Statistics applied to Biology

A very short introduction

Oreste Affatato*

April 3, 2023

An Awesome Publisher

* A not so awesome author

Disclaimer
This book contains the material for the lab of statistics applied to biology, for master students in Biomedicine at Uppsala University. The text is purely intended as an introduction to more sophisticated and rigorous methods and should not be considered a specialized textbook on the subject.

Colophon
This document was typeset with the help of KOMA-Script and LaTeX using the kaobook class.
The source code of this book is available at:
`https://github.com/fmarotta/kaobook`
(You are welcome to contribute!)

Publisher
First printed in the past by An Awesome Publisher

I know not what tomorrow will bring...

– Fernando Pessoa

Preface

This is the handout for the computer lab on statistics applied to biology. It is a very simple introduction to this field with some R programming. The overall aim is to offer you a valuable approach to Statistics that might help you in your future studies. Moreover, my intention was also to guide you towards more professional and better texts on the subject, and hopefully to show you the beauty of Statistics.

I am particularly attached to this handout, because it has been based on my own mistakes and misunderstandings of Statistics (the ones I could recognize, of course). Anyway, if I was able to see further it has been only because I was sitting on the shoulders of the giants. This handout is inspired by countless readings (books, articles, [toilet] papers...) and discussions with colleagues and friends. I am particularly grateful to the mentors I never met, in particular Danielle Navarro, David Spiegelhalter, Andrew Gelman and Richard McElreath. Their understanding of Statistics and their pedagogical skills are just off the charts. I am even more grateful to the mentors that I've met. In particular, I would like to thank Lars Lindhagen, my first Statistics teacher here in Uppsala and a constant inspiration. Moreover, I would like to thank dr. Schmidt and dr. Sokolov for our endless talks in pursuit for an endless self-improvement.

This handout has also been written under the influence of the many biases of the author. I hope that the kind reader would forgive me for that: it was all intentional. Anyway, I will be grateful to our students that will indicate to me mistakes in the text. I will also greatly appreciate any suggestion to improve the lab.

Oreste

Contents

Introduction 1

The White Rabbit put on his spectacles. "Where shall I begin, please your Majesty?" he asked. "Begin at the beginning," the King said gravely, "and go on till you come to the end: then stop."
- Lewis Carrol, Alice in Wonderland

Statistics may be defined in many ways. For us, it will suffice the following definition: statistics is the art of learning from data.
Why bother with that? As scientists, we ask many questions on how the universe works, what are its underlying mechanisms, from the motion of an election in an atom to the galaxies falling apart. To verify our claims and to check whether our ideas about the universe are indeed correct, we then have to run experiments. From these experiments we gather data. The data usually[1] are characterized by a certain level of noise. We would like to detect a particular signal, but data are rather dispersed. That's why we need statistics, to understand the data, to separate the signal from the noise, to divide the real phenomenon from the disturbances of randomness.

1.1 Probability and statistics

Probability and Statistics are two fields strictly related. They both are fundamentally mathematical disciplines and they both deal with randomness.
In Probability, we start from a theoretical data generating phenomenon and then we try to predict some of its features. For example, we imagine to flip a coin, with head or tail. We already establish that there is a 50% chance to have one of the two and then we may ask ourselves, flipping the coin 10 times in a row, how it would be likely to get 7 times tail. More generally speaking, we know already the distribution generating the data and its parameters, and then we would like to predict the likelihood of some kind of outcome.
Statistics works the other way around. We already have the outcome, basically the results of our experiment. Then we make some assumptions on what the generating distribution might have been and then we try to estimate the parameters of the distribution. Following the previous example, imagine we flipped a coin 10 times and we've got 7 times tail. We may ask ourselves what is the generating phenomenon, what was the probability to get tail in one flip, making some due assumptions.
When it comes to Statistics, in particular, the methods of inference are indeed solid, but not unambiguous. Statistics relies heavily on mathematical results to work, however the same problem can be tackled with very different approaches. Anyway, whatever is the approach that you would like to use, it is important that you have good argumentation for that and that you fully understand what the meaning of the tools the you use.

[1] Here I was uncannily generous: they are ALWAYS noisy!

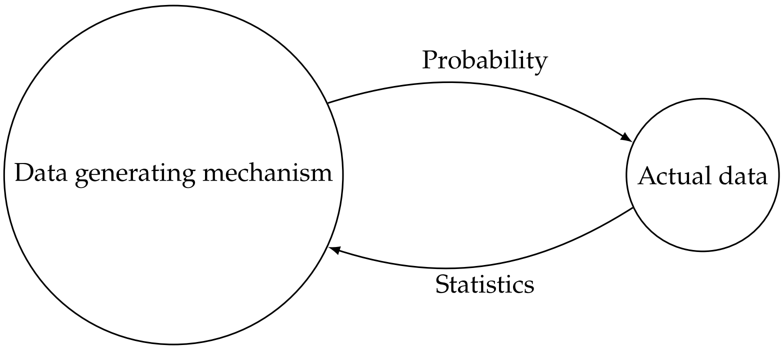


Figure 1.1: Relationship between Probability and Statistics.

1.2 Probability and distributions

In this section we are going to briefly recall some major concepts from probability theory. Some of these will be discussed again in Chapter 3. In particular, here we will give the formal mathematical definition of probability and leave the discussion on its interpretation for later.

So, probability is a way to measure uncertainty, randomness. It doesn't make sense to talk about these concepts *per se*, therefore probability refers always to an event (or a set of events). And in particular all the events are always to be understood in the bigger framework of the space of all possible outcomes. Let's start from here.

Considering an experiment in which we cannot know for sure what would be the actual outcome, we call Ω the set containing all possible outcomes. It is also called the **space of events**. An event E is any subset of Ω, so in mathematical terms $E \in \Omega$. With the concept of probability of an event we try to capture the uncertainty around the fact that, once done the experiment, that event will occur. The probability of an event is defined according to a set of axioms proposed by the Russian mathematician Andrej Nikolaevič Kolmogorov. These axioms are the following

- $P(E) \in \mathbb{R}, P(E) \geqslant 0$;
- $P(\Omega) = 1$;
- $P(\bigcup E_i) = \sum_{i=1}^{N} P(E_i)$.

Where $\mathbb{R}$ is the set of real numbers, Ω is the space of events and E_i is a general event[2].

From these axioms, one can derive a lot of properties. We are just going to mention some that might be useful for future discussion.

- $0 \leq P(E) \leq 1$;
- if two events A and B are *mutually exclusive*, $P(A \cup B) = P(A) + P(B)$;
- if two events A and B are *stochastically independent*, $P(A \cap B) = P(A)P(B)$.

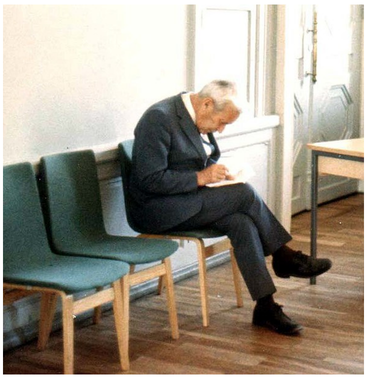


Figure 1.2: Andrey Nikolaevich Kolmogorov (25 April 1903 – 20 October 1987). Kolmogorov probably captured in a unlikely state of uncertainty, maybe before a conference. Source: Wikipedia.

These are some very basic properties on probability. Another important concept is the **random variable**. A random variable is anything that can take some particular values each one with some given probability. For example, the outcome of rolling a die is a random variable. The possible outcomes are the numbers from 1 to 6 each one with the same probability,

[2] Remember that these are all sets, therefore the usual properties and operations defined for sets hold true also in this case.

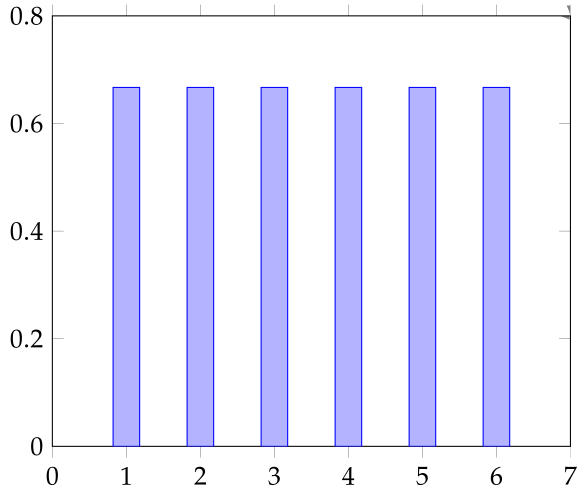


Figure 1.3: Probability distribution of the rolling die experiment.

i.e. 1/6. A **probability distribution** is a function that shows all the possible values of a random variable with the given probabilities. In the case of rolling the dice, the probability distribution is a uniform function, since all the outcomes have the same probability. In more mathematical terms, X is the random variable that represents our experiment of rolling the dice. In particular, X can take any value in the set $\{1, 2, 3, 4, 5, 6\}$. They all have the same probability. So $P(X = 1) = 1/6$, $P(X = 2) = 1/6$ and so on.

The probability distributions can be divided into two main categories: **discrete** and **continuous**. The discrete distributions, like in the case of the rolling die experiment, derive from random variables that can take only a finite number of possible outcomes[3]. The continuous distributions derive from random variables that could take any possible value among the real numbers. For example, the distribution of the height can be an example.

One last general thing about the distribution. For the discrete distributions, the sum of all the probabilities is one. In the case of the continuous distribution, the total are under the curve is one[4].

There are many important theoretical distribution. Among those, we would like to mention just the **Gaussian** or **normal distribution**. It is a very particular type of continuous distribution and it is very important for many of its properties[5].

What are the fantastic features of this curve?

- it is *completely* described by the two parameters, the mean μ and the variance σ^2;
- it is symmetric;
- increasing the mean would move the center to the right, while decreasing it would move it to the left;
- a bigger variance would spread more the bell shape.

Many of the phenomena that we encounter in nature follow the normal distribution and this is its major power. Because of the central limit theorem, many natural occurring random variables, not matter what was the actual stochastic generating process, they will always follow a normal distribution, provided that the sample size is big enough.

[3] Just kidding. They can be infinite, but in such a way that can be numbered using natural numbers.

[4] You can imagine the reason. The probability that anything can happen is one.

[5] And also because it can be proven that other distributions, other certain conditions, as the sample size increases they become normally distributed.

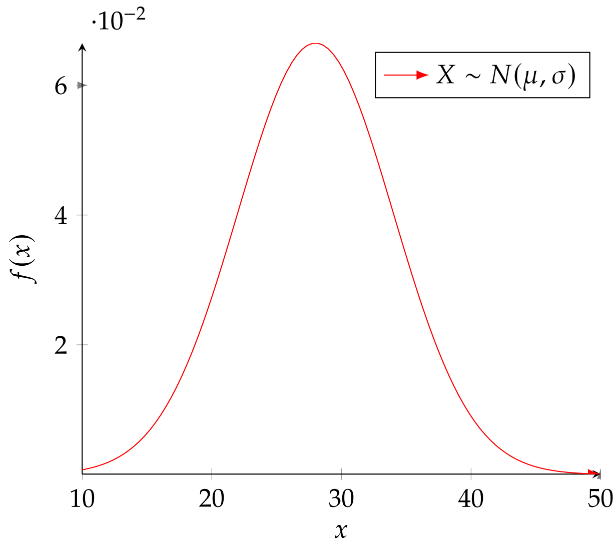


Figure 1.4: Normal distribution centred in μ and with variance σ^2.

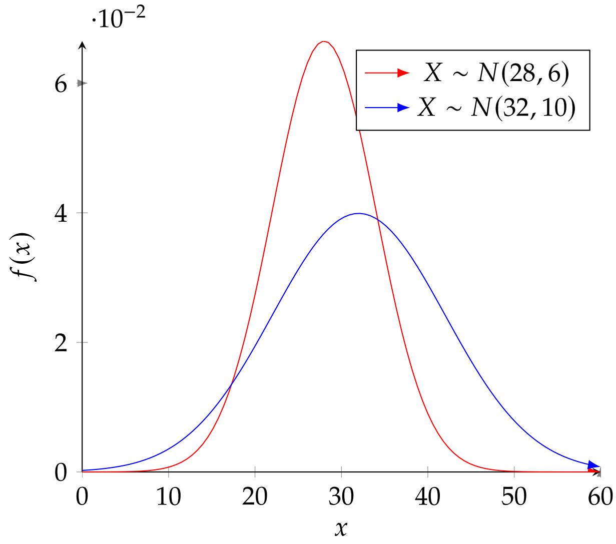


Figure 1.5: Comparison between two Gaussian distributions.

1.3 Population, sample *et al.*

Before we get started with Statistics, I would like to remind you an important distinction: the one between **population** and **sample**[6].
The population is the target, all the possible representations of the phenomenon that you want to study. Let's imagine you want to study the blood concentration of a particular protein. What is your population? Here is not easy to establish, as the research question was too broad. It could be all humanity, just the adults in Sweden *et cetera*. Let's try to be more specific, and just focus on the overall Swedish population. Therefore, our target population comprises all the Swedes. Now that we know our target, can we draw blood to all the Swedes and measure the blood concentration of that protein? Of course not. In general, from many reasons, it is impossible to perform an experiment on the entire target population. Therefore, the only thing we can do is to draw a sample from the target population. The sample is a subset of the population that is hopefully representative of the entire target.
So we use what is practically available, the sample, to draw conclusion on what is not available, the population. That is the process of statistical

[6] Following an old tradition in statistics, I'll use Greek letters for the population parameters and Latin letters for the sample statistics.

inference: using the sample to understand the population. This distinction is very important.

One last thing to remind is the difference between the **systematic** and the **random error**. In the process of sampling, in different stages, we can introduce bias in the process. Imagine you are measuring a length, but the ruler has some defects and it overestimate the lengths. It doesn't matter how many time you measure the same length, all the measures are going to be overestimated. This is a systematic error, as the measures are pushed in one specific directions, due to the methods/tools you are using. The only way to deal with this source of error, is to change methods or tools.

It may also happen that, repeating measures you may commit some mistakes. Sometimes you may overestimate, sometimes underestimate your measure. In this case, the error happens randomly. Repeating the measures a lot of times, you improve your precision as the overestimations are going to balance the underestimations in the long run. This is the random error.

With the statistic tool we are going to study we will be able to deal with the random error, the one due to chance. The systematic error is due to the methodology, so it is not possible to reveal it using purely statistical methods. Keeping these things in mind we'll later look at some ways on how to describe the sample.

Before proceeding further, some more concepts about data. These are observations of certain **variables** of interest, like blood pressure, intracranial volume *et cetera*. Variables can be of two main types: **categorical** or **numerical**.

- *Categorical variables*. This type comprises nominal and ordinal variables. A **nominal** variable is characterized by mutually exclusive categories with no specific order (e.g. sex: male or female), while **ordinal** variables have intrinsic ordering (e.g. symptom severity: mild, moderate, severe);
- *Numerical variables*. This type comprises discrete and ordinal continuous. A **discrete** variable is characterized by integer values (e.g. days of sick leave), while **continuous** variables have any real value (e.g. height, measured in certain units).

Descriptive Statistics 2

Est pittoresque tout ce qui est accidenté.
- Roland Barthes, Mythologies.

In this chapter we are going to deal with very basic descriptive concepts. We will discuss some of the most important *statistics** useful to address this issue. In particular, we will emphasize the strengths and the limitations of such tools.

To do exercise with descriptive statistics we will use a synthetic dataset. The story goes like this. An important Swedish company, the Yolo AB, decided to do something to promote the health of its own employees. As a start, they decide to have a medical screening of all the people working there. Before downloading this precious dataset, first you might want to set a working directory

```
setwd("Insert here the path to your working directory")
```

Now we can download the dataset from my GitHub account

```
Data = read.csv("https://raw.githubusercontent.com/
 OresteAffatato/Statistics-lab-UU/main/YoloAB.csv")
```

Probably you would get one column more, the one with the row numbers. To delete that, run the following code

```
Data = Data[, -1]
```

We are asked by the CEO of the Yolo AB to run some descriptive analyses of the health situation of the employees. As you can already see having a quick look at the dataset, the physicians have measured several health-related parameters from a sample of 40 participants.

This table contains all the detailed information about each participants, but we can't really use this knowledge in this format. To make sense of all of this it is just not possible to only look at the table. We need to find some smart way to summarize the informative content of this dataset and in doing that we will loose some information. There is always a trade-off in statistics: we have to sacrifice the specificity of all the values of the table to create some summary information that could help us if there is any emerging pattern, any signal.

Detecting the signal is not the end of the story, of course. We would also expect some sort of variation (due to whatever reason, from the normal biological variation to measurement errors). Therefore our task is actually two-fold: assess the presence of a signal and also account for the variation, for the noise in the data.

* A *statistic* is anything that can be calculated from the data.

2.1 Looking for the signal

The mean and the median are two widely used statistics that measure the tendency of numerical data to cluster around some sort of center. As we already said, data are expected to vary, but maybe they are generated around some preferred values. The mean and the median try to measure this in two different ways.

2.1.1 Mean

Let's assume we have a sample of N numerical observations, i.e. $x_1, x_2, ..., x_N$. The **arithmetic mean** is defined as

$$\bar{x} = \frac{x_1 + x_2 + ... + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

What is the interpretation of this? You can imagine the arithmetic mean as the centre of mass of your data. That's also the reason why the mean is very sensible to big numbers (i.e. outliers). Any big number (compared to the others) would attract the centre of mass towards itself, like putting some extra weight on a double-armed scale.
How can we calculate the arithmetic mean with R? Let's imagine we would like to calculate the mean age. We can use the code

```
mean(data$Age)
```

We could obtain the same result creating first the vector with all information about the age of the participants, and then taking the mean of this vector.

```
Age = c(data$Age)
mean(Age)
```

2.1.2 Median

There is no mathematical formula to calculate the **median**. It is just the central value (in an ordinal sense), the observation that divides the data set into two equal parts. Exactly half of the observations are going to be below the median and the other half above it.
To compute the median age (continuing the previous example) with R we can use the code

```
median(data$Age)
```

or, since we already created the vector

```
median(Age)
```

The median, as you can easily notice, is not affected by outliers. It doesn't care about the actual value of each observation. It just divides the observations into two equal subsets, independently from their numerical value. The mean, on the contrary, is usualy attracted by extreme values. It is important to keep in mind this difference as it is pretty clear when we deal with asymmetric/skewed data.

2.2 Taming the noise

So far so good. Mean and median are undoubtedly useful tools to summarize data and to assess the center of our distributions. The problem is that they don't tell anything about how sparse the data are. Let's consider a very stupid example. Imagine to measure the same variable (e.g. diastolic blood pressure) in two groups of people, in this example with three participants in each group.

Group 1: 78, 79, 80 mmHG
Group 2: 59, 79, 99 mmHG

In both cases, mean and median are equal to 79 mmHg. The difference is that in Group 1 the data seem to be more clustered around 79 mmHg than in Group 2[1].
Mean and median are not sufficient to have a complete picture of our data. We need than a way to measure the dispersion.

2.2.1 Variance and standard deviation

The mean is one of the most widely used measures of central tendency, for many practical and theoretical reasons[2]. Therefore, it makes sense to assess dispersion in our data set measuring the discrepancy between each data point and the mean. Let's consider a set of N data, $x_1, ..., x_N$, with mean $\bar{x}$. The discrepancy d_i for each data point x_i would measure how far is this particular observation from the mean, i.e. it is the difference between mean and x_i

$$d_i = \bar{x} - x_i$$

This for one data point. If we want to know the general discrepancy, the general distancing between each data point and the mean we could just take the mean of the discrepancies

$$\bar{d} = \frac{\sum_{i=1}^{N} d_i}{N} = \frac{\sum_{i=1}^{N} \bar{x} - x_i}{N}$$

Guess what? Because of the definition of mean, the mean discrepancy is always zero[3].

$$\bar{d} = 0$$

It is clear why. The mean is the center of mass of our data. It is defined to be on average at the same distance far away from all the x_i. Therefore, the positive contributions to the sum will be always cancelled out by the negative ones and on average we get zero as a result.
To avoid that, instead of calculating the mean of the discrepancies, we calculate the mean of the squared discrepancies. In this way all the contributions are going to be positive. We define this quantity **variance**.

$$Var(x) = \frac{\sum_{i=1}^{N} d_i^2}{N} = \frac{\sum_{i=1}^{N} (\bar{x} - x_i)^2}{N}$$

[1] I am actually worried about the health condition of the people in Group 2.

[2] We'll see why later on.

[3] It is an easy mathematical exercise, you should try. ;-)

The variance has some properties

- it is strictly positive;
- the closer the observations x_i are to the mean $\bar{x}$, the more the variance is close to zero;
- the larger the sample size N, the smaller the variance.

To calculate the variance in R we can use the code

```
var(Age)
```

The variance is a good indicator of dispersion, also according to our common sense. Smaller the discrepancies, smaller the dispersion (i.e. smaller the variance). And the more data we have, the better.

The main problem with the variance is that its units of measure do not match the ones of the mean. In fact, variance has the squared units of measures corresponding to the mean. To overcome this problem we can use another good estimator of dispersion, which is the **standard deviation**, defined as follows:

$$SD(x) = \sqrt{Var(x)}$$

To calculate the standard deviation in R we can use the code

```
sd(Age)
```

Pretty straightforward, right? The standard deviation has the same properties as the variance, but it has also the same units of measure as the mean[4].

A final note. $Var(x)$ is also written as $s^2(x)$ and for obvious reasons $SD(x)$ is also written as $s(x)$.

2.2.2 Interquartile range

Another measure of dispersion widely used is the **interquartile range**. It is the difference between the 25^{th} and the 75^{th} quantiles[5]. It is a measure of the interval that contains the central 50% of the data.

Using R we can calculate the extreme values of this interval

```
quantile(x = Age, probs = c(.25, .75))
```

And we can also directly calculate how wide it is.

```
IQR(x = Age)
```

2.3 Final remarks

Here I would just briefly discuss few points.

- *A description should be descriptive.* The statistics we discussed are very important, and you should use them as much as needed to get a better understanding of your dataset;
- *Right tool for the right context.* All the tools we discussed have pros and cons. It is very much important that you keep in mind when you should use one or the other. They can convey the right information only when used adequately.

[4] Variance and standard deviation have also the same problems of the mean. Since the are both defined as an arithmetic mean they are also sensible to extreme values.

[5] I think I haven't said what a quantile is. :-(
Anyway, the median is the 50^{th} quantile, i.e. the number that has 50% of data below and 50% of data above itself. In the same way, the 25^{th} quantile has 25% of data below and 75% above. And so on.

2.4 Exercises

1. Consider the other variables of the dataset. Calculate some descriptive statistics. For each case try to reflect which statistics is more helpful to describe your variable and try to argue why.
2. You might want to consider to plot some variables to have a graphic representation. You could use the command *hist()* to create histograms and/or *plot(density())* for density plot.
3. For some variables you might want to consider to use the command *boxplot()* to show the median, the first and the third quartile (as well as minimum and maximum values). You can use *boxplot(variable group variable)* to compare groups.
4. The Yolo AB is not a very nice company to work in, after all. To pay for this medical screening they asked for a small contribution from each employee. Each person has to contribute with 1% of the monthly salary to cover the medical costs. How much is each participant going to pay *on average*? What would work better to answer this, the mean or the median?

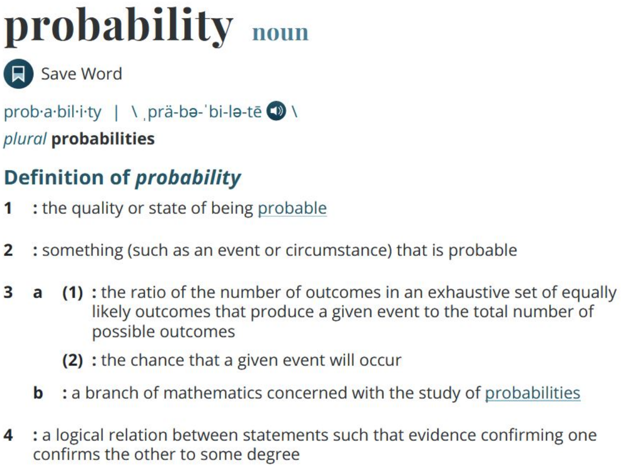
Interlude 3

O, be some other name!
What's in a name? That which we call a rose,
By any other name would smell as sweet;
- William Shakespeare, Romeo and Juliet

In order to understand what is about to come, i.e. the process of *statistical inference*, we need to have some concepts clear in mind. In particular, the concept of probability.

3.1 Frequentist or Bayesian?

What do we mean when we talk about probability?
We have an intuitive sense for it. We even may think to have a look at the Merriam-Webster[1]



probability noun

Save Word

prob·a·bil·i·ty | \ ˌprä-bə-ˈbi-lə-tē \
plural **probabilities**

Definition of *probability*

1 **:** the quality or state of being probable

2 **:** something (such as an event or circumstance) that is probable

3 **a** **(1)** **:** the ratio of the number of outcomes in an exhaustive set of equally likely outcomes that produce a given event to the total number of possible outcomes

(2) **:** the chance that a given event will occur

b **:** a branch of mathematics concerned with the study of probabilities

4 **:** a logical relation between statements such that evidence confirming one confirms the other to some degree

It still doesn't sound clear, does it? It is also a bit tricky when they write something like "chance" and "equally likely" in the definitions at point 3, as they define probability using the concept of probability.
Mathematics has an entire field on this, the theory of probability. We may have a look there to see if there is a better answer to our question.
In mathematics, probability is defined by the Kolmogorov axioms. Let's consider a set of events Ω and $E \in \Omega$, an event in Ω*. The probability of an event E is a real non-negative number

- $P(E) \in \mathbb{R}, P(E) \geqslant 0$

The second axiom tells us that the probability of Ω is 1[2]

* Put it simple, Ω is the set of all possible outcomes of an experiment and E is one of the outcomes. Consider the experiment of flipping a coin. In this case $\Omega = \{H, T\}$, as you can get either head (H) or tail (T). An event could be flipping the coin and getting head, i.e. $E = \{H\}$.

[1] On the website it is written "since 1828", therefore it must be reliable.

[2] This means that 1 is the probability that something is going to happen.

- $P(\Omega) = 1$

The third axiom tells us that the probability of the union of all disjoint events is equal to the sum of their own probabilities

- $P(\bigcup E_i) = \sum_{i=1}^{N} P(E_i)$

From this we can create the theory of probability, a coherent and sound mathematical area. Still we have the same problem. What does it mean? What is the probability? These axioms give us a quite abstract idea of it, how can we translate this to the real life?
Trying to answer to these questions, in the field of statistics have emerged two major schools of statistics: frequentism and Bayesianism[3].
Let's try to see how the frequentists define probability. Let's consider a practical example, the experiment of flipping a coin. What is the probability of getting head? If you flip the coin N times in same cases you are going to have heads (n_H) and some others tails (n_T). We define the **relative frequency** of success the number of times we get head (the success) divided by the number of trials.

$$f = \frac{n_H}{N}$$

The point is that we use the frequencies to approximate the probability. And this is actually a good way to go, because the more the trials we do the more the frequency is going to be an accurate predictor of a future experiment. Why is that? There is a powerful mathematical tool that proofs that this is always the case, given that N is sufficiently large. It is the *law of large numbers*. This theorem tells us that, as the number of trials becomes infinitely large, so the frequency is going to converge to a finite number, non-negative number. This number is the probability of success.
The Bayesians have a totally different approach. In this case, the probability is not a measure of frequency, but rather a measure of the degree of our belief that something is going to happen. Something based on past experience and prior knowledge of the phenomenon. It is rather a *subjective* approach.
It might seem that this is a rather abstract discussion, with no real impact on your research. This is very much wrong. This is the foundation of the two statistics schools and both have different methods that can provide answers to your research questions. And, most of all, the interpretation of your analysis is different, depending on the methods you are using. Part of the scientific misconduct in statistics derive from the fact the researchers use frequentist methods, but then interpret the results in a Bayesian way. This is a real problem, as this is just not what your analysis is telling you.

3.2 Sampling theory

In this last section we are going to discuss an important result concerning the sampling process. In particular, we are going to discover why the arithmetic mean is such a powerful tool in statistics.
In general, we know some things about the process generating the data, but we definitely don't know the details. In particular, we usually ignore the exact values of the parameters that generate the data we gather in

[3] You probably are not aware, but you belong to the first one.

our experiments. That's why we do statistics, to try to measure these parameters. However, for the moment, and for pedagogical reasons, let's imagine that you are some sort of divine entity and that you know everything. You might imagine to be Athena, the Greek goddess of wisdom. Up there, from Mount Olympus you observe and know everything.

Figure 3.1: Athena (some century before Christ - no clue how is she doing right now, probably still alive). Here is a picture of Athena clearly pissed. She knows everything, and in particular she knows what you do with p-values and stuff. Source: Wikipedia.

Some scientists would like to measure the average height of a certain adult population, living close to Mount Olympus. You know the truth, you know that the height of these people are normally distributed with mean $\mu = 175$ cm and standard deviation of $\sigma = 10$ cm. The true distribution looks like this

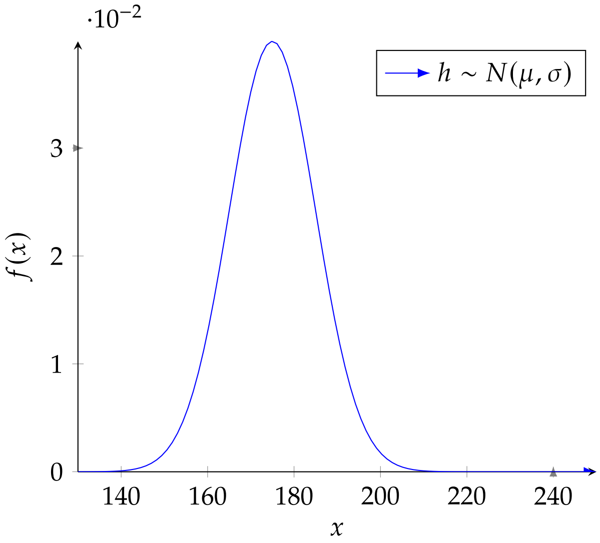


Figure 3.2: Distribution of the height in a population.

These scientists decide to run an experiment, the recruit 5 people at random, measure their height and calculate the mean and the standard deviation.

```
# Setting the generating function for the height distribution

Height = rnorm(10000, mean = 175, sd = 10)

# Simulating the first experiment. We extract randomly 5 people
    from the
# population

FirstSampling = sample(Height, size = 5)

# Calculating the mean

SampleMean = mean(FirstSampling)
SampleSD = sd(FirstSampling)
```

You get a sample mean $m = 175$ cm and a standard deviation $s = 5$ cm. As Athena you notice that the sample mean, the one calculated by the scientists, is very similar (even equal, in this case) to the population mean, the true mean. However, you want them to understand the bigger picture. Hidden here there is one of the most astonishing secrets of nature.
You appear in the dream of one of them and suggest to run again this experiment multiple time. The day after, they ask all the available

scientists to work on this, and they repeat the same experiment 100 times and keep the record of all the means.

```
# Setting the generating function for the height distribution

Height = rnorm(10000, mean = 175, sd = 10)

# Simulating the repeated experiments. We extract randomly 5 people from the
# population for 100 times

SampleMean = c()

SampleSize = 5

for (i in 0:100) {

  Sampling = sample(Height, size = SampleSize)
  SampleMean[i] = mean(Sampling)

}


hist(SampleMean, density=20, breaks=20, probability = TRUE,
     xlab="Sample Means", ylim = c(0, 0.19), xlim = c(160, 190), main="True distribution over histogram")
curve(dnorm(x, mean = 175, sd = 10),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
abline(v = 175, col="red", lwd=3, lty=2)

# Calculating the mean and standard deviation of means

MeanMean = mean(SampleMean)
SDMean = sd(SampleMean)
```

Plotting all the means from all the experiments over the true distributions, we should get something like this

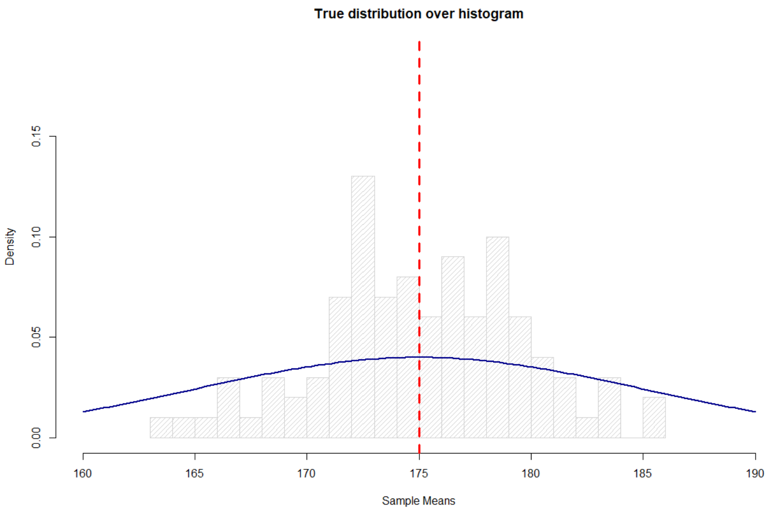


Figure 3.3: Pooled results from the many simulated experiments.

We can see that the means we've got from the different experiments also seem to have some sort of distribution, usually called *sampling distribution*. In particular, we notice that the means tend to cluster around the true mean. Calculating the mean and the standard deviation for the sampling distribution we get $m = 175.1$ cm and $s = 4.6$ cm.

You are a very kind goddess, the story goes on, and you really want to help them to uncover the truth. You then decide to appear amongst them in disguise. You look exactly as one of them and suggest to re-run all the experiments, but with a larger sample size per experiment[4].

Trying with a sample size of 100 or 1000, the scientists start to see the pattern. The means tend to cluster more and more sharply around a certain value (that you know it is the true value) and tend to be less sparse, i.e. the standard deviation of the sampling distribution gets smaller and smaller, as the sample size increases.

What we have discussed here with this example it is actually one of the most powerful (albeit mysterious) laws of nature. It is actually a theorem, i.e. the fact that this is true can be proven mathematically. It is known as the central limit theorem and it is the main reason why the mean is such an important measure used in statistics. According to this theorem (and as you saw from yourself), the mean of the sampling distribution is equal to the population mean. Moreover, the standard deviation of the sampling distribution shrinks as the sample size increases. This is the mathematical proof of the fact that more observations are always better. Also, the sampling distribution is actually a normal distribution.

The good part of all this is that you do not need to be Athena to make use of this theorem. Again, in real life you never know the true population mean. However, the central limit theorem guarantees you that anyway the mean you calculate from your experiment is a good bet for the population mean. And the larger the sample size, the better is the approximation.

[4] You can run the previous code, just change the variable *SampleSize*.

Estimating parameters 4

Estimation is one of the most important areas of statistics. Much research is devoted into constructing good estimators of the true population parameters. But why does it mean to provide an estimation of a parameter? And most of all, what does make something a *good* estimator for a certain population parameter?

4.1 Point estimators

Working for the Yolo AB was not really your thing. Now that you have learnt one of the deepest secrets of nature, the central limit theorem, you would like to use it to unveil some other hidden pattern in the universe. There is a call for a research expedition. A team of physical anthropologists is going to travel to a remote island to study the people that live there.

Figure 4.1: Jane Morris Goodall (3 April 1934). Goodall is probably the most famous physical anthropologist. You have her blessing for your field study. Source: Wikipedia.

Once you are there, you and your team decide to measure some basic traits, to have a first general idea of the people living there. You manage to get some volunteers and then measure height, weight, and some other things. Here is the code to download the dataset from GitHub

```
Data = read.csv("https://raw.githubusercontent.com/
 OresteAffatato/Statistics-lab-UU/main/Expedition.csv")
```

You would like to start your analysis on the variable weight. What you have in your hands is some measure from a sample drawn from a more general population, which comprises all the people living in that island. What you want to do is to make an *inference*: from the sample you have you would like to make a statement on the general population. In this case, you have measured the weight of the people in your sample, but you would like to generalize this somehow to the weight of the entire population[1].
As a start you might want to see how the data look.

```
Weight = Data$Weight
hist(Weight, xlab = "Weight")
```

The distribution looks quite normal. This make sense, also from a biological point of view. One might argue that too little or too much weight have detrimental effect on the health of the individual and therefore we would expect the weight to be somehow clustered symmetrically around some most typical values, while the most extreme are less likely. Therefore, we would also expect, for theoretical reasons, that the true population distribution in a normal curve. We could then write that $W\ N(\mu, \sigma)$, i.e. the weight of the general population is distributed according to a normal distribution with parameters μ and σ.

So far so good. Now we have to find a way to estimate the true parameters of the population distribution, i.e. μ and σ. One simple way to estimate the population mean is to use the sample mean[2].

[1] We are assuming here for simplicity that the volunteers were all picked at random, and that there is no systematic distortion of any sort in our data.

[2] This is the one that you can calculate from your data.

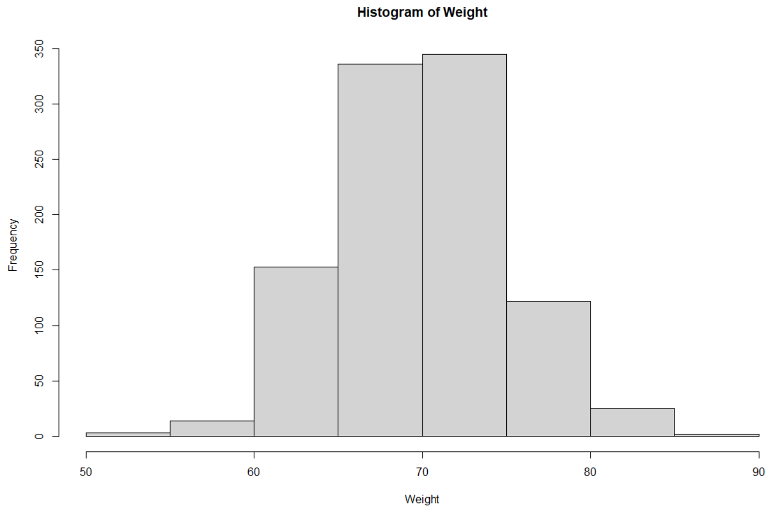


Figure 4.2: Histogram of the variable Weight.

$$m \simeq \mu$$

Is it a good choice? Yes, for several reasons. First, if we did everything correctly, and all the volunteers were picked at random and all accepted to participate, we would expect that the distribution in the sample would somehow be representative of the distribution of the population. However, we have even a stronger argument. Because of the central limit theorem, we know that our sample mean is a good choice to estimate the population mean. All the means gathered from all the similar experiments would tend towards the population mean. Ok, but how do I know if my estimation gives a good approximation? Easy. You know, still because of the central limit theorem, that the larger the sample size is the sharper is the peak of the sampling distribution. Ok, but this looks quite vague, is there any objective way to measure how close is my estimate to the true parameter? YESSA! Still because of the central limit theorem[3] you can calculate the standard deviation of the sampling distribution, also known as the *standard error* (sometimes written as SE).

$$SE = \frac{\sigma}{\sqrt{N}}$$

Where N is you sample size and σ is the usual standard deviation of the population. This all looks beautiful to you, but you are a very alert anthropologist. You notice that you still don't know σ and therefore you just cannot calculate the standard error. However, there is a way to estimate the σ.

$$s \simeq \sigma$$

You could use the sample standard deviation to estimate the standard deviation of the population[4], and therefore you can compute the standard error.

Let me spend some more words on the standard error. This is really the standard deviation of the sampling distribution, i.e. the distributions of all the means of all the repeated identical experiments. As we already

[3] Isn't it amazing?

[4] This is not super-straightforward, you might need some corrections, but the software is going to take care of that. No worries.

discussed, the larger the sample size is, the smaller the standard error. Therefore, most of the means would tend to stay close to each other and close to the population mean. This closeness is a measure of statistical *precision*. We would say that a measure is precise if it is close to the real measure, the thing it is suppose to represent. And that is what the standard error tells us, how close we should be to the population mean, and therefore how precise is our estimation.

What makes an estimator a good estimator?

There are some desirable properties that we would like our estimators to have. In particular, we would like it to actually represent somehow the real value at the population level.
An estimator is said to be **unbiased** if its expected value is the parameter it is intended to estimate. This means that if we would repeat the exact experiment many times, and each time calculate the estimator, this would tend towards the estimand, i.e. the parameter we would like to estimate.
This was just an example of properties we would like our estimators to have. Much research is devoted to somehow prove this and other properties by professional statisticians and mathematicians, to help us, more applied scientists, not to get fooled by the jungle of data.

4.2 Interval estimators

Everything seems to work well for you, you found a convincing way to estimate the population mean μ. However, you do feel that the point estimator alone is not the end of the story. Another valuable information would be how sure are you about this measure, or, in other words, how precise do you think your measure is[5]. The SE is already a measure of this precision, however people often use this information to compute an *interval estimator*. In frequentist statistics, we usually use the **confidence interval**.
Ok, but what is that?
The idea is simple. From the start, you don't know if your sample mean is exactly the population mean[6]. In our case, the mean is around 69 kg. So we are quite sure that 69 kg is a good approximation of the population mean weight. However, also 70 kg seems like a plausible value for that. One thing that we could do is to gather all the numbers that we think are also plausible values for the population mean weight. A single number cannot do that for obvious reasons, and therefore we need an interval of values. So we would like to build an interval that includes all the most plausible values and that somehow traps the true population value. We don't know in general the true value, so we have to decide how safe we would like to play this game. We should establish a *confidence level* κ, which means how sure we are that our interval captures that population mean. The most fundamental building block on this *confidence interval* looks like this

$$P(\mu \in [a, b]) = \kappa$$

[5] As we already discussed variation should be expected in our experiments and we should account for that.

[6] I continue with the previous example on the mean weight, but this applies to any kind of estimators.

This means that we would like to build an interval $[a, b]$ from our sample in which we are $\kappa\%$ sure that μ is inside it.In many application, especially in medicine and biology, the confidence level is usually set at 95%. With this we mean that we would like to be very sure that the true parameter will be captured by this interval[7].

Once this has been established, we have to build the interval. In particular, this interval is going to be centred on the sample mean, because this is our best estimation of the population mean. But how wide? From the theory of the normal distribution we know that the 95% of the probability mass around the mean is 1.96 standard deviations[8] away from the centre. Therefore, the interval should be

$$[\mu - 1.96SE, \mu + 1.96SE]$$

Inside this interval the area of the normal curve is indeed 0.95. This means that if you would pick at random numbers from this distribution, 95% of the times they will be extracted from this very interval.

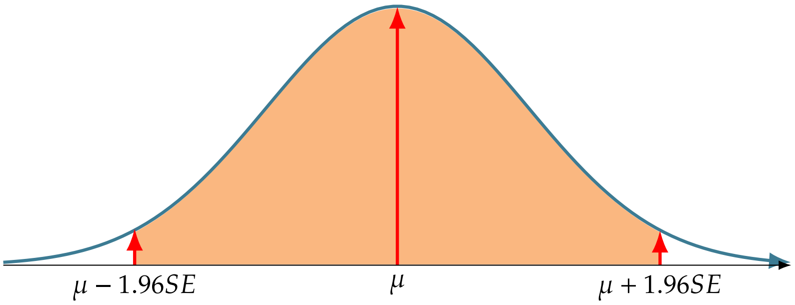


Figure 4.3: Area under the normal corresponding to 95% probability.

Everything seems very beautiful, but it doesn't really work this way. This works only if we know σ (and μ, I guess), but we don't know it. We just use the sample standard deviation of the sample to estimate that. Therefore our SE is also not the proper one. So, there is a price to pay. Instead of 1.96 we have to use another number, $t_{0.05}$. This comes from the *t-distribution*. It is a slightly bigger number, but that's life. Everything comes with a price.

Before going to discuss the R code, just some final things. When we calculate a confidence interval there is always a trade off between *precision* and *reliability*. The width of the confidence interval is a compromise between precision and confidence. Let's have a look again

$$[m - f(\kappa)SE, m + f(\kappa)SE]$$

Where $f(\kappa)$ is some sort of function of the confidence level, depending on the context. According to this equation, the center of the interval is the mean m (or in general our best estimator for the population parameter). But then the width depends on two things: the confidence level (via $f(\kappa)$) and the precision (via the standard error). Therefore your confidence interval could be wide for two reasons (not mutually exclusive): either the confidence level is very high (and therefore you really want to be sure to capture the true parameter) or your measure is really imprecise. Therefore, in general a too wide confidence interval can be really poorly informative.

[7] You will encounter this 95% confidence level virtually everywhere. But remember that this is just a convention, a social norm, there is nothing inherently special about this particular number.

[8] Remember that the standard deviation of the sampling distribution is the SE.

Now let's go back to R. You can do it in two ways. You can calculate it manually or using an R package.
In the first case, just check the $t_{0.05}$ value for a t-distribution with $n-1$ degrees of freedom, where n would be the sample size. Then you can calculate the upper and lower bounds

$$UB = m + t_{0.05}SEM$$

$$LB = m - t_{0.05}SEM$$

The other way would be the following[9].

```
install.packages("lsr")
library("lsr")
ciMean(x = Weight, .95)
```

4.3 General remarks

Estimation is one of the most powerful tools in statistics. In many (if not all) cases, we are interested in an effect size and this is a parameter of a certain distribution. I don't have any particular remark for the point estimator, as it is pretty straightforward. The confidence intervals are a terrific instrument, but they are a bit tricky. They are very informative, in the sense that you get a set of values compatible with the population parameter you want to estimate, i.e. you have a certain confidence level that the true parameter is amongst them (being your point estimator your best shot).
But here is the tricky part. Suppose you calculated a 95% confidence interval, e.g. $[69.6, 70.2]$ for the weight example. A common misconception is that this means that the probability that the population parameter lies in that interval is 95%. In other terms, that $P(\mu \in [69.6, 70.2]) = 0.95$. This is of course wrong. Remember that you are a frequentist, and therefore a probability for you makes sense only in terms of frequencies. The real interpretation of that 95% confidence level would be that if you were to repeat the experiment 100 times and calculate the corresponding confidence intervals, 95 of them would actually contain the true value.
A simulation can help us understand better. Let's imagine to repeat 20 times the same identical sampling of people and measure the mean weight and 95% confidence interval every time.

```
install.packages('lsr')
library(lsr)
install.packages('ggplot2')
library(ggplot2)

N = 20
SampleSize = 50
m = matrix(nrow = 20, ncol = 3)
Experiments = as.data.frame(m)

# Repeating the experiment N times

for (i in 1:N) {
```

[9] This package has been developed by the author of this highly recommended book on statistics [1].

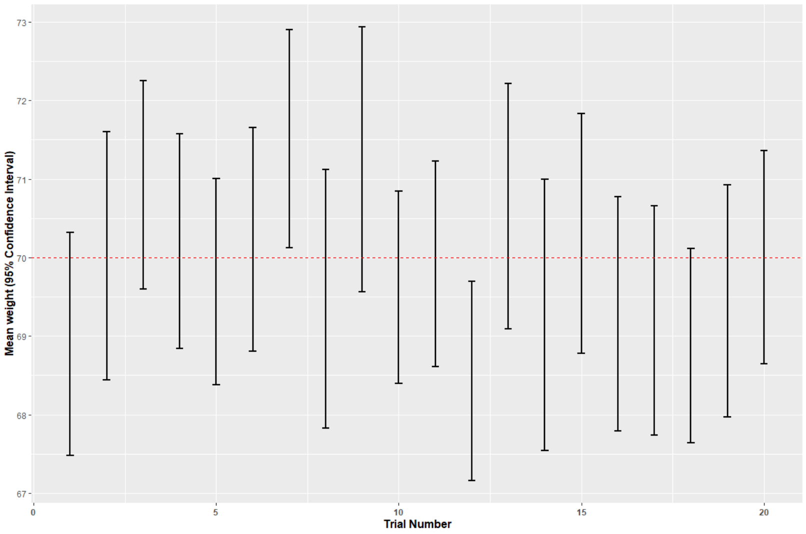


Figure 4.4: Simulation of 20 repeated identical experiments. As you can see, sometimes you just miss the real mean.

```
  Weight = rnorm(SampleSize, mean = 70, sd = 5)
  Experiments[i, 1] = mean(Weight)
  CI =ciMean(Weight)
  Experiments[i, 2] = CI[1,1]
  Experiments[i, 3] = CI[1,2]
}

Trial_Number = seq(from = 1, to = 20, by = 1)

Experiments = cbind(Experiments, Trial_Number)

colnames(Experiments)[1] = "Mean"
colnames(Experiments)[2] = "Lower_Bound"
colnames(Experiments)[3] = "Upper_Bound"

# Creating the forest plot

Forest_plot = ggplot(data=Experiments,
           aes(x = Experiments$Trial_Number,y =Experiments$Mean, ymin = Experiments$Lower_Bound, ymax = Experiments$Upper_Bound ))+
  xlab('Trial Number')+ ylab("Mean weight (95% Confidence Interval)")+
  geom_errorbar(aes(ymin=Experiments$Lower_Bound, ymax= Experiments$Upper_Bound),width=0.2,cex=1) +
  theme(plot.title=element_text(size=16,face="bold"),
        axis.text.x=element_text(face="bold"),
        axis.title=element_text(size=12,face="bold")) +
  geom_hline(yintercept=70, linetype="dashed", color = "red")

# Printing the plot
Forest_plot
```

We should get something like this. Notice that I have highlighted the true population parameter with the red dashed line, but remember that of course you never know this parameter. We know it only in this case because we artificially simulated the process.

From the forest plot in Figure 4.4 you can see why confidence intervals are so tricky to understand. Most of the times, they manage to capture the true population parameter. However, sometimes they miss it. If you set the confidence level at 95% you are fairly sure to be able to trap it. The point is that in reality you just don't know which one is the true parameter and you do just one experiment. Therefore, you never know for sure if the confidence interval you calculated contains the population parameter or not. All you can tell, is that you are quite sure to have it, depending on the confidence level that you set. But once you have calculated the confidence interval it's done, the population parameter is either inside or outside.

4.4 Exercises

1. Plot the other variables and see if the normal distribution is a good approximation;
2. Discuss the best estimation of the true parameter (to mean or not to mean?) and the values in the 95% Confidence Interval also in a biological perspective;
3. Try also other confidence levels and see what happens;
4. Bonus: run the simulation again and change some parameters and see how many times the confidence intervals miss the true value depending on the confidence level;

Hypothesis testing 5

To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.
- Ronald Fisher.

Hypothesis testing is the other important approach in statistical inference. It is based on decision-making as its relies on assessing whether there is enough evidence against a certain statement, in favour of an alternative one. Still given the fact that it is widely used and important, it is also quite controversial, especially for some misuse of these methods. We will discuss the general principles and some examples of hypothesis testing in this chapter and leave the criticism for a later discussion. The important issue here is that we understand what the tools offered in this framework actually mean.

5.1 General principles

We will start discussing the general ideas. We will follow the discussion in [1], a reading the I really recommend anyway[1].
After your field expedition you decide to dedicate yourself to some more controversial subject. You would like to see whether clairvoyance is a real thing. To do that you enroll, let's say, 100 individuals and you have them undergo a very simple experiment. In a room there are two chairs and a table between them. You sit in one of the chair and put a box on a table. The box might be empty or not. The participants enter the room one at the time. You decide whether to put something inside the box and only you have the access to the content of the box. After everything is ready you just ask each participant if the box is empty or not.
Before looking at the data and find a way to analyze them, there is a first step. As you might have noticed we have a *research question* or *research hypothesis*. In our case this would be that clairvoyance is real. Problem is: how do we connect this to the data? To solve this, as a first step, we have to translate our research hypothesis into a *statistical hypothesis*, stated in a precise mathematical language. One way to do that, in our case, is that if clairvoyance was real the probability of correctly guessing the content of the box should be different from pure chance, i.e. $\pi \neq 0.5$[2].

- Research hypothesis: clairvoyance is real.
- Statistical hypothesis: $\pi \neq 0.5$

To somehow prove this we use a rather counterintuitive argument. The idea is that we have to focus on the opposite hypothesis, called *null hypothesis* H_0, instead of our starting hypothesis (which is referred to as the *alternative hypothesis* H_a). The reasoning is: instead of measuring the evidence in favor of H_a, we measure the evidence against H_0. In our case, since we want to prove that $\pi \neq 0.5$, so

[1] The author provides the best explanation of hypothesis testing I've ever read so far.

[2] Again, I use the Greek letters for the population parameter. So this is the probability at population level.

- H_a : $\pi \neq 0.5$
- H_0 : $\pi = 0.5$

Why did the statisticians created this weird setting? The main reason, I think, is that calculations are a lot easier if our task is to measure the evidence against the null hypothesis. And this was a great advantage a century ago when these methods were developed.

This is the basis of the hypothesis testing machinery: we want to measure the evidence against the null hypothesis H_0 to somehow disprove it and therefore prove our alternative hypothesis H_a[3]. Later I will explain better what *measure the evidence* and *disprove* mean.

The problem is that in statistics we are never sure about things, and even though we did everything correctly, we might make some mistake, just by chance. In particular, our H_0 could be either true or false and we might reject it or not, based on our analyses. So these are the possible scenarios

Table 5.1: Possible outcomes for H_0

	reject H_0	keep H_0
true H_0	👻	✓
false H_0	✓	👻

The ghost means that you made a mistake, your discover is gone forever. In particular, there are two types of error, according to the sacred scriptures:

- if you reject H_0 when it is actually true, we have the *type I error*;
- if you keep H_0 when it is actually false, we have the *type II error*;

These two errors don't share a balanced importance. I would say that it really depends on the context[4], but in general and as a common practice, the type I is considered the most important. So, in general we are likely to make to make such errors. Let's call α the probability for the type I error and β the probability for the type II.

Table 5.2: Probability rates

	reject H_0	keep H_0
true H_0	α	1 - α
false H_0	1 - β	β

As I said, usually we want α to be very low. In general, for a test we set $\alpha = 0.05$, according to a tradition that goes back to Fisher, the statistician that created the null hypothesis testing. This level is also called the *significance level* of the test and it is the maximum tolerable value for the type I error.

Ok, let's have a look at your experiment. Here is the code to simulate it.

```
# Simulating the experiment

N = 100

Experiments = c()

Experiments = rbinom(n = 100, size = 1, prob = 0.5)

# Calculating the successes
Successes = sum(Experiments)
Successes
```

[3] This statement it is a bit dangerous, from a philosophical and statistical point of view. But I guess it is easier to just present the "orthodox" view and then take a giant shit on it.

[4] These errors are related to the false positive/false negative classification. My argument is simple. Ask yourself: if a person gets a result from an HIV test, what is more dangerous the false positive or the false negative?

I've got 60 successes over 100 trials. Now I think we are supposed to weight this evidence. Of course, if you were to get something around 51 or 52 you were pretty safe in assuming that maybe this all clairvoyance thing is all a made up story. On the other hand, if you were to get something like 98 or 99 successes this would have look like compelling evidence that maybe you should revise your understanding of nature.

Before going to that, we need some assumptions on the data generating process. That is the basis to measure the evidence against the null hypothesis. Of course, looking at the code you know exactly which distribution has generated the data and all its parameters. In real life, you never know for sure, and therefore you need to make assumptions on the possible distribution that generated the data you measured in your experiment.

In our case, each data point was extracted from a pool of only two possible outcomes, success or failure, each with a given probability. Therefore, we can safely assume that the generating function is a binomial distribution

$$\mathrm{X} \sim \mathrm{Binomial}(\pi, N)$$

Where N is our sample size and π is the probability of success, the parameter we wanted to measure, remember?

What one usually does at this point, to make a decision on your particular case, is to establish the *significance level*. Again, the general idea is that we want to set a maximum level of tolerance for the type I error. This has to be established before you start the experiment really, but anyway you should do it before you calculate any inferential statistics. Let's set the significance level at 0.05.

Now, the last part. The core idea behind the hypothesis testing is to calculate how likely is it to have such result or more extreme, given the fact the H_0 is true (and of course given all our model assumptions). This is generally fairly easy[5] to calculate and it is called *p-value*.

$$\text{p-value} = p(D \mid H_0)$$

In our case, this would be the probability to witness 60 success (or more) over 100 trials, assuming H_0, i.e. that $\pi = 0.5$. We can calculate this using R

```
binom.test(x = 60, n = 100, p =.5)
```

According to the calculations, the resulting p-value is 0.057[6]. This means that the probability of having 60 (or more)[7] successes in 100 trials, assuming that the real probability of success is 0.5, is 0.057. This is above the significance threshold that we set and therefore we say that there is not any statistically significant evidence against the null hypothesis[8].

I just want to stress, independently of the significance level, that the p-value is intrinsically a probability. Therefore, one could still notice that the probability, for such a weird experiment, to register such a result, assuming H_0 to be correct, is pretty low. Therefore, we might suspect that H_0 is not so solid.

In Figure 5.1 we summarize the core of hypothesis testing.

In the next sections, we will discuss a couple of tests that might be useful to you. There are of course many more, but I want to stress that I don't want to be comprehensive in the discussion. I just want to transmit an

[5] That is one of the main reasons why this method is so popular.

[6] This is a tragedy for the median professor.

[7] Or 40 successes (or less), as this also would somehow prove that people somehow get things wrong systematically.

[8] People would say that you fail to reject the null hypothesis, but in reality there is nothing to reject. The only thing we've got is the weight of the evidence against H_0. I'll explain this better later.

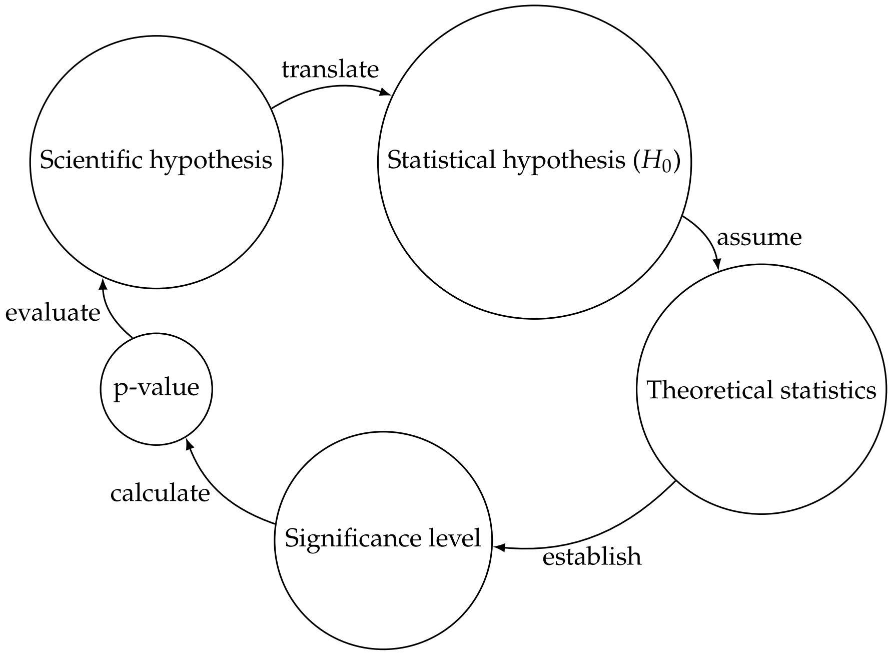


Figure 5.1: Summary description of the hypothesis testing process.

approach: open the machinery of statistics, and see for yourself what is inside. I want you to be aware of the meaning and the implications of the tools that you are using.

5.2 The Student's t-test

According to the legends, the origin of this test is lost in a sea of alcohol. It was in fact designed to measure the quality of the beers.
The mechanism of functioning is quite simple: compare mean values. Let's assume[9] that we know that a given type of beer should have a precise density. To check if the beers I am producing have the same high quality I should compare somehow the average density of my beers with that value, established by the company. One way to do that, is to calculate the mean density and calculate the difference with the established high-quality level. To standardize the procedure, and to take into account the uncertainty of my measures, we could measure the difference in terms of standard errors. What we obtain is the t statistics, defined as

$$t = \frac{M - \mu}{SE}$$

Where M is the mean reference value, μ is the mean you get from your sample and SE is the usual standard error. It can be proven that this statistics follows the so called *t distribution*. So far so good. And now? We have to define our hypothesis.

- H_0: the population means of the two groups are equal, $t = 0$;
- H_a: the population means of the two groups are different, $t \neq 0$;

The t statistics tells us also the direction of the difference of the means. As a matter of fact, $t > 0$ would imply $M > \mu$, while $t < 0$ would imply $M < \mu$.
Ok, now we can proceed with our R coding.

[9] I will explain the one-sample t-test.

```
t.test(x,y)
```

This test works under the assumption of normal distribution of your data, so you should check that. Moreover, as all the other tests, it is used in the general process of hypothesis testing (establish statistical significance, calculate p-value...).

5.3 Pearson's χ^2 test

There are no legends about the origin of this test, as it is quite boring, I guess.
The old good pal Karl Pearson designed this machinery to measure how much our observations differ from an expected theoretical outcome. For example, imagine we are flipping a coin and we would like to know if the coin is fair and not biased. Imagine we flip it 100 times. Theoretically, we would expect around 50 tails and 50 heads. We could use this test to measure the evidence of the adherence of the observations to the theoretical expected outcomes. How do we do that? We calculate the χ^2 statistics[10]

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

Where the sum refers to the number of types, in our case we have only two types, head or tail. O_i is the number if observed instances of the type i while E_i is the number of expected observations of the type i. I guess this might be rather abstract, doing exercise will definitely help you understand. You can already see that if the observed outcomes are the same as the expected, the χ^2 would be zero or close to zero. Let's state this more clearly

- H_0: the proportion of observed outcomes are the same as the expected outcomes, $\chi^2 = 0$;
- H_a: the proportion of observed outcomes differ from the expected outcomes, $\chi^2 > 0$;

This is not the only use of this test, the curious reader could find some more in the books and in internet.

5.4 General remarks

Hypothesis testing is indeed one of the most common inference frameworks available in Statistics. However, one should always keep in mind the limitations of these methods. I'm not strictly referring to the low adherence of your data to the prescribed assumptions. These models can work fairly well even if you cannot strictly meet the requirements. I am mostly referring to the general meaning of the tools you are using. This machinery is nice for making decisions between two competing hypothesis, but in most of the cases you are interested more in the real effect, rather that the pure rejection of an hypothesis. Therefore, you should keep clear in mind the distinction between statistical significance

[10] Which follows a χ^2 distribution. Surprising?

and biological or clinical significance. When it was created by Fisher, the term *significance* was referring to the fact that well performed experiments would for sure hit the significance threshold. Those ones that couldn't did not imply no effect, but just the fact that we needed more evidence or more experiments to be able to conclude anything. Always remember that the final judge it's you. There is no substitution for your biological intuition and this is what is going to decide ultimately what is really important.
We will discuss in a later chapter other problems with this framework.

The eternal war between Fisher and Neyman

The harsh reality is that we don't have a unified theory of the null hypothesis testing, we never had. What we use nowadays is a mixture of two very different approaches, one developed by Ronald Fisher and the other by Jerzy Neyman.
Fisher developed the concept of null hypothesis (and for him that was it) and significance level, and much of his efforts were devoted into calculating the p-values for all the different situations. For him the important thing was to measure the evidence against the null.
Neyman developed the concepts of alternative hypothesis and error type. So the core idea was to measure the weighted evidence of two competing hypothesis for then making a decision.
What is ironic is that nowadays we use a mash-up of both approaches even though back then Fisher and Neyman didn't like each other's approaches[11].

5.5 Exercises

You are still into these borderline science and before getting lost in the endless blissful knowledge of regression, you stay around to do some more fringe research.

1. It turns out that one of your colleagues is a vampire. To help her thrive, you decide to check whether in the town in Transylvania where she lives there is the same distribution of blood type as in the rest of Romania. You get a sample of 1000 individual and you know that in the general population there should be a 40% of individuals with the A type, 10% B, 4% AB and a 46% with the O type. You can find the dataset in GitHub

```
Data = read.csv("https://raw.githubusercontent.com/
    OresteAffatato/Statistics-lab-UU/main/TrueBlood.csv")
```

   In this dataset you registered the expected frequencies of each blood type and the observed frequencies that you measured. What do your data suggest?
2. The Earth is not your thing after all, and you decide to join the Galactic Republic. In Tatooine, there is a screening going on to see whether a group of Jedi has an average midi-chlorian blood density higher than a certain level, which is 152 mg/mL. It is known that Jedi warriors with concentration higher than that limit have above average powers, and this might be useful in hard times. You sample

[11] For a better explanation, and for my source, I suggest to read [1]

100 individuals from the group and you measure the midi-chlorians blood levels. The data can be found in GitHub as usual.

```
Data = read.csv("https://raw.githubusercontent.com/
   OresteAffatato/Statistics-lab-UU/main/MaytheForce.csv")
```

Here you find the blood densities for each participant. What do your data suggest?

3. Things are not so good in the outer space after all, and therefore you come back to the good old Earth. You retire in your house in the countryside and dedicate your time to gardening and stuff. After some time, you notice that there is something going on with your rowans, some infection is spreading. According to the legends, the rowan, or *Sorbus aucuparia*, has the power to ward off evil powers of different sorts. Therefore you would like to keep these plants alive and healthy. You know that, in general, you would expect 20% of your plants to be infected, that's life, that's normal. But you suspect that someone is actively infecting your rowans. You take your measures from a sample of 120 rowans and you register the infection status (1 is healthy, while 0 is infected). You can find the data in my GitHub

```
 Data = read.csv("https://raw.githubusercontent.com/
   OresteAffatato/Statistics-lab-UU/main/WingardiumHerbosa.
   csv")
```

So, what do you say?

Regression 6

The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning.
- Nate Silver, The Signal and the Noise

In this Chapter, we will discuss a bit how we can assess associations. In particular, we will introduce two important tools: the correlation coefficient and the linear regression. These are widely used, but one must always keep in mind the old *mantra*: correlation does not imply causation. Even though this particular catchphrase annoys me immensely, it is of vital importance to understand its meaning.
Back to our story: after all of your wandering, you finally arrive at the library of Babel. This immense library contains an unimaginable quantity of books. All the books have 410 pages and contain all possible texts, written permuting the usual letters of the alphabet (all of the alphabets). Most of the books contain pure gibberish, some other novels that were already written, and some others prophecies of future events. Everything, literally everything, that could have been written, has been written and it is kept safe in the library of Babel. It's now up to you to find sensible information.

6.1 Correlation

So far we have studied random variables separately, in the sense that we didn't assess their co-dependencies. In particular, we would like to have a statistic that measures to what degree two random variables tend to vary together. There are countless examples. One can easily think that the height and the weight of a person tend to vary together somehow, for obvious reasons. The problem is that usually we do research on far more complicated things, and in these cases, one should be cautious.
You grab a book from one of the shelves and you start reading. With great surprise, you realize that the text is intelligible. It is a dataset from an experiment, in which someone has collected several medical-related data from a sample of 500 people. You notice that some of the variables look really strange in the overall context, they don't seem to be related at all. Anyway, you would like to see if there is any pattern.
The most basic machine that you could use is the *covariance*. In general, imagine you have two random variables X and Y[1]. We have our two samples as usual, with $x_1, ..., x_n$ and $y_1, ..., y_n$ data points. We can then calculate the sample means m_X and m_Y. The **covariance** is thus defined

$$cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_X)(y_i - m_Y)$$

Here in this formula we should use the population means, however we used the sample means as estimators of the corresponding ones at

[1] You can imagine we are talking about height and weight.

population level. This formula might look a bit complicated, so we can re-write it, to try to understand better

$$cov(X,Y) = \frac{1}{n}\left[(x_1 - m_X)(y_1 - m_Y) + (x_2 - m_X)(y_2 - m_Y) + ...\right]$$

As we can see, the covariance is constructed by multiplying the deviation of the single data point from the mean in one variable, i.e. $(x_i - m_X)$, with the deviation from the mean in the other variable, i.e. $(y_i - m_Y)$[2]. Three main scenarios can happen at this point

- overall, positive deviations in one variable are systematically associated with positive deviations in the other one, X and Y tend somehow to grow together. In this case, the majority of the terms in the addition are positive and therefore $cov(X,Y) > 0$;
- overall, positive deviations in one variable are systematically associated with negative deviations in the other one, so X tend to increase when Y decreases[3]. In this case, the majority of the terms in the addition are negative and therefore $cov(X,Y) < 0$;
- overall, the are no systematic associations between the deviations in the two variables from the mean, i.e. X and Y tend to vary independently. In this case, the positive terms would balance the negative ones and therefore $cov(X,Y) = 0$;

Let's use the covariance to find some patter in our dataset. First, let's go fetch it in my GitHub page

```
Data = read.csv("https://raw.githubusercontent.com/
  OresteAffatato/Statistics-lab-UU/main/TheBook.csv")
```

We do believe that, for example, height and weight are correlated. For obvious reasons, taller persons tend to weight more than shorter ones.

```
Height = Data$Height
Weight = Data$Weight

# Calculating the covariance

cov(Height, Weight)
```

I've got −0.003[4]. Interesting result. Remember that we are in the library of Babel, we should always use our mind to distinguish between the reality and the games of chaos.
You notice that the author of the study, used the measured heights and weights of the participants to calculate the BMI. The formula is the usual

$$\mathrm{BMI} = \mathrm{w}/\mathrm{h}^2$$

Where w is the weight and h is the height. You suspect that there should be now an association between, for example, BMI and weight.

```
cov(BMI, Weight)
```

[2] Then dividing everything by the sample size.

[3] Or vice versa.

[4] Two things. Technically, the unit of measure of this should be m^2, which makes it rather abstract. It has the same inconvenience of the variance, you might recall. Second, does this result make sense to you?.

[5] Before proceeding any further, I suggest you to calculate the covariance between BMI and height. What do you suspect the result to be? Does the actual result match with your hypothesis?

I've got 4.94. It is positive, as we would expect. Is it large, though?[5] In general, the covariance is quite useful, but it is difficult to interpret the magnitude. Dividing the covariance by the products of the population standard deviations we obtain a normalized version of the covariance, known as the **Pearson correlation coefficient**.

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

This has the same properties of the covariance, but it takes values in the interval $[-1,1]$. What this measure, in reality, is the degree of *linear* association between two variables. In particular, when there is perfect linear relation between X and Y, the coefficient is either equal to +1 or −1. To be more specific, one can prove mathematically that when $Y = a + bX$, then $\rho(X,Y) = 1$, while when $Y = a - bX$, then $\rho(X,Y) = -1$. Remember every time you use this machinery that what you are actually assessing is the level of linear association, or linear correlation.
Let's see in the case of BMI and height.

```
cor(BMI, Weight, method = "pearson")
```

6.2 Linear regression

In this section we are going to show some more sophisticated machinery that you can use to study linear relationships. There are some quantities in the real world that are *linearly related*, i.e. the increase (or decrease) of one is associated to the increase (or decrease) of the other according to a fixed proportion. Many other phenomenon are intrinsically not linear, but the good news is that on a proper scale everything is approximately linear. This obsession for linearity is well motivated, both from a pure mathematical and practical point of view. Again, not because it represents faithfully reality, but because it can be extremely useful[6].
The general equation of a line, you may recall it from past studies, is the following

$$y = mx + q$$

Where m is the angular coefficient and q represents the intersection of the line with the y axis. The m coefficient is going to be very important. In particular, it tells us the slope of the line. A big slope means that for tiny variation of the independent variable we have huge changes in the dependent one. We can also see it in a similar way. Let's consider two points on the line $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$. They both satisfy the equation of the line. So

$$y_1 = mx_1 + q$$

and

$$y_2 = mx_2 + q$$

[6] *All models are wrong, but some are useful* - George Box.

are always true. Let' calculate $\Delta y = y_2 - y_1$.

$$\Delta y = y_2 - y_1 = mx_2 + q - mx_1 - q = m(x_2 - x_1) = m\Delta x$$

If we have a variation of 1 unit in the independent variable[7], this means that $\Delta y = m$, i.e. the angular coefficient is exactly the variation in the dependent variable. This is the interpretation that is going to be useful in the future.

So far so good. What we have seen is about theoretical curves, when we know exactly the relation between the variables. What happens in statistics is the contrary. We have a bunch of data and we assume that they have a linear relation. The question is: assuming that my idea is correct[8], which one is the best line that fits my data?

So your typical situation is going to be something like the following. Imagine you are performing a study assessing the relationship between level of understanding of the real world and number of papers published. Your data may look like this

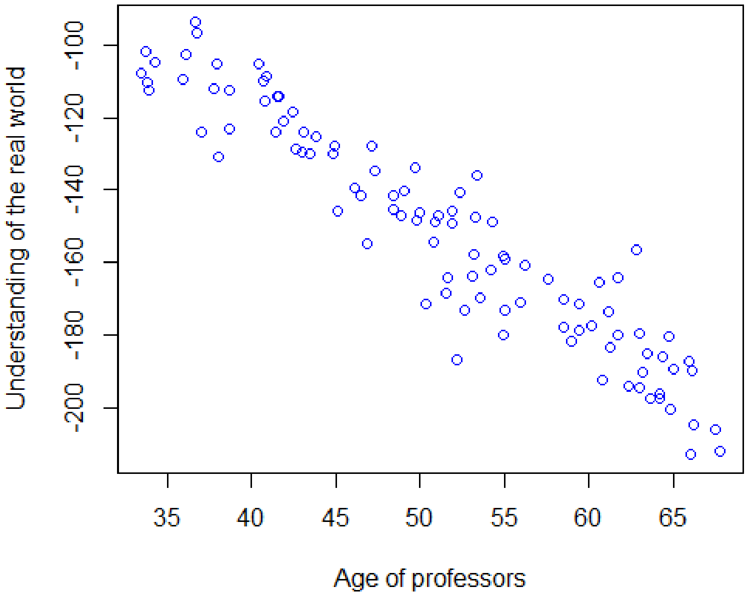


Figure 6.1: Association between age of university professors and general understanding of the world. I hope I'm not getting too many enemies with this joke...

They look a bit messy, as they are real world data, but you can see the pattern, right? They seem to cluster around a straight line. Problem is: which one?

As you know, there is one and only one line passing through two points. But you usually have many and non collinear data points, i.e. there will be of course no line that would pass through all of them.

What we do at this point is that we leave aside the "perfect" line and we look for the best line that can fit our data.

Our equation would have basically the same form as the usual, with some differences

$$y = mx + q + \varepsilon$$

[7] I.e. $\Delta x = 1$.

[8] This means, assuming that the variables are linearly related.

In this case, m and q have the same meaning, but not in the "exact" sense we discussed before. They are the best estimate for our true coefficients. ε is the random error, the noise that spreads our data all over the place. So what we are going to do (and what one usually does) is to use the data to get and estimate of the m coefficient, as this is usually the most valuable information. This tells us how much changes the dependent variable for one unitary change in the independent variable[9].

6.2.1 Simple linear regression - one predictor

Enough of talk! Let's apply this to some examples from the book you found in the library of Babel. One thing that you could study is the difference in IQ between people born in Uqbar and people born in another place[10]. You might wonder how the simple linear regression could help you answering this question. It turns out it very much can. Linear regression could be used to measure mean differences between two groups.

How does it work? So the variable X_{uqbar} can take two possible values, zero or one, as we said. Y is going to be our variable of interest, in this case the IQ of the participants.

$$Y = A + B_{uqbar}X_{uqbar} + \varepsilon$$

When considering the people born outside Uqbar, the above equations becomes...

$$Y\mid_0 = AB_{uqbar}X_{uqbar}(0) + \varepsilon = A + B_{uqbar} \times 0 + \varepsilon = A + \varepsilon$$

However, if you evaluate the expression on the people born in Uqbar

$$Y\mid_1 = A + B_{uqbar}X_{uqbar}(1) + \varepsilon = A + B_{uqbar} \times 1 + \varepsilon = A + B_{uqbar} + \varepsilon$$

Now we are interested in the difference between them, so

$$\Delta Y = Y\mid_1 - Y\mid_0 = A + B_{uqbar} + \varepsilon - A - \varepsilon = B_{uqbar}$$

So the coefficient B_{uqbar} tells us exactly the expected mean difference in IQ between these two groups. Let's see this in action. Our theoretical model is the following

$$Y = A + B_{uqbar}X_{uqbar} + \varepsilon$$

The code to estimate A and B is

```
IQ = Data$IQ
Uqbar = Data$Uqbar

# Creating the model for the line fitting
Regression = lm(formula = IQ ~ Uqbar)
```

[9] Imagine that in the case in Figure 6.1 the slope of the line would be -2. Does this mean that if I publish one more paper my understanding of the world would decrease by 2 points?

[10] The variable *Uqbar* is coded this way, 1 if you were born there and 0 otherwise.

```
summary(Regression)
confint(Regression)
```

As an output I get many things. Once again, remember that in doing research you are the leader, not the follower. Just ignore the outputs you are not interested in.

In particular, we are interested in the estimates of A and B. So $A = 110.1$, which means that if you are born outside Uqbar the predicted IQ would be 110.1. $B_{uqbar} = 7.6$ and this means that if you were born in Uqbar the model predicts a IQ higher by 7.6 points than a person who was born outside this place. So, people born in Uqbar have on average an IQ higher than people born outside, the difference being 7.6 points. Don't forget the confidence interval. In particular, we have 95% CI [3.2, 12.0][11].

6.2.2 Improving the model - multiple linear regression

The linear regression is a quite versatile tool, as it can be easily generalized to study more complicated phenomena. In particular, you might want to include the influence of more than one predictor, so the model would look like this

$$Y = A + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + ... + \varepsilon$$

We are usually interested in the estimation of one predictor. We include the others because we would like to have a mild form of *segregation* of the data. This means that we would like to assess the prediction of one variable, one predictor, *while the levels of the others are held constant*. An example would probably make this point clearer.

Ok, people born in Uqbar are on average smarter, but maybe the IQ level might also depend on the education level of the parents. We could easily imagine that parents with higher education level would somehow educate their children giving more importance to the intellectual and rational sphere of the kids. Therefore we would like to assess the difference in IQ, once again, but comparing them within equal levels of education of the parents[12].

Going back to our example, the theoretical model would be

$$Y = A + B_{uqbar}X_{uqbar} + B_{edu}X_{edu}\varepsilon$$

Let's see what our data would tell us.

```
IQ = Data$IQ
Uqbar = Data$Uqbar
Education_level = Data$Education_level

# Creating the model for the line fitting
Regression = lm(formula = IQ ~ Uqbar + Education_level)
summary(Regression)
confint(Regression)
```

Let's have a look at our estimates. The intercept is usually not interesting, however I'll recall that it represent the value of Y, the IQ in our case, when all the other predictors are equal to zero, in our case when the participant was born outside Uqbar AND the education level score of

[11] Don't forget to round the numbers! Of course you don't know these quantities up to the precision of the R output.

[12] This variables takes value from 0 to 20, and somehow scores the level of education attained by the parents of the participants. The higher the better, *ça va sans dire*.

the parents is zero.
Usually we are interested in one estimate, one slope, and this for reasons that will be clear later on. Anyway, let's have a look at both of them. B_{uqbar} = 9.3 with 95% CI [6.5, 12.1]. This means that on average, comparing people whose parents have the same education level score, we would expect people born in Uqbar to have 9.3 IQ points higher than people born outside. The confidence interval looks quite narrow, therefore the population parameter is anyway close to 9.
The other estimate is B_{edu} = 3.2 with 95% CI [2.9, 3.4]. This means that, amongst people in the same group[13], i.e. with the same value of X_{uqbar}, we would see a difference of 3.2 in IQ points for a unit difference in education level score. So for example, amongst all the people in Uqbar, a participant whose parents have 27 points in education level is expected to have 3.2 higher IQ than a participant whose parents score 26 in education level scale. The confidence interval, again, looks quite narrow, so this is quite a precise estimate.
It is a lot of information, I suggest you to take a break and come back to this later.

6.3 Final remarks

I love the metaphor of the library of Babel, because what we do, when we do research, it is more or less to wander around a universe of symbols. We look for meaning, but sometimes we can mistake a trick of chance for a real sign, we disregard what is important in favor of what is futile instead.

The linear regression, as well as its generalized version are incredible tools, but they have clear limitations. Here we discuss some of them, following a book[14] by Andrew Gelman and others, gods and goddesses of statistics [2]. I will report them exactly as the authors did, in order of decreasing importance. I suggest to read the original, because I can only worsen the arguments.

- he most important feature of our model is the *validity*. This means that our data should really reflect the research question we would like to answer. And therefore our model should include all the important predictors that are vital to understand the phenomenon under study. This is probably the hardest task;
- the second important feature is *representativeness*. Our sample has been created with the purpose to make an inference to the general underlying population;
- the third feature is *linearity*[15]. The relationship between the predictors and the outcome of interest should be linear, at least within a reasonable approximation. This, however, is not a strict requirement and one could also add non-linear terms to the model;
- the fourth feature is the *independency of errors*, which means that the deviations of each data point from the estimated line, also called *residuals*, should be independent from one another;
- the fifth feature is the *equal variance of the errors*, i.e. for each predictor the variance of the residual should be approximately the same as for the other predictors;

Figure 6.2: Jorge Luis Borges (34 August 1899 – 12 April 1586). Borges is probably one of the most famous Croatian poets, known for his poems on prose, contradictions, lists and lies. Source: Wikipedia.

[13] This means either all the people born in Uqbar, or all the others.

[14] A super recommended reading!

[15] Notice that Gelman and others put the first purely mathematical and technical assumption at the third place.

- the last feature is the *normality of errors*, i.e. the residuals should be normally distributed.

The last two are rarely important when we are interested in causal inference.

6.4 Exercises

1. Have a look at the dataset contained in *TheBook.csv*. Pose at least 4 research questions and try to use the linear regression to answer. Some things in this dataset do not make sense, some others do. Remember that is your intuition and your knowledge that should guide you, not the R output.

Calling bullshit 7

The only function of economic forecasting is to make astrology respectable.
- John Kenneth Galbraith.

The title of this chapter refers to a famous book [3]. It is in general a book I strongly recommend to anyone and in particular those ones who would like to continue to work in science, but I would say that it is nice to read if you are interested in knowing a bit more about how our society works, with respect to understanding data.

Anyway, why am I mentioning this book in this context? Because it is a general approach to science that I strongly recommend and that I would like to spread. Not only the scepticism, but also the awareness that some of the tools offered by Statistics are often misused (sometimes even on purpose!). That is why it is of vital importance to learn more and to be aware of the meaning and limitations of the tools we are using.

I don't have any particular truth to share. I've just come across some issues, mainly my own mistakes, my own misunderstanding, and I would like to share those episodes, with the hope that they might be useful to you.

7.1 The (in)famous Table 1

This is where it all started for me. Not so long ago I was doing an epidemiological study and I was asked by someone to do the usual Table 1, the table with the descriptive statistics of the cohort whose data I was analysing. So far so good. I put some descriptive statistics, but then I was asked to "put the p-values". Translated in more proper terms, to do some t-tests to assess any difference between the two groups. The usual argument is that if the two groups were to differ significantly (in the statistical sense), that might affect our results. Is that so? Let's look at one example.

So here the authors report the classical Table 1 with the demographic data. So far so good. They also perform some t-test to assess the difference between the two groups. In this case, the patients have atypical anorexia nervosa, a condition in which the patient has the normal symptoms of anorexia nervosa, but weight in the normal range. Let's consider the age difference. What a t-test would tell us? A t-test would tell us if there is any difference between the means of the populations involved, not between the groups! And in this context, one should specify which is the target population. If you use a t-test in this context you are implying the mean age of the population of the people with atypical anorexia nervosa is 14.8 and the mean age of the control population is 14.5[1]. We use hypothesis testing to make inferences on the populations of interest. Therefore, it is generally not recommended to perform tests for the descriptives of the cohorts.

But this is not a massive problem, one should just think a bit about the reference population. The issue I mentioned at the beginning I think is

[1] Of which already your parents are massive outliers.

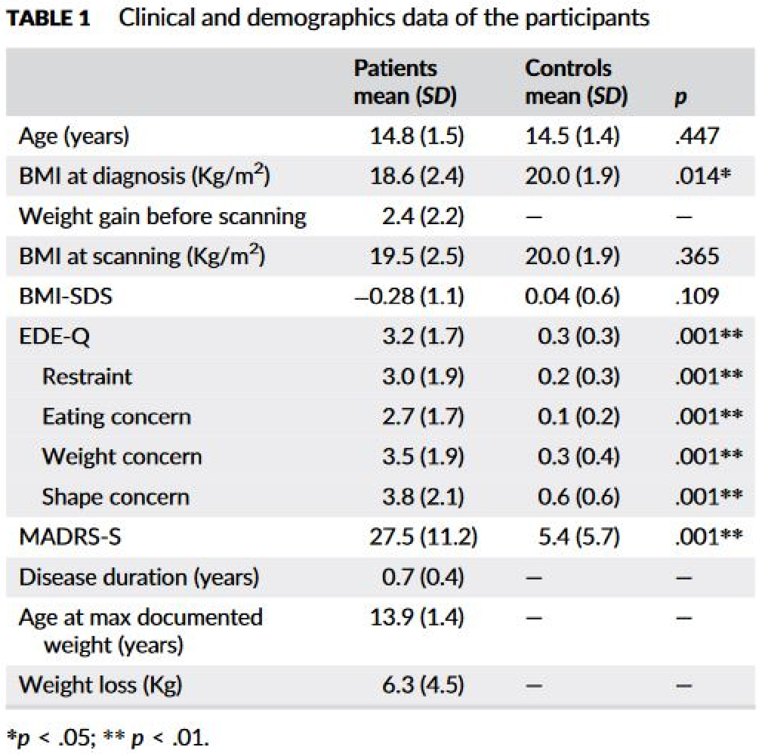


TABLE 1 Clinical and demographics data of the participants

	Patients mean (*SD*)	Controls mean (*SD*)	*p*
Age (years)	14.8 (1.5)	14.5 (1.4)	.447
BMI at diagnosis (Kg/m^2)	18.6 (2.4)	20.0 (1.9)	.014*
Weight gain before scanning	2.4 (2.2)	–	–
BMI at scanning (Kg/m^2)	19.5 (2.5)	20.0 (1.9)	.365
BMI-SDS	–0.28 (1.1)	0.04 (0.6)	.109
EDE-Q	3.2 (1.7)	0.3 (0.3)	.001**
Restraint	3.0 (1.9)	0.2 (0.3)	.001**
Eating concern	2.7 (1.7)	0.1 (0.2)	.001**
Weight concern	3.5 (1.9)	0.3 (0.4)	.001**
Shape concern	3.8 (2.1)	0.6 (0.6)	.001**
MADRS-S	27.5 (11.2)	5.4 (5.7)	.001**
Disease duration (years)	0.7 (0.4)	–	–
Age at max documented weight (years)	13.9 (1.4)	–	–
Weight loss (Kg)	6.3 (4.5)	–	–

*p < .05; ** p < .01.

Figure 7.1: A Table 1 from this paper. This is just an example, I'm not saying anything bad about the whole paper (which I haven't read thoroughly), I would just like you to think about the use of statistics in Table 1.

most worrisome. What a test like this would tell you is the measure of the evidence against the hypothesis that certain group means are equal. That's it. It doesn't tell you that this has an impact on the outcome of interest. You should neither use this argument to justify putting this variable in your statistical model. In the first case, a regression coefficient would tell you the impact of a variable on the outcome (conditioning on the complete set of predictors). In the second case, you should build you model using your *a priori* understanding of the phenomenon, and not the very data you are going to use to make an inference[2].

The take-home message here is simple: always ask yourself why you are to do certain things. You would like to have in your research the exact number of statistics that would help you understand nature and to sustain your arguments, no more, no less[3].

7.2 Putting things in perspective

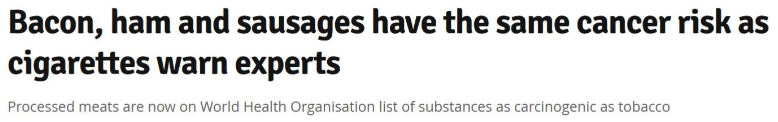
Here, to explain my point I will follow an example and the arguments that I found in [4].

In November 2015, the International Agency for Research in Cancer (IARC), an institution of the World Health Organization, included processed meat in the Group I carcinogen, the same category that contains cigarettes, just to give you a reference. Later on, we find this article in an English newspaper. In this article, they mention the IARC report in which it has been claimed that eating 50 g of processed meat per day was associated with an increased risk of bowel cancer of around 18%. It seems quit a high increasing of relative risk, is it that so dangerous?

What is mentioned in the press release is in fact the relative risk. To have better understanding of what is going on we should always take

[2] Another issue that we haven't discussed arises when doing multiple testing and it is called *multiplicity problem*. I suggest you to read more about it, but the message I want to convey is the following: don't do a test on something you are not interested in in the first place

[3] Less would be insufficient to prove your point, more would be deceiving.



Bacon, ham and sausages have the same cancer risk as cigarettes warn experts

Processed meats are now on World Health Organisation list of substances as carcinogenic as tobacco

Figure 7.2: The article from Daily Record's.

also the absolute risk into account. This last one can give us the general magnitude of the phenomenon we are studying. If we check the absolute risk, we discover that in the general population 6 individuals every 100 would develop the bowel cancer. Now, if we consider a similar population in which all the individuals eat 50 g of processed meat per day, they would have an increase of 18% in relative risk, implying that the absolute risk would reach 7 individual over 100. Therefore, we can conclude that eating 50 g of processed meat per day increases the absolute risk in a population from 6% to 7%. Does it sound so terrible now?

The whole point is that, in some cases, a huge increase in relative risk might imply very little for the growth of the absolute risk because the phenomenon that you are studying is very rare. Reporting both absolute and relative measures is important to put everything in perspective. Moreover, the importance of these statistics depend on the context, right? For example, if you were studying the pathophysiology of cancer, which one would interest you? And if you were a public health expert, whose job is to allocate funding for new places in oncology department, which one would be most informative?

7.3 Predictors of success

For this section, I will follow a classical example found in [3] adding some glimpses of personal experience[4].

In the previous chapters, we mentioned some ways to measure associations between two random variables and we briefly completed the discussion saying that two variables are indeed associated when they tend to systematically vary together. The crucial point is that, in many cases, these variables tend to vary together because there is a common mechanism influencing both. It is the intervention of this (and others, of course) mechanism that provokes that trend. Associations are useful because they have a sort of predictive power and, when are well performed, they are intrinsically hypothesis-generating and allow us to further explore causative relationship. However, we should keep in mind that causation is a totally different business. When we talk about causation, we talk about actual intervention. A real life action changing one variable has a more or less direct impact on the other one. This is of course not true in the case of pure associations.

Many studies have been performed in the field of psychology assessing the relationship between delayed gratification and success later on in life. In one of these iconic studies, 4-year kids were offered two options: either have one marshmallow at any time or to wait and get two of them[5]. What this study found is that kids that waited until the end, had better scores in high school and general good relationship with their parents. I just mention the fact the the paper concluded that delayed gratification was a predictor of later success in the academic career and emotional well-being during adolescence. So far so good. But then what is the problem? You can guess it by now. Massive rampage of press releases and

[4] To find out how dangerous can be random correlations, I would suggest you to visit this web page: https://www.tylervigen.com/spurious-correlations

[5] The overall waiting time set by the researchers was 15 minutes. I guess an eternity for a small kid.

books promoting delayed gratification to be successful in life. I imagine guys reading these pieces of garbage and the preventing their kids to eat marshmallow so they will one they become president of the US or leaders at NASA.

Majestic the human mind, isn't it? From a normal study in psychology, many people inferred causation and promoted a behaviour that cannot *per se* grant the success in life that they wanted. The point is that in this, as in other context, there was a common cause influencing both the variables. Just to conclude our examples, further research has been done, suggesting that was the stability of the family to explain the relationship between delayed gratification and academic success.

Associations are nice tools *per se*, we just have to keep in mind what they imply and how we can use them. Causation, in general, does imply association, whereas the contrary is just not true. However, if there is no association between two variables, in no way there can be a causative pathway. For this reasons associations can be used as evidence to further explore possible causative relationships.

Once, a famous scientist told me that to hire a technician in a lab is not a predictor of success. And therefore this scientist was not willing to hire one. What would happen if I were to hire a technician in my lab tomorrow?

7.4 Confidence or credibility?

Another typical misconception comes from the interpretation of confidence intervals. We already discussed this a bit before, so I won't cover it too much here.

A common idea is that the calculated 95% confidence interval is used to claim that the probability of the true parameter to be inside the interval is 95%. Indeed, these intervals are built this way, but the major problem is that probability has to be interpreted in a frequentist way. Therefore, probability refers to the frequencies of repeated identical experiments. From our perspective, once the experiment is done, that's it.

The problem originates from the fact that we tend to interpret our interval estimators in a Bayesian way. In that case, the claim makes sense, however, the intervals are calculated differently. Therefore, we cannot confuse one with the other, as it is not only a matter of first principles, but also of actual calculations.

7.5 No effect whatsoever

Another pernicious misconception related to confidence intervals[6] happens when disgracefully the zero falls inside the confidence interval. The median professor is highly triggered as she might lose her tenure because of this.

Enough of jokes, what is at stake here[7]? So the problem is that, if the confidence interval contains zero people usually claim that there is no effect, even in cases in which the estimated effect is quite big. One should avoid confusing low precision, i.e. having a large standard error (and therefore a wide confidence interval), with the claim that there is no

[6] And to p-values, but we will come back to that.

[7] Decency.

effect underlying. Here in Figure 7.3 is an example

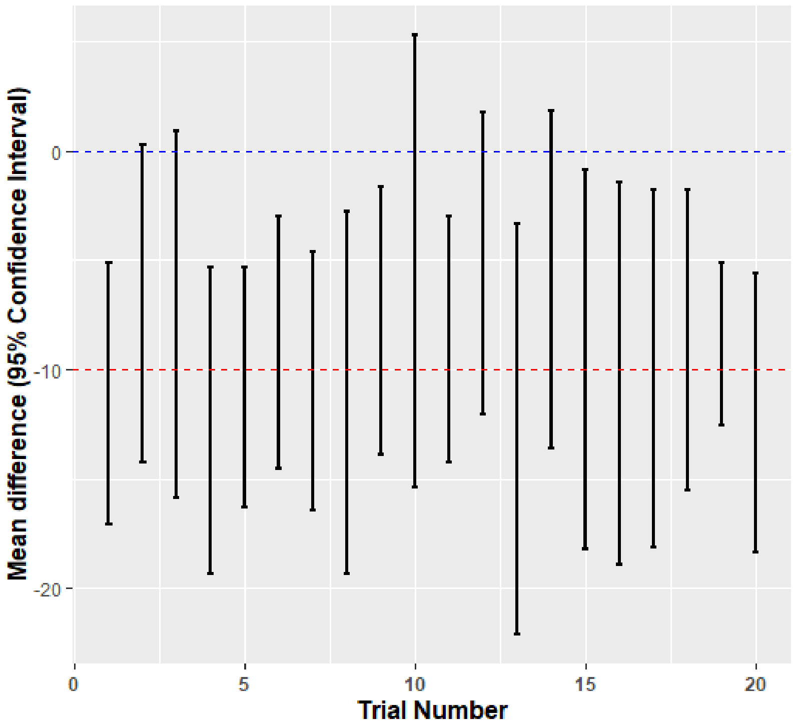


Figure 7.3: Simulation of repeated experiments estimating a mean difference. Sometimes the confidence intervals cross the zero, but the real effect is non-zero.

As you can see in Figure 7.3, sometimes the confidence intervals we calculated include zero. Some people would claim that there is no effect. But this is wrong, as you can see. It just means that given our data and our model the zero is a compatible value for the mean difference, but still our best estimate (in this case is more or less the center of the interval) is in general non-zero. In this case it is easy to know, because I artificially simulated the experiments, drawing data from a normal distribution with mean -10. So we know the correct answer, while in reality we don't know. But here I just wanted to prove the point that having the zero inside the confidence interval doesn't mean no effect.

In almost all applications there is always an effect. No one really believes in null effects. The universe is a big, crazy place, and you should expect some sort of effect, even tiny. But this issue cannot be solved using statistics. Since the effects are always there, you have to decide if it is something important or not. One should always avoid throwing the responsibility on statistics. Statistics is there to help you understand uncertainty, not to tell you what's important or meaningful.

7.6 The Sun is shining (p < 0.05)

Now, finally, we reached the root of all evil. That damned p-value! Go to Pubmed or Google Scholar and look for papers about p-value and its misuses. They are just countless! What we will try to do in this section is to clarify its meaning, what it can tell and, most importantly, what it cannot tell.

First of all, its definition. From a purely mathematical point of view, the p-value is a conditional probability. It is the probability to have the observed or more extreme outcomes (D, our data) assuming that the null hypothesis is true (H_0).

$$\text{p-value} = (D \mid H_0)$$

As we mentioned earlier, in our hypothesis testing framework, we would like to make a simple decision: can I reject the null hypothesis on the basis of the evidence I have? Ultimately you want to know if this H_0 is true or not, based on your data. However, the p-value is a sort of backward thinking to reach that point. What you do is assuming that H_0 is true and then try to assess then how it is likely to have my test statistics (or more extreme results). However, this is not a direct method to assess if H_0 is true at all. What you would like to know[8], really, is if H_0 is true assuming that you have the data produced in your experiment. So the probability you would like to compute is

$$\pi = (H_0 \mid D)$$

You don't have to be a gifted mathematician to know that p-value $\neq \pi$. Confusing these two probabilities is known as the *prosecutor's fallacy*. I'll explain this using an example found in [3], modified a bit.
We are in the US and someone has been killed. In particular, they find some fingerprints and they found a match in the FBI databases: you have been called for the trial as the main suspect.
The major evidence against you is the match of the fingerprints in the FBI database. In particular, this database is quite huge so let's assume that the person really guilty is included in the database. In particular, this matching process is very precise, you can get a 1 in 10 million chance to have been falsely classified. Another important piece for our argument is that we have overall 50 million people data included and the guilty piece of shit is amongst them. Your situation looks more or less like the following.

	Match	No match
Guilty	1	0
Innocent	5	50 000 000

Ok, so far not so good. The mechanism seems pretty much precise, and you might as well end up in jail for the rest of your life. But luckily you are a super gifted statistician! You would like to prove to be innocent, given the fact that they found a match. That's your aim, to calculate this conditional probability. But first, let's calculate our beloved p-value.

$$\text{p(match} \mid \text{innocent)} = 1/10\,000\,000$$

So the probability to find the match, given the fact that you are innocent seems very small. The adepts of the p-value church would be forced to reject the fact that you are innocent and therefore deem you guilty. The p-value seems pretty much against you. However, let's calculate the other one

$$\text{p(innocent} \mid \text{match)} = 5/6$$

[8] But this is not possible with hypothesis testing, and in general in the frequentist framework.

So it seems in the end that you were the nice guy, after all. So, even when the p-value is obscenely small, it might still be the case that you are totally wrong about your inference. This problem cannot be solved in the frequentist framework. If you are a Bayesian you can always perform the calculation I just did. Fantastic, right?! So why not everyone is a Bayesian statistician? Because there is a huge price to pay: you need to know the prior distribution. You need to have already some prior knowledge about your phenomenon, that you can then update with your experiments. Problem is: in many case you don't have a reliable prior. This is not always a big deal, but we just have to be aware of this limitation. So, if you know your prior you can actually calculate the following

$$p(H \mid D) = \frac{p(D \mid H)p(H)}{p(D)}$$

This formula is known as the **Bayes' theorem**. This condense in one line what we would like to know as researchers: the probability of our hypothesis being true, given the data we've actually got.

To finish this discussion, I would just like to mention some other problems related to the p-value. If you know its definition, everything should follow pretty much straightforward. Check [4] and [3] for further discussions with examples, or any textbook. For this final remark, in particular, I will follow a bit [4]. Summing up, remember that:

- from the p-value alone you cannot conclude directly that the null hypothesis is true or false. As a matter of fact, when you calculate the p-value you are even assuming that H_0 is true. This is the prosecutor fallacy, it all originates from the fact that $p(H_0 \mid D) \neq p(D \mid H_0)$;
- the p-value alone is never the end of the story, you need the point estimator and the confidence interval in order to have a better understanding of the inference you are making;
- a statistically significant results is by no means also a biologically or clinically significant one, that is up to you to assess.

This is not really the end of the story. The machinery of hypothesis testing is quite spread in science, and in particular it is more likely for you to be published if you manage to hit the magic significance threshold. The scientific publishing system is in general biased towards these so called "positive results"[9]. From a strictly scientific point of view this is meaningless. You just want to run an experiment and see if you have enough evidence to make a decision. From the perspective of estimation inference it makes even less sense, because you anyway end up with a measure of where your effect is likely to be. So for hypothesis testing it might be troublesome in principle, but in estimation you always gain some sort of knowledge. Anyway, in real life the editors are more interested in publishing your work if you managed to hit the significance threshold. You are a very honest and gifted statistician, so you would just do your job and stand your ground with arguments. Some other guy might have shady intentions. In particular, they might decide to change the statistical test, the assumption or some other trickery of sort in order to have a lower and significant p-value. This is called **p-hacking** and it has tremendous consequences, especially for the replication of the studies[10].

The take home message here is very simple: be honest. Decide from the

[9] When you don't get a statistically significant result, you might guess, it is considered negative.

[10] The basic argument is simple: it is hard to reproduce the same results if the previous guys were cheating, isn't it?

very beginning your research question, the design of the experiment and the statistical analysis. In particular, when it comes to Statistics, you should be able to argue for the method you chose before even starting, you need to have arguments in favour of the methods you think are the best. Remember that you are a scientist and ultimately you are interested in gaining knowledge, not in cheating. If you are smart enough, you can use the flexibility in the statistical methods to prove whatever you want. I sincerely hope that with this lab we had together I was able to convey the message that science stands in a complete different ground. It was never about the papers, but about the science.

7.7 Final remarks

These are just some of the many troubles that you may encounter in your scientific journey. I hope that with this Chapter I was able to convey the importance of these issues. If I were able to see further, it was only because I stood on the shoulders of the giants. Therefore, I leave here some suggested readings and with these I wish you the best for your future:

- *Calling Bullshit - The Art of Scepticism in a Data-Driven World by Carl T. Bergstrom and Jevin D. West.* As the subtitle suggests, it is mostly on refining our sense of scepticism, to protect ourselves against major bullshits advertised across many contexts.
- *The Art of Statistics - Learning from Data by David Spiegelhalter.* This books offers a general overview on Statistics, given by a deeply passionate statisticians. Here you really get the feeling of what being a statistician means: helping the others to get things right.
- *Factfulness by Hans Rosling.* A very important book showing how our vision of the world could be strongly biased and eventually just plainly wrong. A refreshing reading.
- *Statistical Rethinking by Richard McElreath.* A great example as a scientist, a wonderful teacher of statistics.

Causal Inference | 8

Felix, qui potuit rerum cognoscere causas.
- Virgilius.

So far we discussed some common ways to make sense of data. In this chapter, we will see that data by themselves are not the end of the story (they never are). In particular, when we want to estimate the causal impact of a variable on another pure statistical methods are not enough. We need a sort of *calculus of causation* that can help us remove all the spurious associations and leave, to the best of our knowledge, only the statistical association due to a causal path.
The debate about causality is endless and probably as old as humankind. However, some important progress has been made in recent years, especially by Judea Pearl and co-workers. Their tremendous work elucidated the role of causation in statistical inference and can help us solve complicated statistical problems easily. In particular, we will see that, even though all their work has strong mathematical foundations, you don't need to master complicated mathematics to use their tools. The material of this chapter is inspired by their work and will follow [5, 6].

8.1 Once upon a time, a paradox...

You are working in a pharmaceutical company and they have just sent you the results of an observational study you designed. Your drug *DAGavir* looks not so promising in treating the terrible *Confounding syndrome*. The data look like the following

	Drug	No drug
Combined results	780 out of 1 000 (78%)	830 out of 1 000 (83%)

Table 8.1: Table with the aggregated recovering rates.

After a while, one of your colleagues comes by and suggests looking at the data, stratifying by sex. You can see that the picture already looks quite different

	Drug	No drug
Women	232 out of 250 (93%)	670 out of 770 (87%)
Men	548 out of 750 (73%)	160 out of 230 (69%)
Combined results	780 out of 1 000 (78%)	830 out of 1 000 (83%)

Table 8.2: Table with the segregated recovering rates, stratification by the variable sex.

As you can see, when you segregate your data according to sex, the results seem to point to the opposite conclusion. But how is it possible that a drug is not beneficial in the overall population, but it is beneficial in each sub-population? These sorts of phenomena, when a statistical trend holds for the entire population but is reversed in each sub-category, are all examples of the *Simpson's paradox*.
Ok, that looks amazing, we found a paradox, but we still have people to treat. Should we prescribe this drug or not? Should we trust the

aggregated or the segregated results? The point that Judea Pearl makes is that the answer to these questions is NOT in the data. We should rather ask ourselves what we know about the process that generates the data.

Suppose that your friend comes back and tells you some additional facts we know about the drug. Apparently, men are more likely to be prescribed the drug, because of other medical reasons (as you can see from yourself looking at Table 8.2). Moreover, from other studies, we know that testosterone decreases the likelihood of recovery, independently of the action of any drug. We could summarize this by saying that being a man is both a common cause of taking the drug and of a more difficult recovery.

Does this solve our problem? Absolutely! Now we know that to have a more fair comparison we need to segregate the data, thus comparing women with women and men with men. In this way, we are more confident that the differences in recovery rates are due to the action of the drug and not some other external factor.

This story has several important features that I would like you to consider when you are interested in estimating the causal effect of exposure to an outcome.

- the data and our statistical methods only provided us with some associations between certain variables;
- we needed some extra information and some non-statistical modelling of the process that generated the data to solve the paradox;

8.1.1 Exercises

1. According to a recent study published in the *Journal of Baywatching*, accidents in the sea are positively correlated with the number of lifeguards patrolling the beaches. Scientists suggest decreasing the number of lifeguards to decrease also the number of accidents. Do you agree?
2. A pioneering study, published in the *Journal of the Caribbean Sea*, proved that global warming is highly negatively correlated with the number of pirates. As the number of pirates decreases over the years the average global temperature is rising. They conclude that to save the planet we should all become pirates. Do you agree?
3. Apparently there are only two physicians in your town that can treat kidney stones: one is a very young medical doctor while the other has way more experience. It appears that the young doctor has a higher success rate in the surgery for kidney stones, overall. However, kidney stone surgeries are not all alike. They can be divided according to the size of the stone (small or large) and therefore to the severity of the disease. When we segregate the data according to the size of the stones we find that the trend is the opposite, the more experienced doctor has a higher success rate. Why is that? Would you consider the segregated or the aggregated data to establish which physician is better? Imagine you have kidney stones and you don't know the size. Which physician would you prefer to treat you?

8.2 Writing down our causal assumptions

It turns out that there is a way to formally describe our causal assumptions for a given phenomenon under study. Remember that our main goal is to estimate the pure causal association between a given exposure and a given outcome and therefore we would like to remove spurious associations that mess up with our estimand. In other words, we would like to remove the *confounding* effect from our target estimation.
The language to draw our causal assumptions is the one of the *causal Directed Acyclic Graphs*. Let's have a closer look.

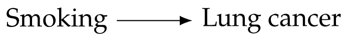


Figure 8.1: Example of a causal DAG

The graph in Figure 8.1 represents a causal DAG. Exposure and outcome are the *nodes* of the DAG while the arrows are called *edges*. This DAG implies that smoking *causes* lung cancer. Therefore we would expect to see a statistical association between smokers and lung cancer diagnosis, for example. This should also clarify a bit the name, causal DAG. *Causal* because it summarizes our causal assumptions. *Directed* because all the edges are arrows that display the direction of the causative effect. *Acyclic* because a closed path, a cycle, would imply that something causes itself. Our causal assumptions should avoid that. *Graphs* because they are graphs.
Our example in Figure 8.1 is pretty much straightforward. The problem is that things are not always that easy. Indeed, such a causal path would imply a statistical association between smoking and lung cancer. However, there are some other paths that might create a statistical association even when there is NO direct causal path between exposure and outcome. Our goal is to remove this effect that *confounds* our results and therefore, just as in Simpson's paradox, to decide whether we should segregate our data or not according to a certain variable.

8.3 The common cause

Imagine we are trying to study the association between having yellow fingers and lung cancer. To the best of our knowledge, yellow fingers are not causally related to lung cancer and therefore our causal DAG would look like this

This example is taken from Hernán. You should definitely check this course, it is amazing. And Miguel Hernán is a world-leading expert in causal inference.

Yellow fingers
Lung cancer

Figure 8.2: First causal model.

Imagine now that we have a huge cohort, with many participants, and that from these people we gather a lot of data, in particular the color of the fingers and the diagnosis of lung cancer. Would you expect to see some association between these two variables? According to our causal model, there should be no association, i.e. the proportion of people with lung cancer should be the same amongst people with and without yellow fingers. However, you do see an association. Why is that? There might be that another variable is somehow creating an association between having yellow fingers and a lung cancer diagnosis. Consider, for example, the variable smoking. Smoking causes people to have yellow fingers and we also know that it increases the risk of developing lung cancer. We say that smoking is a *common cause* of our two variables. The causal DAG now would look like this

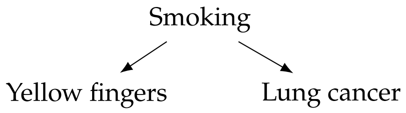


Figure 8.3: Smoking as common cause.

This causal DAG matches better with our intuition and with our understanding of the phenomenon. We do believe that the association between having yellow fingers and lung cancer is entirely explained by the fact that smoking is a common cause of both. Ok, but how do we get rid of this confounded association? We understood that smoking is the problem, so we might consider stratifying by this variable and comparing smokers with smokers and non-smokers with non-smokers (just as we did in the Simpson's paradox example). After stratifying, consider the group with only non-smokers. Would you expect to see and association between yellow fingers and lung cancer? Most likely not, because now the fact that they have yellow fingers and/or lung cancer is unrelated to smoking. Stratifying, as well as other segregating procedures, is a way to eliminate the confounding due to the common cause. When we do such a thing we say that we *condition our model within the levels of a certain variable*. In our case, stratifying by smoking status we conditioned our model within the levels of smoking, i.e. we compared smokers with smokers and non-smokers with non-smokers. In more complex models, this is very hard to visualize, but the process is exactly the same. Conditioning within the level of a certain variable is reported in the DAG by drawing a box around the target variable

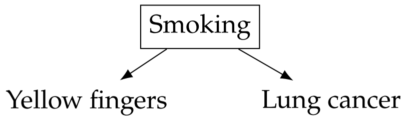


Figure 8.4: Conditioning within the levels of smoking.

What we discussed here is quite general. A common cause makes the information flow between the two variables that it causes, thus generating a statistical association. If we are interested in pure prediction, that is fine. However, if we are interested in estimating the causal effect this is dangerous and we should block this flow.

- a common cause always creates an association between the two variables that it causes;
- conditioning within the levels of the common cause removes this association;

If you are not convinced completely here is an R script that might clarify this a bit better. First, we have to download some packages.

This example, and the code, is taken from Richard McElreath. Watch this for a better explanation and further details. Of course, I recommend his amazing book as well [7].

```r
install.packages("Rlab")
library(Rlab)
N = 1000
```

We are imagining having a trial of N observations. Now we are going to simulate a Bernoulli process.

As the code shows, the Bernoulli process is when we have N observations and only two possible outcomes, 0 (failure) and 1 (success). Success has a certain probability p, while failure has (obviously) probability $1 - p$.

```r
CommonCause = rbern(N, prob = 0.5)
Exposure = rbern(N, (1 - CommonCause)*0.1 + CommonCause*0.9)
Outcome = rbern(N, (1 - CommonCause)*0.1 + CommonCause*0.9)
```

As you can see, the common cause is generated by a normal Bernoulli process. The exposure and the outcome do depend on the common cause via the probability. Depending on the value of the common cause, they have a different probability of success. We can summarize the result of the simulation in a contingency table and measure the correlation.

```
table(Exposure, Outcome)
cor(Exposure, Outcome)
```

As you can see from the diagonal of the table exposure and outcome appear to be quite correlated. This is further confirmed by the correlation coefficient which in this case is equal to 0.63. Let's condition within the levels of the common cause and then calculate again the correlation coefficient.

```
cor(Exposure[CommonCause == 0], Outcome[CommonCause == 0])
cor(Exposure[CommonCause == 1], Outcome[CommonCause == 1])
```

In the first case, I consider all the cases with the common cause equal to zero, while in the second equal to one. The correlation in the first case is 0.02 while in the second is −0.01. The association basically disappeared, from around 0.60 (which is commonly regarded as a medium effect) to a virtual zero.

8.4 The collider

With the previous study, we were successful in solving the problem of confounding due to the common cause. You then receive a call to solve an even harder problem. The story goes like this. Our genes and the environment in which we live play a major role in our health. In particular, we think that there is no causal association between some genes you have and the environment in which you live.

This is a standard example, for sure I read it somewhere, but I can't remember where. Probably from Hernán.

Gene
Environment

Figure 8.5: No causal path between genes and environment, so far so good.

However, apparently, you know from the literature that the particular genes and the environment you are studying are both separately cause of a certain disease. Therefore, the complete DAG should look like

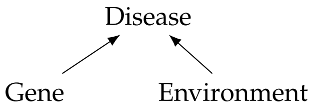


Figure 8.6: The disease as collider.

The two arrows are pointing towards one variable, they are colliding into it. Therefore, we call this node a *collider*. As usual, you assemble your cohort and you gather all your data. Would you expect an association between the genes and the environment, even when you know the causal paths in Figure 8.6? Put in another way, even though you know the causal mechanism, do you expect an association between genes and environment in the *aggregated* data? You probably suspect no association in the aggregated data. The flow of information is going from gene to disease and there it stops, it doesn't continue towards the environment. The fact that someone has this gene doesn't tell you anything about the environment.

So far so good, but imagine I want to segregate the data anyway, conditioning within the levels of the disease.

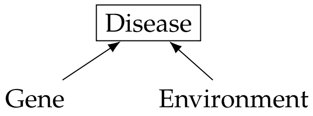


Figure 8.7: Conditioning within the levels of the collider.

So let's stratify, all the people with the disease on one side and all the people without on the other side. Consider, for example, just the group with the disease. Would you now expect an association between the genes and the environment? In this case, yes. You know already that in this group everyone has the disease, therefore something must have caused it, and it can only be either the gene or the environment. In fact, if you know that someone in this group doesn't have the gene, then you know that he was living in that specific environment that causes the disease. So in this group, knowing something about the genes tells you something about the environment. Conditioning on the collider, you opened the flow of information between gene and environment. But this is of course a troublesome association, because it is not due to a causal direct path between genes and environment. So if you are interested in estimating a causal effect, this path should remain closed. From this example, we learned some general things.

Again, this is just an example. Let's assume that our DAG in Figure 8.7 is the correct one.

- a collider doesn't create an association between the two variables causing it;
- conditioning within the levels of the collider opens the flow of association between the two variables causing it;

Once again, if we are interested in estimating the causal effect, we should not condition within the levels of the collider, as this would introduce bias.
As before, let's try to understand this better with a code.

The code is similar to the previous one, and still taken from here or here [7].

```
install.packages("Rlab")
library(Rlab)
N = 1000
Gene = rbern(N, prob = 0.5)
Environment = rbern(N, prob = 0.5)
Collider = rbern(N, ifelse(Gene + Environment > 0, 0.9, 0.2))
```

As you can see, the distributions of genes and environment are completely independent, while the value of the collider depends on the value of both gene and environment. We can have a look at the contingency table and the correlation coefficient

```
table(Gene, Environment)
cor(Gene, Environment)
```

The table is quite sparse, we don't see major patterns and the correlation between gene and environment is around 0.04, basically zero. What happens if we condition within the level of the collider? Let's calculate the correlations separately for the two possible values of the collider

```
cor(Gene[Collider == 0], Environment[Collider == 0])
cor(Gene[Collider == 1], Environment[Collider == 1])
```

The two correlation coefficients are respectively around 0.50 and −0.30. As you can see, segregating the data created an association which is not corresponding to any causal structure.

8.5 Final exercises

1. You want to study the effect of a certain drug to treat a disease. You run an observational study. You collect the data on the drug (whether someone is taking it or not) and on the disease (whether someone has it or not). Moreover, you know that the drug lowers the blood level of C-reactive protein (CRP) to treat the disease. The effect of the drug flows entirely through the CRP concentration (this is also known as *mediator*). You collect the data on CRP as well (high or low concentration). Would you condition within the levels of blood pressure to estimate the effect of the drug? Justify your answer.
2. Think about an experimental setup. Define the variables you want to study and the causal association you would like to estimate. Draw a DAG that portrays the causal model that you think it generates the phenomenon under study. Which variables you need to collect to have a complete picture? Which one would you condition on? Use DAGitty to double check your reasoning.

8.6 Further readings

Causal inference is a fundamental tool to help us removing bias from our studies. There is already a huge literature on causal inference and I can't really guide you through that, as I am not a real expert. I would like to recommend you some readings that I found very much valuable.

- There is a slim book [5] which is a nice introductory guide, written by the authors that developed the theory of causal inference that we discussed.;
- Pearl also wrote a popular science book [6] on his work, you might want to check it out as well;
- I would also have a look at the course by Hernán, you can find it for free here. The ideas we discussed in this chapter have all a strong mathematical foundation, but you don't need to deeply understand the math behind it in order to use the DAGs, and Hernán really helps you to improve your causal intuition;
- and of course, check the literary masterpiece [7];

Appendix

Codes A

In this Appendix are included all the scripts that I used to generate the synthetic datasets. Here is the code I used to create the YoloAB.csv file.

```
# Creating the variables for the dataset of the YOLO AB
install.packages("Rlab")
library(Rlab)

setwd("Here you should put the path to your working directory")

N = 40
ID = c(seq(1:N))

# Age variable is uniformly distributed between 23 and 35 years old
Age = runif(N, min = 23, max = 35)
# Rounding to get all integers
Age = ceiling(Age)

BMI = rlnorm(N, meanlog = 25, sdlog = 1.5)
BMI = log(BMI)

# Man is coded as zero while woman is coded as 1
Sex = rbern(N, prob = 0.5)

# Diastolic blood pressure, measured in mmHg
Diastolic_BP = 80 + rpois(N, lambda = 2)

# Glucose level, measured in mg/dL
Glucose_level = 90 + rpois(N, lambda = 2)

# Healthy is coded as zero while 1 means that the participant has at least
# one health-related problem
Health_status = rbern(N, prob = 0.2)

# The monthly salary is measured in SEK
Monthly_salary = 20000 + rlnorm(N, meanlog = 2, sdlog = 4)


# Putting all the variables in the same dataset
Dataset = data.frame(ID, Age, Sex, Diastolic_BP, Glucose_level, Health_status,
                     Monthly_salary)
```

Here is the code I used to create the Expedition.csv file.

```
    # Creating the variables for the dataset of expedition
install.packages("Rlab")
library(Rlab)

setwd("Here you should put the path to your working directory")

N = 1000
```

```
ID = c(seq(1:N))

# All the following variables are going to be normally distributed

# Age, measured in years

Age = rnorm(N, mean = 35, sd = 10)
Age = ceiling(Age)

# Height, measured in cm

Height = rnorm(N, mean = 170, sd = 10)

# Weight, measured in kg

Weight = rnorm(N, mean = 70, sd = 5)

# Brain volume, measured in cubic mm

Brain_Volume = rnorm(N, mean = 1490287, sd = 7370)

Database = data.frame(ID, Age, Height, Weight, Brain_Volume)
```

Here is the code I used to create the TheBook.csv file.

```
# Creating the variables for the dataset

setwd("The path to your directory")

N = 500
ID = c(seq(1:N))

# Age variable is uniformly distributed between 23 and 35 years
    old
Age = runif(N, min = 23, max = 35)
# Rounding to get all integers
Age = ceiling(Age)

# Weight variable is normally distributed
Weight = rnorm(N, 73, 4)
# Rounding
Weight = round(Weight, digits = 1)

# Height variable also normally distributed
Height = rnorm(N, 173, 5)
# Rounding to get all integers
Height = ceiling(Height)
# Converting into meters
Height = Height/100

BMI = (Weight)/(Height*Height)

# Education level of the parents
Education_level = runif(N, min = 0, max = 20)

# Born in or outside Uqbar
Uqbar = rbinom(N, 1, prob = 0.6)
# Uqbar = as.factor(Uqbar)

# Creating the error terms
epsilon = rnorm(N, 0, 10)

# Creating the IQ variable
IQ = 80 + 10*Uqbar + 3*Education_level + epsilon

# Belief in astrology

BeliefAstrology = runif(N, min = 0, max = 10)

# Depression score

DepressionScore = rnorm(N, 40, 2)

# Amygdalar volume
VolumeAmygdala = 8000 -100*DepressionScore - 10*Age + epsilon

# Narcissism score

NarcissismScore = 16 + 2*BeliefAstrology + epsilon


# Putting all the variables in the same dataset
Dataset = data.frame(ID, Age, Weight, Height, BMI, IQ, Uqbar,
                     BeliefAstrology, DepressionScore,
                     NarcissismScore, VolumeAmygdala)
```

Bibliography

Here are the references in citation order.

[1] Danielle Navarro. *Learning Statistics with R: A tutorial for psychology students and other beginners*. 2018 (cited on pages 23, 27, 32).

[2] Andrew Gelman, Jennifer Hill, and Aki Vehtari. *Regression and other stories*. Cambridge, 2021 (cited on page 41).

[3] Carl T. Bergstrom and Jevin D. West. *Calling Bullshit - The Art of Scepticism in a Data-Driven World*. Penguin Books (cited on pages 43, 45, 48, 49).

[4] David Spiegelhalter. *The Art of Statistics - Learning from Data*. Pelican Book (cited on pages 44, 49).

[5] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics - A Primer*. Wiley, 2016 (cited on pages 51, 57).

[6] Judea Pearl and Dana Mackenzie. *The book of why*. Penguin Books, 2018 (cited on pages 51, 57).

[7] Richard McElreath. *Statistical Rethinking - A Bayesian course with examples in R and Stan*. Second edition. CRC Press, 2020 (cited on pages 54, 56, 57).