the normal distribution squares the differences from the mean.

9.1.3 *Diagnostics*

Even if assumptions are justifiable from a theoretical point of view, visual diagnostics must always be performed to check for any serious violations. If none are found, then the assumptions hold and the model can be used for parameter estimation or hypothesis testing.

A QQ-plot (section **5.2.1**) can be used to assess whether the assumption of normality is reasonable (fig. 9.2). Some deviations beyond the confidence intervals can be observed in the upper-right part of the QQ-plot. However, using the guidelines from **5.2.1**, the overall trend is very similar to the line representing the theoretical normal distribution and the deviations occur in the tail, which is less worrisome than near the center.

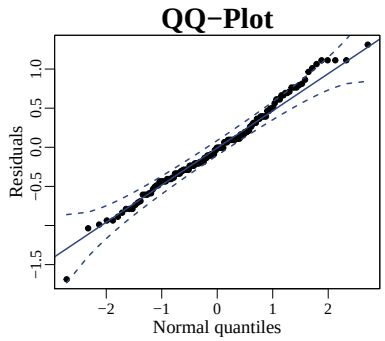


Figure 9.2: QQ-plot to check for deviations from normality of the residuals.

To check whether groups' variances are approximately equal, the scale-location plot can be used (fig. 9.3). This plot shows the relative differences from the mean at each group. As with all assumptions, slight deviations such as that observed in the example can be ignored. A steep increasing or decreasing trend (as judged by the blue smoothing ling) indicates violation of equal variance.

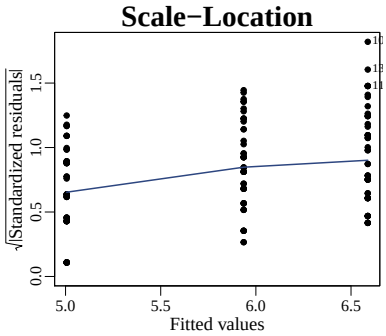


Figure 9.3: Diagnostic plot to check for heteroscedasticity of the residuals.

Tests for heterogeneity of variance are also available, albeit with the same inherent problems as tests for normality. Hence, a plot of the residuals of each category next to each other, such as in figure 9.3 is more informative.

Since ANOVA is not robust to outliers, the presence of any outliers or strongly influential observations should be assessed. Most commonly, the Cook's distance (a measure of outlyingness) is plotted. Figure 9.4 shows this diagnostic plot for the ANOVA performed on the `iris` dataset. Any observations beyond 0.5 Cook's distance are noteworthy and any observations beyond 1 Cook's distance are considered statistical outliers.

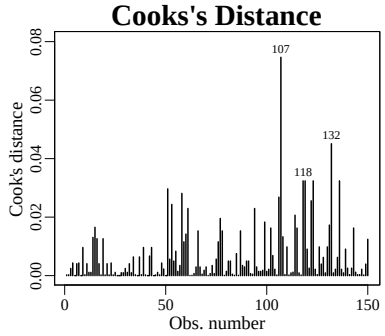


Figure 9.4: Diagnostic plot to check for potential outliers. The higher the value, the more 'outlying' the observation is.

In case there are statistical outliers, one should consider the cause of outlyingness and consult the labjournal if available. An outlier caused by an accidental error in experimentation should be removed from the dataset. Outliers of unknown origin should never be removed from the dataset. In the example, all Cook's distances are below 0.5, so there are no outliers.

9.1.4 *Confirmation*

The ANOVA can also be used to perform hypothesis testing. Most statistical software automatically perform an *omnibus* test: Is there any difference among means? In the `iris` example, the null-hypothesis is that all species' means are equal:

- $H_0: \quad \mu_1 = \mu_2 = \mu_3$

R reports the omnibus test in an ANOVA summary (table 9.1).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species	2	63.21	31.61	119.26	1.7×10^{-31}
Residuals	147	38.96	0.27		

Table 9.1: ANOVA table for comparing the species from figure 9.1. The omnibus test is significant at $\alpha = 0.05$ (see Pr(>F)).

The p-value in table 9.1 demonstrates that not all species' means are equal. Individual comparisons (A vs B, A vs C and B vs C) require a *post hoc* analysis.[7] This is an analysis that occurs after the ANOVA.

[7] Simply using separate t-tests is wrong without multiple testing correction (covered in Elements of Biostatistics).

9.1.5 Post Hoc: Tukey's HSD

Tukey's honest significant difference (HSD) is a *post hoc* analysis.[8] It is used when all individual comparisons are of interest. Continuing from the previous example, the ANOVA showed there was a significant difference in means, but not which species differed from which. Tukey's HSD can tell which individual differences are significant.

[8] Post hoc means after the (initial) analysis.

- **Performing more tests is *not* better. Only test for differences which are of primary interest to the research question.**

	diff	lwr	upr	p adj
versicolor-setosa	0.93	0.69	1.17	3.39×10^{-14}
virginica-setosa	1.58	1.34	1.83	3.00×10^{-15}
virginica-versicolor	0.65	0.41	0.90	8.29×10^{-9}

Table 9.2: Differences between groups (diff) with lower (lwr) and upper (upr) bounds of 95% confidence intervals and (Bonferroni adjusted) p-values (p adj).

The results of Tukey's HSD on the example ANOVA are summarized in table 9.2. The difference, range for the confidence interval and adjusted p-value can be seen for each comparison.[9]

[9] P-values adjusted for multiple testing (covered in Elements of Biostatistics).

The confidence intervals can also be plotted (fig. 9.5). In the `iris` example, all three differences do not include zero in the confidence interval and are thus significant.

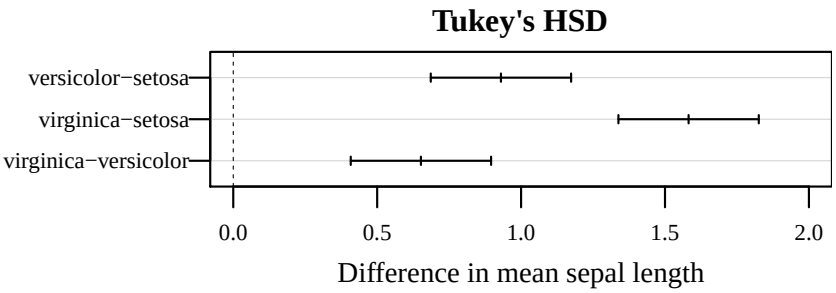


Figure 9.5: 95% confidence intervals for difference between groups. In this example, none of the intervals contain 0 and thus all differences are significant.

9.1.6 *Alternatives to ANOVA*

Non-Normal Alternative (*)
If observations come from a known, but non-normal distribution in the exponential family, a generalized linear model (GLM) can be used instead of ANOVA. This method is covered in depth in Elements of Biostatistics.

Non-Parametric Alternative
If observations come from an unknown, non-normal distribution, a non-parametric model can be used instead. Non-parametric models are beyond the scope of this book, but there is a simple non-parametric alternative to the omnibus test: The Kruskal-Wallis test is an extension of the Wilcoxon test (section ***7.1.3***) to more than 2 groups.

In R, the function `kruskal.test()` is available. This function's input is comparable to that of ANOVA in R (`aov()`). Also see the helpfile by running: `?kruskal.test`.

Heterogeneity of Variance (*)
In case variance is not equal in all groups, the response variable (y) might be on a different scale (e.g. logarithmic, exponential, quadratic). If this is the case, a transformation can often yield homogeneous variance. Transformation is covered in Elements of Biostatistics.

Dependent Data (*)
The direct extension of ANOVA to deal with paired or time-dependent data is repeated measures ANOVA. However, this method is favoured only by those who are unfamiliar with linear regression.

In linear regression, there is a superior method to deal with dependency: **mixed models**, covered in Advanced Biostatistics.

Unbalanced Data (*)
In case of unbalanced data, a linear model can be used instead of ANOVA (see paragraph ***8.2***). Rather than estimating and intercept and a slope for a single line, individual lines without a slope (9.13) are estimated. Each lines' height corresponds to the mean of a category.

To perform individual comparisons, one can use the R function `relevel()` to change the order of the categories of the factor. The first category is the intercept and the rest is compared to this category. An example of this procedure is included in paragraph ***8.3***.

9.2 Simple Linear Regression

Simple linear regression is a model for the intercept and slope of a line (fig. 9.6). The intercept is where the line crosses the y-axis (i.e. $x = 0$) and the slope is its steepness. Whereas ANOVA is used to compare a continuous response variable (e.g. length) to a discrete explanatory variable (e.g. species), simple linear regression is used to assess the relationship between two continuous variables, like length and width.

In mathematical notation, a simple linear model looks like this:

$$\underset{\text{response variable}}{y_i} = \underset{\text{intercept}}{\beta_0} + \underset{\text{slope}}{\beta_1} \cdot \underset{\text{explanatory variable}}{x} + \underset{\text{residual}}{\epsilon_i} \tag{9.4}$$

Read equation 9.4 as follows: The circumference (y) of tree number i is equal to the intercept plus the slope times its age (x) plus its particular deviation from the regression line (ϵ_i) (fig. 9.7). The slope and intercept are the same for every observation (systematic part) and hence have no index (i).

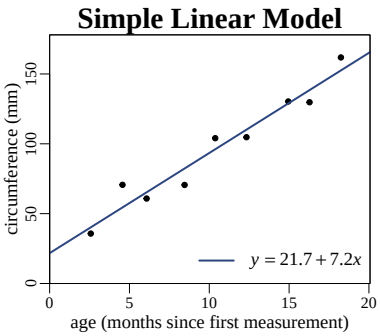


Figure 9.6: Linear relationship between the age of 12 trees and their circumference. The blue line depicts a linear model fitted to the data.

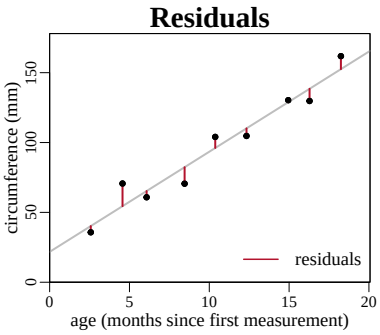


Figure 9.7: The residuals of a linear model are the vertical distances to the regression line (i.e. the random deviations ϵ_i).

9.2.1 Assumptions

Simple linear regression implicitly makes the following assumptions:

1. **Linear relationship** between the response variable and explanatory variable;
2. **Normality of the residuals**;
3. **Homogeneity of variance** (i.e. equal variance of the residuals along the regression line);
4. **Independence of observations**;
5. The explanatory variable **X is fixed**, or measured without error.

The last assumption is not as intuitive as the others and is covered in greater depth in Elements of Biostatistics. For now, remember that the explanatory variable is generally set by the scientist (e.g. dose and type of treatment) or measured without error (e.g. gender, age).

9.2.2 Estimation

The assumptions of linear regression are very similar to that of ANOVA because the estimation is also by least squares, minimizing the squared residuals ($\sum \epsilon_i^2$). As with ANOVA, this implies simple linear regression is not robust to outliers. Refer to ***8.1*** for a justification of least squares estimation.

- **Least squares is covered in depth in Elements of Biostatistics.**

9.2.3 *Diagnostics*

Since a simple linear model makes more assumptions than ANOVA, the diagnostics are also more elaborate. The four standard diagnostic plots reported by R assess violations of linearity, normality, heteroscedasticity and can reveal potential outliers.

Figure 9.8 is the first of four diagnostic plots reported in R. It is used to check for deviations from a linear relationship. In case of a perfect linear relationship, the blue smoothing line would be a straight line through $y = 0$. Here, there is no strong deviation. If the true relationship were quadratic for example, a clear parabolic shape should be visible. This can be remedied with a transformation, covered in Elements of Biostatistics.

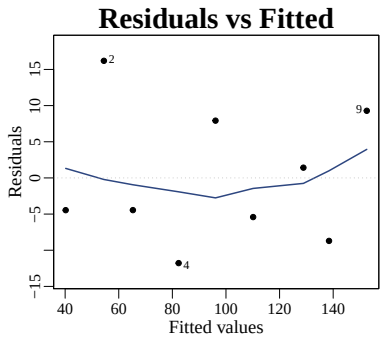


Figure 9.8: Diagnostic plot to check for deviations from a linear relationship.

A QQ-plot (section ***5.2.1***) can be used to assess whether the assumption of normality is reasonable (fig. 9.9). Although there are minor deviations from the blue line representing a theoretical normal distribution, none are beyond the confidence intervals.

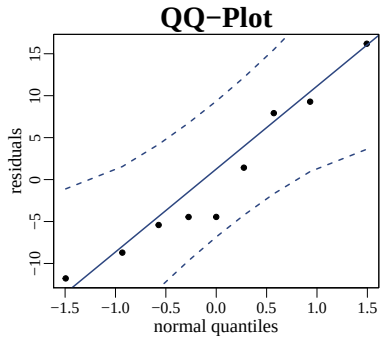


Figure 9.9: QQ-plot to check for deviations from normality of the residuals.

Heteroscedasticity, or approximately equal residual variance along the regression line, can be checked using the scale-location plot (fig. 9.10). This plot shows the relative differences from the regression line, as x increases. As with all assumptions, slight deviations such as that observed in the example can be ignored. The blue smooting line is approximately flat, neither increasing nor decreasing. Hence the assumption of homoscedasticity is reasonable.

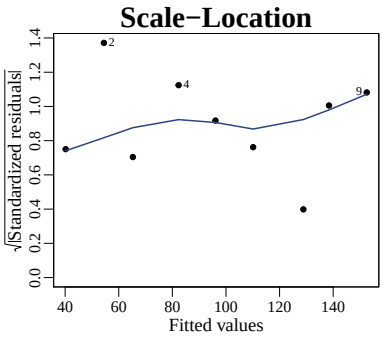


Figure 9.10: Diagnostic plot to check for heteroscedasticity of the residuals.

Simple linear regression is not robust to outliers, so the presence of potential outliers or strongly influential observations should be assessed. The Cook's distance (a measure of outlyingness) is the shown in the last of the default diagnostic plots in R (fig. 9.11). Any observations beyond 0.5 Cook's distance (the first boundary) are noteworthy and any observations beyond 1 Cook's distance (the second boundary) are considered statistical outliers. Although there are no outliers in this example, observation number 2 is beyond the first dashed boundary, meaning it is between 0.5 and 1.0 Cook's distance. However, since no further information is given about observation 2 and it is not beyond 1 Cook's distance, no further action needs to be taken.

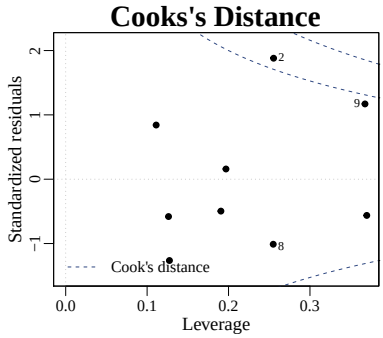


Figure 9.11: Diagnostic plot to check for potential outliers.

The rules for dealing with outliers are the same as those for ANOVA: Always consider the cause of outlyingness and consult the labjournal if available. Only outliers caused by an accidental error in experimentation should be removed from the dataset. Outliers of unknown origin should never be removed from the dataset.

- **Examples of serious violations, their potential causes and remedies are covered in Elements of Biostatistics.**

9.2.4 Confirmation

Provided its assumptions are met, a simple linear model can be used to perform two hypothesis tests:

1. $H_0: \quad \beta_0 = 0$
2. $H_0: \quad \beta_1 = 0$

These hypotheses correspond to the significance of the estimates for the intercept and the slope. Interpret these as follows:

1. If the intercept is insignificant, then the line might as well pass through the origin $(0,0)$ (fig. 9.12);
2. If the slope is insignificant, then there is no significant effect of the explanatory variable (x) on the response variable (y). In other words, y neither increases nor decreases significantly along the values of x (fig. 9.13).

Significance of the slope is often reported in scientific papers, because it is a statement about the significance of the relationship that is being modelled. In some cases, the hypothesis about the intercept is also of interest, but generally it can be ignored.

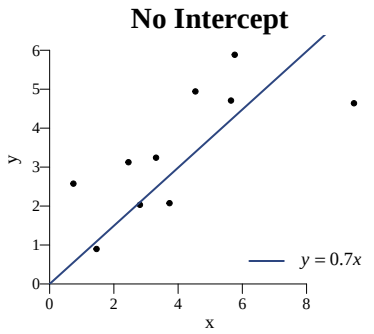


Figure 9.12: Simple linear model without an intercept (i.e. the line goes through the origin).

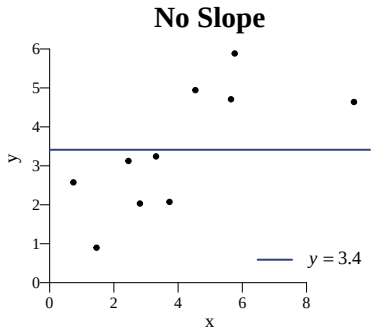


Figure 9.13: Simple linear model without a slope (i.e. the line is flat).

9.2.5 Prediction (*)

Linear models can also be used for prediction. Continuing from the previous example, the research question might be:

> *What circumference are trees expected to have after 12 months?* (prediction)

It is important to realize that—although related—this question is a fundamentally different from:

> *How large is the effect of age on the circumference of trees?* (estimation)

> *Is there a significant effect of age on the cirumference of trees?* (confirmation)

If the sole purpose of an experiment is prediction, then there is no interest in the true value of a parameter, the significance, or the strength of association between variables. Only one thing matters: prediction accuracy. Hence, the data are divided into a train and test set. The train set is used to fit the model, and the fitted model is used to predcit the outcome of the test set. An estimate of the out-of-sample prediction error is then obtained by comparing the predicted outcome of the test set, to its actual outcome variable.

9.2.6 *Alternatives*

A mathematician might argue that a linear model restricts us to a narrow class of functions, but as you will find out throughout this book series, linear models can actually capture many kinds of relationships by means of transformation or generalization. In fact, the number of possible adaptations of linear models to different problems is so large that linear regression is the main focus of Elements of Biostatistics and Advanced Biostatistics. Some examples are discussed below.

Non-Normal Response Variable

If observations come from a known, but non-normal distribution in the exponential family of distributions, a generalized linear model (GLM) can be used.[10]

If observations come from a known, non-normal distribution *not* in the exponential family, a generalized additive model (GAM) can be used instead.[11] The number of possible distributions for this type of model are numerous, but keep in mind that the distribution still requires a theoretical justification.

[10] Covered in Elements of Biostatistics.

[11] Covered in Advanced Statistics.

Heteroscedasticity

In case variance is not approximately equal along the regression line, the response variable (y) might be on a different scale (e.g. logarithmic, exponential, quadratic). If this is the case, a transformation can often resolve the problem.[12]

[12] Covered in Elements of Biostatistics.

Dependent Data, or x is Not Fixed

In case of repeated measures, longitudinal data, spatially correlated data, nested study designs or otherwise dependent data, *random* effects can be incorporated into the linear model. A linear model that contains both fixed and random effects is called a mixed model.[13]

[13] Covered in Advanced Statistics.

More Than One Explanatory Variable

Multiple Linear Regression is the extension of the simple linear model to any number of explanatory variables.[14]

[14] Covered in Elements of Biostatistics.

Numerical & Categorical Explanatory Variables

This is a special case of multiple linear regression, sometimes referred to as analysis of covariance (ANCOVA).[15]

[15] Covered in Elements of Biostatistics.

9.3 Examples in R

Although the actual estimation is beyond the scope of this introductory book, ANOVA and simple linear regression are quite easy to perform in R.

- **If you receive an error when running `library(...)` , install the missing package by running: `install.packages(...)`.**

9.3.1 ANOVA

An ANOVA model can be fitted with the function `aov(y ~ x)`, where `y` is the response variable and `x` the explanatory variable. If these are in a dataframe, the `data` argument can also be used as shown below:

Estimation

```
ANOVA <- aov(Sepal.Length ~ Species, data = iris) # Fit an ANOVA
coef(ANOVA)    # Report the coefficients
confint(ANOVA) # Report 95% CI for the coefficients
```

The coefficients must be added to the intercept to obtain the estimated means per species. The intercept is mean of the 'invisible' species (here: *I. setosa*).

These estimates are only meaningful if the assumptions of the ANOVA hold. Hence, diagnostics must be performed:

Diagnostics

```
library('car')           # Required for qqPlot()
qqPlot(residuals(ANOVA)) # QQ-plot with 95% CI
plot(ANOVA, which = 3)   # Scale-location plot
plot(ANOVA, which = 4)   # Cook's Distance
```

The default `plot()` function returns four diagnostic plots, not all of which have been covered. Below are all available diagnostic plots for those interested:

```
plot(ANOVA)            # Four default diagnostic plots (1, 2, 3, 5)
plot(ANOVA, which = 1) # Residuals vs Fitted (see simple linear model)
plot(ANOVA, which = 2) # QQ-plot (no confidence intervals)
plot(ANOVA, which = 5) # Constant leverage (heterogeneity of variance)
plot(ANOVA, which = 6) # Cook's distance vs leverage (detect outliers)
```

Provided the assumptions hold, the esimates and (optionally) significance can be reported:

Confirmation

```
summary(ANOVA) # Perform and report the omnibus test
```

Confirmation: *Post-Hoc* (Individual Comparisons)

```
post.hoc <- TukeyHSD(ANOVA) # perform Tukey's HSD test
post.hoc       # Print the result (very small p-values rounded to 0)
plot(post.hoc) # Plot the confidence intervals for difference
```

Regardless of the objective, it is always wise to attempt to visualize the result; is this in line with what was expected? Are the (significant) differences of relevant size?

Visualization

The base `plot()` function in R will produce a boxplot that you can add to your report:

```
plot(Sepal.Length ~ Species, data = iris) # Boxplot
```

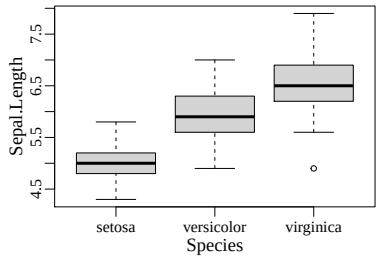


Figure 9.14: Boxplot of species.

Using a popular R package for visualization (`ggplot2`) and an extension (`ggpubr`), the significance of the comparisons can be added with just a few lines of code:

```
library(ggplot2)
library(ggpubr)
compare <- list(c("setosa", "versicolor"),
                c("setosa", "virginica"),
                c("versicolor", "virginica"))
ggboxplot(iris, x = "Species", y = "Sepal.Length") +
  stat_compare_means(comparisons = compare, method = 't.test',
                     label = "p.signif")
```

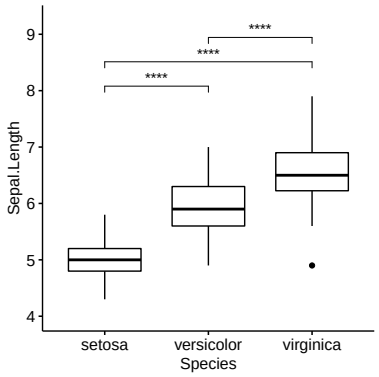


Figure 9.15: Individual comparisons visualized through `ggplot2`.

Scientific journals often include the notations n.s. (non-significant), * ($p < 0.05$), ** ($p < 0.01$) and *** ($p < 0.001$). This is achieved by using the argument `label = "p.signif"`.

A cautionary note to the reader: Be ware that 'very significant' does not mean 'very relevant'. With large enough sample size, anything will become significant. Always take the effect size into consideration.

9.3.2 *Simple Linear Regression*

A simple linear model can be fitted with the function `lm(y ~ x)`, where `y` is the response variable and `x` the explanatory variable. If these are in a dataframe, the `data` argument can also be used as shown below. For this example, only *I. virginica* will be used:

Estimation

```
subset <- iris[iris$Species == "virginica", ]        # Create subset
LM <- lm(Sepal.Length ~ Petal.Length, data = subset) # Fit the model
coef(LM)    # Report the coefficients
confint(LM) # Report 95% CI for the coefficients
```

These estimates are only meaningful if the assumptions of the simple linear model hold. Hence, diagnostics must be performed:

Diagnostics

```
plot(LM, which = 1)   # Residuals vs Fitted (assess linearity)
library('car')        # Required for qqPlot()
qqPlot(residuals(LM)) # QQ-plot with 95% CI
plot(LM, which = 3)   # Scale-location plot
plot(LM, which = 5)   # Cook's Distance
```

The default `plot()` function returns four diagnostic plots, not all of which have been covered. Below are all available plots for those interested:

```
plot(LM)            # Four default diagnostic plots (1, 2, 3, 5)
plot(LM, which = 2) # QQ-plot (no confidence intervals)
plot(LM, which = 4) # Cooks's distance (does not include leverage)
plot(LM, which = 6) # Cook's distance vs leverage (alternative to 5)
```

Provided the assumptions hold, the esimates and (optionally) significance can be reported:

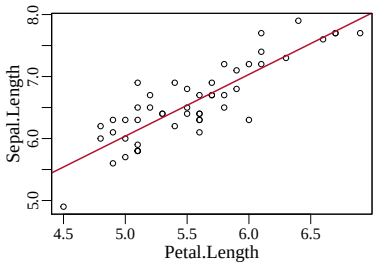
Confirmation

```
summary(LM) # Report the coefficients, std. errors and p-values
```

For now, focus on the `Coefficients` tab of the summary. The rest of the ouput is explained in Elements of Biostatistics.

Visualization

```
plot(Sepal.Length ~ Petal.Length, data = subset)
abline(coef(LM), col = 'red')
```



More elaborate plotting routines are covered in Advanced Biostatistics.

A Example Analyses

(. . .) to be included.

B Further Reading

This is the first book in a series of three books:

- **Introduction to Biostatistics**
- **Elements of Biostatistics**
- **Advanced Biostatistics**

If you want to learn more about *using* statistics, I would simply advice to read the next book in the series. However, the fundamentals will not be further elaborated on. For those who seek further understanding of **probability theory**, I recommend:

- Rice (2007): Mathematical Statistics and Data Analysis

The next book in this series focusses on linear models, but does so in a pragmatic way, respecting the usually short duration of courses on statistics in biology. For a much more comprehensive introduction to **linear models**, I highly recommend:

- Fox (2008): Applied Regression Analysis and Generalized Linear Models

To understand the expressions in Fox (2008), you may want to learn the basics of **linear algebra** first. There is a free video series:

- Linear Algebra by Gilbert Strang (MIT OpenCourseWare, YouTube).

If you have an interest in **predictive modelling**, the essentials will be covered in the last book in this series. However, if you have only read this book prior and you want to move on to prediction, I recommend:

- James *et al.* (2013): An Introduction to Statistical Learning

However, if you already feel familiar with the fundamentals of statistics (including the mathematical notation), I recommend reading:

- Hastie *et al.* (2008): The Elements of Statistical Learning

Acknowledgements

This book would not have been possible without **Dr. H.G.J. van Mil**: A friend, a colleague and a researcher, whose invaluable support and vision to improve the quality of teaching statistics in the life sciences is ultimately motivated me to start writing.

In addition, I would like to thank:

- **drs. B. van Ooijen, MMO** for suggestions on general structure;
- **L.H. Tran, BSc** for proof-reading early versions;
- **V. de Bakker, MSc** for proof-reading early versions;
- **L.M. van Oosterhoud, BSc** for proof-reading early versions.