

Gianluca Baio

# Introduction to statistical concepts

18 October, 2022



---

## Preliminaries

### What are these notes about and for?

The aim of these notes is **not** to provide a comprehensive and detailed introduction to *all* of statistical theory. Of course, that requires more than just a single document/book and you need to work your way through the many lectures you will attend during your MSc Programme in Health Economics and Decision Science. What these notes are meant to provide you with is a critical introduction to the most important concepts, that you will encounter specifically in STAT0015, STAT0016 and STAT0019. Of course, the last two modules are not compulsory, so you may not take them. Nevertheless, it is useful to have all these at hand. Arguably, statistical reasoning and analysis is central to all forms of what can be generically called “Health Economics” — both in terms of modelling for cost-effectiveness/utility analysis and when performing econometric modelling. It is then crucial that the concepts described in these notes are clear to you.

The structure of these notes guides you through the very basics of the philosophy underpinning the ideas of sampling and data collection. Suitable methods for summary and visualisation of the data are also presented in Chapter 1. Then Chapter 2 describes several statistical models that are commonly used in the applications you will encounter. These include Bernoulli and Binomial models to describe sampling variability in individual or aggregated binary data, Poisson models for counts, Normal distributions for continuous, symmetric phenomena and more specific distributions (e.g.  $t$ , Chi-squared and the Gamma-family), which are the basis for many of the procedures you will see during your Programme. Of course, the presentation is far from exhaustive and there are many more models you may be exposed to in specific modules.

Chapter 3 and Chapter 4 present the central tools of statistical inference — the methods of estimation and testing. These are presented while highlighting the fundamental distinctions among different approaches (e.g. Bayesian, Likelihood and Frequentist), which are often confused or conflated (especially the last two) into an integrated theory, which essentially does not exist. Again, the mathematical sophistication is kept to a low level — you do not need to read these notes to learn all about the technical issues. The point is rather to try and help you understand the basic principles and *why* things work the way they do, over and above *how*. Throughout the notes, there are some parts in which it is unavoidable to use mathematics to make the point; but you are not expected to learn proofs for theorems or anything similar — only to understand the process.

Finally, Chapter 5 discusses regression analysis, which is a general tool used in many areas of statistical modelling. Again, we dispense with the most complicated technical details and try to convey the most important ideas underpinning the development of linear and generalised linear models.

## Computer software

Throughout the notes, we demonstrate some of the computational problems using the freely available software R, which you can use on UCL machines. You can also download it on to your own machines from CRAN, i.e. the main repository from which all the relevant “packages”, as well as the main software is stored. This is available at <https://cran.r-project.org/index.html>.

Notice that you **do not** have to learn R when reading or studying these notes. Code and output are typeset in grey boxes, something like the following.

```
# Defines a variable
x=4
# Defines a vector
y=c(1,2,3,4)
# Computes a function of a given input
m=mean(y)
# Returns the output
m
```

[1] 2.5

You are not expected to have learnt it in preparation for the exam you will have to take before starting the Programme. The code is only presented to help you understand what is actually going on — and you can use it to replicate some of the analyses presented in the notes. Moreover, note that while attending the various modules, you will encounter several statistical software, including R, Stata, MatLab and perhaps others. While having their own different syntax and at times idiosyncrasy, if you learn to use one of these proper statistical programmes, then you *will* be able to switch to others — because their common trait is the possibility of *scripting* the workflow, using functions and packages.

This, in addition to their advanced computational engines, is what makes them more appropriate than commonly used spreadsheet calculators, e.g. MS Excel that are often used, particularly in the field of cost-effectiveness modelling. These are **not** ideal and have several shortcomings. So, while you will see them at times in the various modules, you are **not** encouraged to use them for “real” work — and we will see several applications of statistical modelling in the more appropriate software in STAT0015, STAT0016 and STAT0019.

## Scientific writing (hints for your dissertation)

These notes are written using [quarto](#), which can be used to combine plain text with advanced formatting and, crucially, R code. In this way, you can annotate and describe the whole analysis process in a single file, where you describe all the technical details as well as the general presentation of the problem. This is something you may consider for your final year dissertation.

## Symbols, notation, etc

Although, as mentioned above, we keep the mathematical sophistication to a bare minimum level, we do need to use specific symbols and terminology, for the sake of clarity. Generally speaking, statistical notation distinguishes mainly between **observed** or **observable** variables, which we indicate in upper-case Roman letters, e.g.  $Y$ ,  $W$ ,  $T$ ; and *unobservable* parameters, which are indicated using Greek letters, e.g.  $\theta$ ,  $\mu$ ,  $\sigma$ .

When data are *observed* (and thus their realised value is known to us), we usually indicate in lower-case Roman letters, e.g.  $y$ ,  $w$ ,  $t$ . When we consider a *vector* of variables or parameters, we typeset them in bold,

e.g.  $\mathbf{Y}$  indicates a vector of observable variables. We often describe this fully as  $\mathbf{Y} = (Y_1, \dots, Y_n)$ , which can be used to indicate that we have a vector of length  $n$ . We apply this to parameters too; for example, if a model is indexed by two parameters  $\mu$  and  $\sigma$ , we write that the parameters vector is  $\boldsymbol{\theta} = (\mu, \sigma)$ .

As mentioned above, a crucial part of statistical modelling is to associate variables with probability distributions, e.g. to describe uncertainty or sampling variability. We do this using the terminology

$$y \sim \text{Name of the distribution}(\text{Name of the parameters}),$$

where the symbol “ $\sim$ ” is read “is distributed as”, or more appropriately “is associated with a XXX distribution with parameters YYY”. Alternatively, we may write  $p(\text{Name of variable} \mid \text{Name of parameters})$  to indicate the probability distribution associated with a variable and indexed by some parameters. An example is

$$p(r \mid \theta, n) = \binom{n}{r} \theta^r (1 - \theta)^{(n-r)}$$

(see Section 2.1). The symbol “ $\mid$ ” (read “given” or “conditionally on”) is used to indicate that the argument to its left is the main variable of interest, while the argument(s) to its right are used as parameters or known values.

The notation  $\Pr(Y = y \mid \theta)$  indicates the probability that the variable  $Y$  takes on the value  $y$  — so this is a slightly different concept to the probability *distribution*  $p(y \mid \theta)$  — the former is a single value, while the latter is an entire distribution.

When we use sample data to estimate a model parameter, we may use the “hat” notation, e.g.  $\hat{\mu}$  (read “ $\mu$  hat”) can be used to indicate a function of the data  $\mathbf{Y}$  that we use to give our best guess as to what the underlying value for the parameter  $\mu$  is. This is not universal and some other terminology is possible to indicate an estimate for a parameter. When these are used, we will define them appropriately in the text.

Occasionally, we use “text blocks” to include specific bits of text and alert you to their importance. These look something like the following.

### ! Important

This is a block of text that you should read carefully. This is may be because the content is very important, or perhaps it is subtle and requires some thinking before you fully understand its meaning. Or may be it is a technical note, explaining some more advanced details — in which case, you do **not** need to learn all these (possibly mathematical) details by heart. As usual, only try and understand the deeper meaning of the text and maths included in the block.

### List of mathematical symbols

- $E[Y]$ : *expected value* of a variable  $Y$  (see Section 1.5.1). This indicates the mean of a variable and is often indicated with the symbol  $\mu$ .
- $\text{Var}[Y]$ : *variance* of a variable  $Y$  (see Section 1.6). This is often indicated with the symbol  $\sigma^2$ .
- $\sum_{i=1}^n y_i$ : the *sum* of  $n$  values  $y_1, \dots, y_n$ . Here  $y_i$  indicates one such generic value and the index  $i = 1, \dots, n$  (read: “ $i$  goes from 1 to  $n$ ”).
- $\prod_{i=1}^n y_i$ : the *product* of  $n$  values  $y_1, \dots, y_n$ .
- $\int_a^b f(x)dx$ : is the *integral* of the function  $f(x)$  of the variable  $x$ , ranging in the interval  $[a; b]$ . This is used to compute the *area under the curve* described by the function  $f(x)$ , for values of the  $x$ -axis

ranging in  $[a; b]$ . You will not encounter much of this, although this concept is often discussed in STAT0019.

- $\exp$  is the *exponential* function, with properties  $\exp(0) = 1$ ,  $\exp(1) = e = 2.7182818$ .
- $\log$  is the *logarithm* function, i.e. the inverse of the exponential function. This means that  $\log(\exp(x)) = x$ , i.e. if you apply the  $\log$  to  $\exp$  you essentially cancel out these two functions and are left with the argument to the inner function ( $\exp$ ). The  $\log$  function only applies to **positive** numbers; also  $\log(1) = 0$  and  $\log(e) = 1$ .
- $n!$ : the *factorial* function indicates the product  $n(n-1)(n-2)\cdots 1$ , for any positive number  $n$ . This is used in the definition of some probability distributions, including the Binomial (see Section 2.1, the Student's t Section 2.4)) and the Gamma family of distributions (see Section 2.5).
- $x \in [a, b]$ : read “ $x$  is in the interval  $[a; b]$ ”. The symbol  $\in$  indicates group (or set, interval) membership.
- $\rightarrow$ : read “tends to” or “approaches”. This is used in expression such as  $n \rightarrow \infty$  (read “ $n$  approaches infinity”).
- $f'(x)$ : read “ $f$  prime of  $x$ ”. This indicates the *first derivative* of a function  $f$ . This measures the changes in the value of the function for infinitely small changes of the argument  $x$ . Derivatives are crucial concepts in differentiation and calculus and are used to determine maxima or minima of a given function. This is technically the notation introduced by Giuseppe Luigi Lagrangia, an Italian mathematician (actually popular with the French version of his surname, Lagrange), who developed much of the early versions of calculus.
- $f''(x)$ : read the second derivative of  $f$ . This is computed as the first derivative of a first derivative, so  $f''(x) = f'(f'(x))$ .
- $\min_a f(a)$ : the *minimum* of a given function with respect to its argument  $a$ . In other words, the value  $a$  is the one in correspondence of which the function  $f(\cdot)$  reaches its minimum. The obvious counterpart is  $\max_a f(a)$ .

## Basic concepts

Broadly speaking, the objective of statistical analysis is to produce “some summary” of the available data. Sometimes, it is (at least *theoretically*) possible to deal with an entire **population** of observable quantities. We often refer to these quantities as **variables** that may take any one of a specified set of values, for a given individual. Examples are age (of persons), income (of households), socio-economic class (of workers). **Data** are the set of values of one or more variables recorded on one or more individuals or items.

An example of the theoretical construct underpinning the concept of population is represented by a **census**. In such cases, *every* single unit that is present in the population (and is thus of interest to the research question) is actually measured. We can use these measurements to summarise the information provided by the data using what are often called *descriptive* statistics. Notice that, particularly in this idealised case where we have observed everyone in the population, it is important to be able to fully characterise the underlying measurements with easily-interpretable quantities (as opposed to looking at each single measurement).

In addition, as mentioned above, the idea of a “population” is often elusive: in the case of the census, everybody living in a given country is supposed to fill in the questionnaire and thus provide extensive information about themselves. So, surely we collect information about the whole population. Or do we? . . . The problem is that populations are intrinsically *dynamic* — people die and new babies are born. Similarly, people get married or divorced (thus changing their marital status).

Consequently, the very concept of “population” and the idea that we may be able to fully observe everything is pretty much wishful thinking. Moreover, even if we could think of an entire observed population, it may still be impossible to obtain data on each and every individual/unit. There are several reasons for this:

- *Economic reasons*: to measure as many units as there are in the population may cost too much money, thus limiting the usefulness of the information collected;
- *Accuracy reasons*: it may be better to measure **very** precisely a limited number of units, than just investing the same amount of resources to collect less precise information on a larger number of individuals;
- *Physical reasons*: sometimes, the very act of measurement destroys the unit. For example, consider the case in which you want to know the life-time of a population of 100 bulbs. You may light all of them and measure how long it takes before they all burn. And you would have very precise information about this quantity. But at the end of the process, you would not have any bulb left to use. . .

For these reasons, invariably we rely on information obtained from a **sample** of units, drawn from the population of interest, for which a measurement is indeed available. We can use these units to make **inference** about the (theoretical) underlying population.

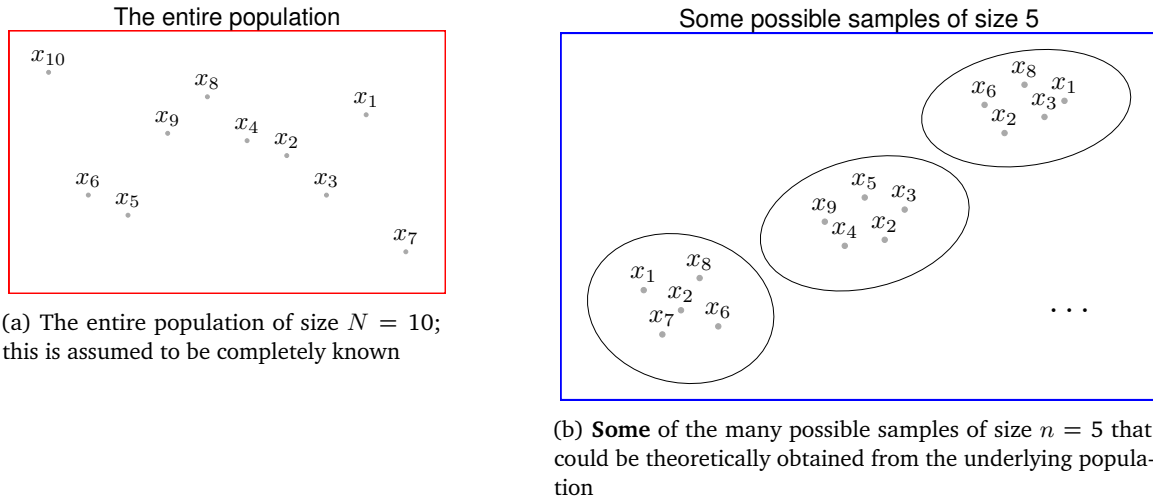


Figure 1.1: A schematic representation of the distinction between probability calculus and statistics. The data-generating process goes “left-to-right”, from the alleged population presented in panel (a), to the (many) possible samples that can randomise from it, through a specified probabilistic mechanism, some of which are depicted in panel (b)

Figure 1.1 shows a schematic representation of the process moving from a *theoretical* population (made by  $N = 10$  units) to **some** potential samples of size  $n = 5$ . Typically, we indicate the *population parameters* (e.g., the “true” mean and standard deviation, that we could compute if we could access the whole population — more on the definition of these quantities later in this chapter) using Greek letters, e.g.,  $\mu$  or  $\sigma$ . Ideally, we would like to learn about these quantities — but as mentioned above, we really cannot observe the whole population (which, again, might not even exist as such!). Thus, we rely on the *sample statistics*, which we indicate using Roman letters, e.g.,  $\bar{x}$  for the sample mean and  $s_x$  for the sample standard deviation (see Section 1.5 for more details on the definition, meaning and use of these quantities).

**Example 1.1** (Many samples?). Consider the following *idealised* situation. You are some kind of God and know all there is to know about a specific population of individuals. You are also a very lucky God, because this population of interest is relatively small and only formed by  $N = 25$  units. The “true” data showing the weight (in Kg) for each individual are presented in Table 1.1, which we indicate as  $y_i$ .

Table 1.1: Whole population data on weight

73.134	54.311	82.485	68.569	47.659
75.259	82.401	52.116	61.638	31.907
69.135	56.550	66.988	85.416	69.868
42.871	93.900	39.348	54.570	49.476
67.182	77.861	20.851	72.541	92.979

Some basic properties of the whole population are the following:

- Total size:  $N = 25$ ;
- Mean weight:



$$\mu = \sum_{i=1}^N \frac{y_i}{N} = 63.561.$$

This is a measure of “central tendency” (more on this later, in Section 1.5);

- Standard deviation:

$$\sigma = \sqrt{\sum_{i=1}^N \frac{(y_i - \mu)^2}{N}} = 18.544.$$

This is a measure of variation around the mean (more on this later, in Section 1.5).

For some reason, you decide that you do not want to share the whole population data — but only make available a smaller sample, randomly selected from it. Of course, there are many possible ways of sampling from this population (and of course, if you had a much bigger population, there would be even more ways — again, more on this later in Chapter 2).

Say you consider three possible samples, as given by the table below. Each sample is determined by the individual Id (a number from 1 to 25) — these are associated with the whole population values in Table 1.1 reading along the rows (so that the first column indicates Ids 1, . . . ,5; the second column indicates Ids 6, . . . ,10; etc.).

Table 1.2: Three possible samples from the entire population

(a)	(b)	(c)
Selected Id Weight	Selected Id Weight	Selected Id Weight
1 73.134	13 66.988	2 75.259
2 75.259	14 39.348	9 93.900
3 69.135	15 20.851	17 61.638
4 42.871	16 68.569	1 73.134
5 67.182	17 61.638	20 72.541
6 54.311	18 85.416	23 69.868
7 82.401	19 54.570	13 66.988
8 56.550	20 72.541	
	21 47.659	

If we consider the equivalent summary statistics to the whole population, we get the following results:

- Sample 1:
  - Sample size  $n_1 = 8$ ;
  - Sample mean

$$\bar{y}_1 = \sum_{j=1}^{n_1} \frac{y_j}{n_1} = 65.105;$$

- Sample standard deviation

$$s_1 = \sqrt{\sum_{j=1}^{n_1} \frac{(y_j - \bar{y}_1)^2}{(n_1 - 1)}} = 12.936.$$

(The reason why the denominator in the sample standard deviation becomes  $(n_1 - 1)$  will be explored in Chapter 3.

- Sample 2:
  - Sample size  $n_2 = 9$ ;

- Sample mean

$$\bar{y}_2 = \sum_{j=1}^{n_2} \frac{y_j}{n_2} = 57.509;$$

- Sample standard deviation

$$s_2 = \sqrt{\sum_{j=1}^{n_2} \frac{(y_j - \bar{y}_2)^2}{(n_2 - 1)}} = 19.408.$$

- Sample 3:

- Sample size  $n_3 = 7$ ;
- Sample mean

$$\bar{y}_3 = \sum_{j=1}^{n_3} \frac{y_j}{n_3} = 73.333;$$

- Sample standard deviation

$$s_3 = \sqrt{\sum_{j=1}^{n_3} \frac{(y_j - \bar{y}_3)^2}{(n_3 - 1)}} = 10.136.$$

Which sample would you say is the “best” one?

Now: samples number 1 and 2 look a bit strange because they have selected consecutive units (1 to 8 for sample 1; and 13 to 21 for sample 2). This is not necessarily suspicious or wrong — but what if people were numbered according to the household in which they live? This would mean that, probably, consecutive Ids are more likely to indicate people living in the same household who are thus potentially more correlated (e.g., parents and children). This may reduce the **representativeness** of the sample with respect to the underlying population. Conversely, the third sample presents selected Ids that look more *random* and so, arguably, may be deemed to be more reliable. Interestingly, despite this desirable property, sample 3 is the one for which the summary statistics differ the most from the underlying population! (More on this later in Chapter 2).

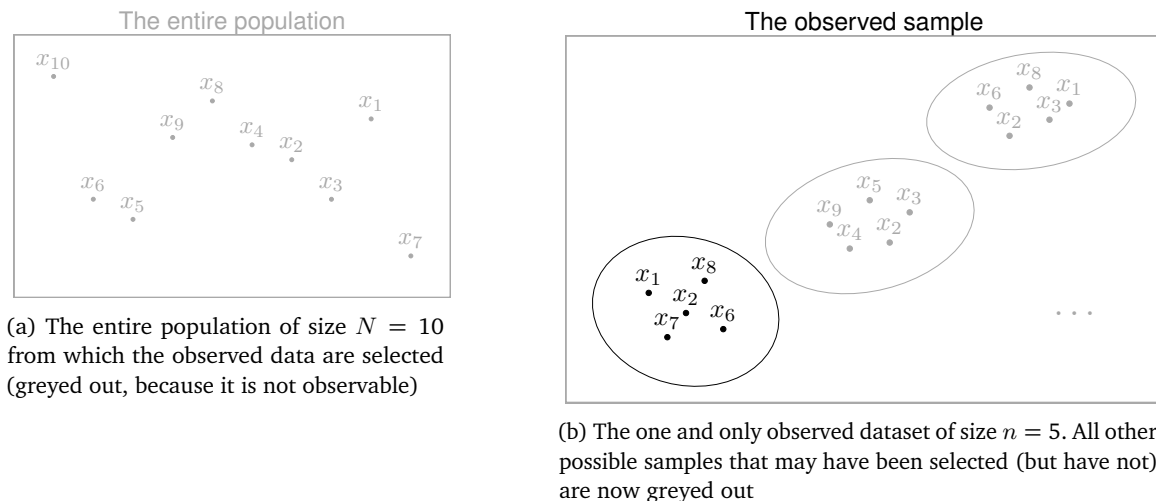


Figure 1.2: A schematic description of the *observed* data, with the now explicitly unknown underlying data-generating process greyed out. This time the process goes “right-to-left”, using the one and only available dataset depicted in panel (b), to learn about the characteristics of the underlying population shown in panel (a)

In addition, in real life (i.e. when we do Statistics), we cannot access all possible samples that could have been drawn from a given population. For example, in the trivial case above, the true population is not really accessible to us (and thus it is greyed out in Figure 1.2). We only have one such sample (and again all but the one in the left-bottom part of the right-hand side panel of Figure 1.2 are also greyed out).

Incidentally, there are in fact 252 different ways of picking at random 5 units out of the population made by 10 — but again, all but one have not been drawn. We want to use the information in the actually observed sample (e.g. the sample mean and standard deviation) to **infer** about the underlying population parameters. That is what Statistics is all about.

## 1.1 How to obtain a random sample

In general, you can think of this process as extracting numbered balls from an urn — a bit like they do when they call numbers in games such as Tombola or Bingo. If you extract the ball with number “14” on it, then you are selecting into your sample the 14-th unit from a complete list of population members.

In reality, we are likely to use a computer to simulate “pseudo-random” numbers. For example, we can use the freely available software R to sample 4 numbers from the set  $(1, 2, \dots, 20)$ , using the following code.

```
# Simulate 4 numbers from the set of numbers from 1 to 20, without replacement
sample(x=1:20,size=4,replace=FALSE)
```

```
[1] 10 20 14 13
```

The resulting values shown here can be used to in fact select the corresponding units in a list of peoples or items of interest. In this case, the option `replace=FALSE` instructs R to sample *without* replacement. This means that if a unit is selected then it is taken out from the list of units that can be selected at a later stage.

## 1.2 Types of data

Once we have identified a suitable procedure to sample from the underlying population, we are then confronted with the actual data to analyse. As mentioned above, data are made by variables, which may have differences in nature. The broadest categorisation in terms of types of data is probably the following.

### 1. Qualitative (non numerical):

- *Categorical*: no actual measurement is made, just a qualitative judgment e.g., sex, hair colour. The observations are said to fall into categories.
- *Ordinal*: there is a natural ordering of the categories such as degree of severity of a disease (mild, moderate, severe), occupational group (professional, skilled manual, unskilled manual).

### 2. Quantitative (numerical):

- *Discrete*: can only take one of a discrete set of values (e.g., the number of children in a family, the number of bankruptcies in a year, the number of industrial accidents in a month).
- *Continuous*: can in principle take any value in a continuous range, such as a person’s height or the time for some event to happen. In practice, all real observations are discrete because they are recorded with finite accuracy (e.g., time to the nearest minute, income to the nearest pound). But if they are recorded sufficiently accurately they are regarded as continuous.

## 1.3 Numerical data summaries

Qualitative and discrete data can be easily summarised using a *frequency table*.

**Example 1.2** (Accident data). Data are collected routinely to describe the type of accident that people have in a given time period. These are grouped according to some broad categories. The *categories* are the types of accident. The number of children dying from each type of accident is the *frequency* of that category. The *relative frequency* or *proportion* of children dying from each type of accident is the frequency divided by the total number of deaths. Multiplying the relative frequencies by 100 gives the *percentages* (i.e., the relative frequencies per 100 cases), as shown in Table 1.3.

In general, it is a good idea to sort out the data in terms of the frequencies, for ease of presentation.

Table 1.3: The number of fatal accidents to children under 15 in the UK during 1987 (source: Action on Accidents, produced by the National Association of Health Authorities and the Royal Society for the Prevention of Accidents)

	Number of children	Percentage
Pedestrians (road; P)	260	30.88
Burns, fires (mainly home; B)	119	14.13
Vehicle occupants (road; V)	96	11.40
Cyclists (road; R)	73	8.67
Drownings (home and elsewhere; D)	63	7.48
Choking on food; C)	50	5.94
Falls (home and elsewhere; F)	40	4.75
Suffocation (home; S)	34	4.04
Other (O)	107	12.71

When we deal with continuous data, frequency tables are not so straightforward, because the number of categories is much larger (and, theoretically, infinite).

**Example 1.3** (Heights). The following data considers individual measurements of span (in inches) of 140 men.

Table 1.4: Individual measurements of span (in inches)

68.2	67.0	73.1	70.3	70.9	76.3	65.5	72.4	65.8	70.7	65.1	66.5	67.5	64.4
64.8	72.7	71.9	73.9	68.3	66.1	69.9	68.5	72.5	67.5	72.1	71.6	65.6	65.7
64.2	71.6	73.4	70.8	71.5	76.0	68.0	65.1	70.1	68.4	71.3	73.9	70.3	72.4
73.9	72.3	67.6	70.2	66.6	75.1	72.2	65.6	72.2	67.0	67.1	70.8	70.7	68.2
69.5	70.0	73.0	65.0	70.0	68.2	69.8	74.8	73.8	68.3	65.4	66.5	67.3	73.2
70.8	71.0	69.9	75.4	72.2	68.6	65.5	68.0	66.3	67.6	68.0	69.8	65.8	68.0
68.4	71.0	71.8	72.3	67.6	69.4	73.2	70.3	70.3	63.9	70.3	73.9	66.0	68.4
72.7	67.4	64.3	71.1	71.2	69.1	64.7	73.2	74.0	66.5	66.7	66.7	72.2	61.5
72.6	68.3	71.5	65.5	70.5	70.7	67.5	74.2	69.4	67.1	70.8	67.8	70.8	66.9
67.5	66.8	70.4	70.6	66.5	70.5	68.2	74.7	69.7	66.9	74.0	67.9	72.1	61.3

One way of getting round the large number of different values in the dataset is perhaps to consider *grouped* frequency tables, such as the that in Table 1.5. This can be obtained using the following process.

1. Calculate the range of the data i.e., the largest value minus the smallest value.
2. Divide the range up into groups. Aim at having between 5 and 15 groups.

3. Calculate the frequency of each group.

Table 1.5: Grouped frequency table for the measurements of span

	Frequency	Relative frequency
61-62.4	2	0.014
62.4-63.9	1	0.007
63.9-65.4	9	0.064
65.4-66.9	21	0.150
66.9-68.4	29	0.207
68.4-69.9	11	0.079
69.9-71.4	27	0.193
71.4-72.9	20	0.143
72.9-74.4	14	0.100
74.4-75.9	4	0.029
75.9-77.4	2	0.014

The resulting table is certainly more informative than the fill list of the original values. But we are losing information by grouping the data — for instance, we only know that 5 individuals are in the range 61-62.4. However, we have lost track of the actual value for the span measurements of these 5 individuals.

## 1.4 Graphical summaries of data

Often, it is preferable to summarise data using pictorial representations. Different types of data are best displayed using different graphs. For example, when considering qualitative data, a good choice is given by a *barplot*. Using software such as R a barplot can be easily obtained using the command `barplot(...)` where `...` is the name of the object containing the frequencies we want to depict (the barplot can be customised to produce nicer graphs than the default version — but that is for another day...).

This type of display also works for discrete data, when we can essentially plot the frequency (either absolute or relative) for each of the possible discrete values.

When we deal with continuous data, it is basically impossible to draw barplots, because it is extremely unlikely (more on this later, in Chapter 2) that two distinct observations are measured to be the **exact** continuous value. In this case, we can use a **histogram**. This is a graph of the information in a grouped frequency table. For each group, a rectangle is drawn with base equal to the group width and area proportional to the frequency for that group. In R can be obtained by the command `hist(...)` — note that, in general you can use the command `help(...)` where `...` is the name of a R function (e.g., `hist`) to visualise extensive help on the function inputs and outputs.

Figure 1.4 shows two different histograms for the height data: the one on the left panel shows the absolute frequencies, while the right panels shows the relative frequencies.

Notice that bar are drawn in correspondence with *grouped* values of heights. In fact, by default, R splits the range of data in intervals. Usually the groups are of equal width, as in the above example, and the height of the rectangle is then also proportional to the frequency. It is common for the vertical axis to be called “frequency” in this situation, which really means “frequency per group width” (sometimes also called “frequency density”).

We could of course modify this and present different depictions of the data according to different groupings along the  $x$ -axis. For instance, we could produce a histogram showing the frequency *per 1.4 inches*, as depicted in Figure 1.5.

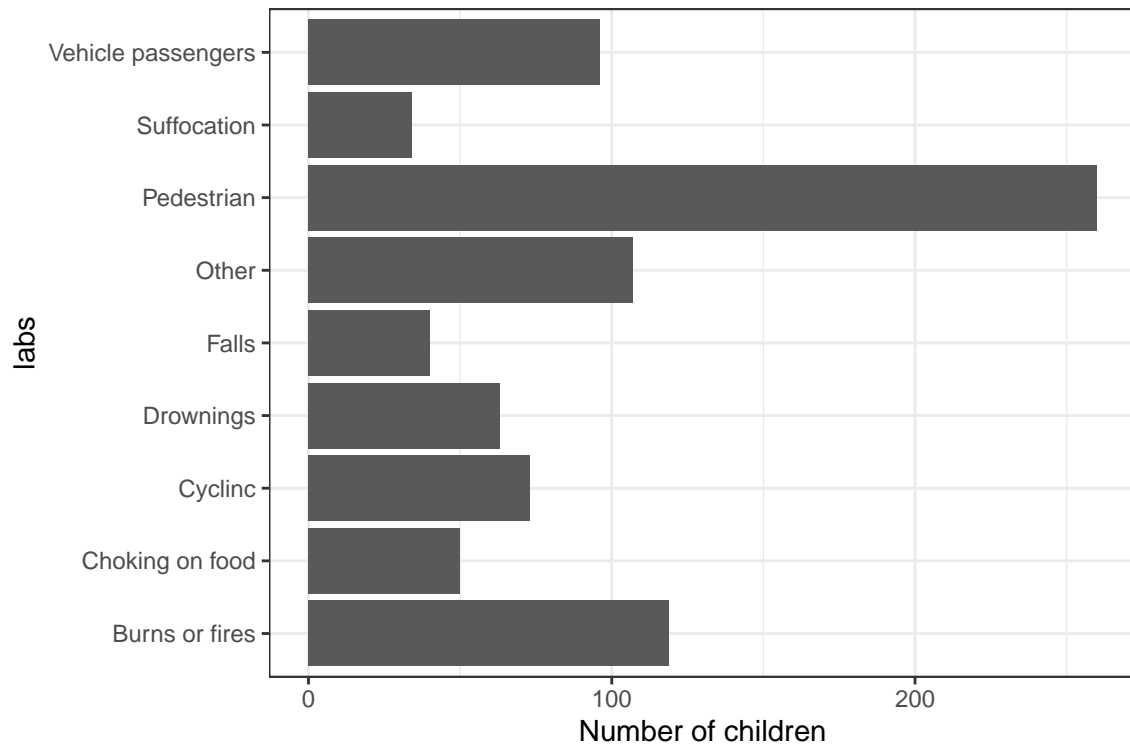


Figure 1.3: Barplot for the distribution of the accident data

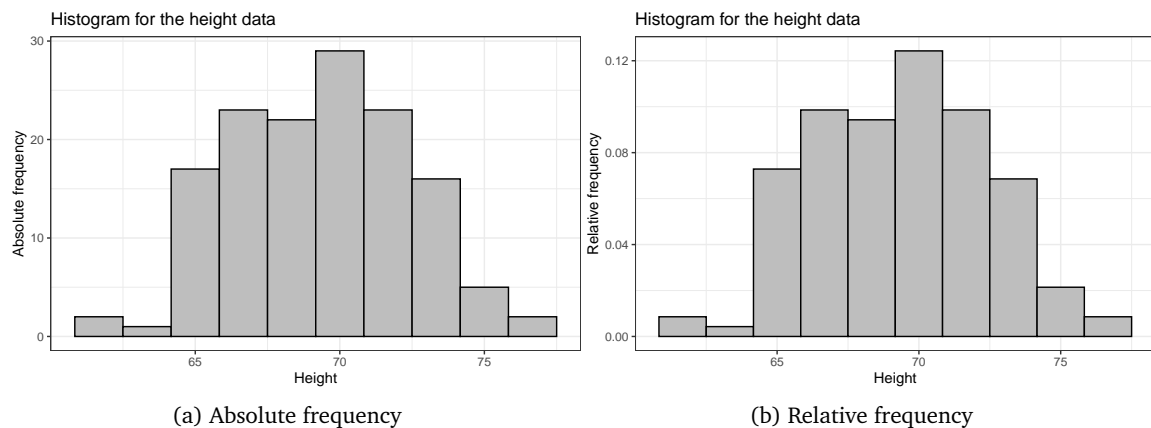


Figure 1.4: Histogram for the distribution of the height data

## 1.5 Summary statistics

In addition to graphical displays it is often useful to have numerical summary statistics that attempt to condense the important features of the data into a few numbers. It is helpful to try and distinguish between *population* and *sample* summary statistics.

### 1.5.1 Measures of Location (or Level)

**Mean.** This is sometimes referred to as the arithmetic mean, to distinguish it from other types of mean, such as geometric mean and harmonic mean. It is defined as

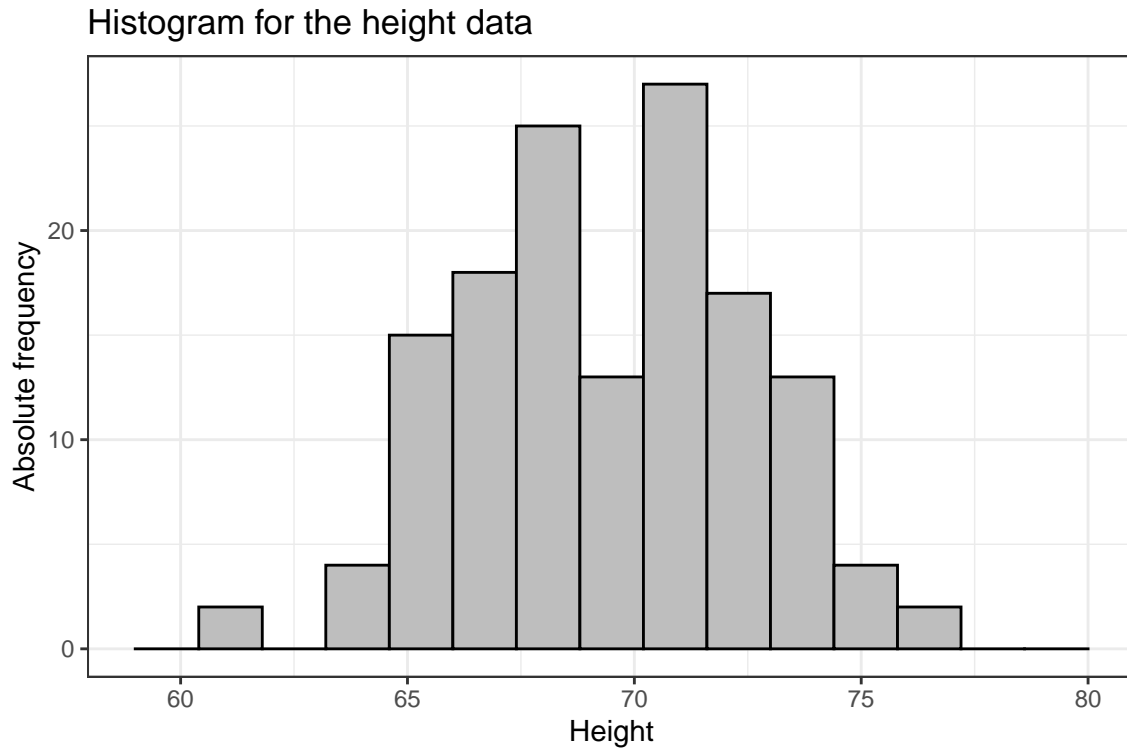


Figure 1.5: Histogram for the distribution of the height data per 1.4 inches

$$\text{mean} = \frac{\text{the sum of all observations}}{\text{the total number of observations}}$$

This is often written in the mathematical notation  $\bar{x}$  which you will find in textbooks and on calculators. Using Example 1.3 to help explain the notation:

- $n$  is the number of observations in the sample, in this case  $n = 140$ .
- $y_1$  is the height of the first individual in the sample, i.e.  $y_1 = 68.2$ .
- $y_2$  is the height of the first individual in the sample, i.e.  $y_2 = 64.8$ .
- $\sum y_i$  is the sum of all the  $\mathbf{y} = (y_1, \dots, y_n)$  values, in this case  $\sum y_i = 9721.8$ .
- $\bar{y}$  is the mean of the sample

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{68.2 + 64.8 + \dots + 61.3}{140} = 69.441.$$

The mean has some properties that it is useful to understand:

1. Imagine trying to balance the data on the end of a pencil. The point on the scale where the figure balances exactly is the *mean*. This helps us understand why if the data are symmetric, the mean is in the middle; and it tells us intuitively where the mean must be if the data are not symmetric.
2. Suppose that you subtract the mean from each data value. Then the resulting differences (sometimes called **residuals**) must **add to zero**. That is

$$(y_1 - \bar{y}) + (y_2 - \bar{y}) + \dots + (y_n - \bar{y}) = \sum_{i=1}^n y_i - n\bar{y} = n\bar{y} - n\bar{y} = 0.$$

**Median.** The median of a set of numbers is the value below which (or equivalently above which) half of them lie. It is also known as the 50-percentile point. To find the median of  $n$  observations, first put the observations in increasing order. The median is then given by:

- the  $\frac{n+1}{2}$ -th observation if  $n$  is odd;
- the mean of the  $\frac{n}{2}$ -th and the  $\frac{n}{2} + 1$ -th observations if  $n$  is even.

For the data in Example 1.3, we can use the R command `sort` to write the observations in increasing order as follows.

```
[1] 61.3 61.5 63.9 64.2 64.3 64.4 64.7 64.8 65.0 65.1 65.1 65.4 65.5 65.5 65.5
[16] 65.6 65.6 65.7 65.8 65.8 66.0 66.1 66.3 66.5 66.5 66.5 66.5 66.6 66.7 66.7
[31] 66.8 66.9 66.9 67.0 67.0 67.1 67.1 67.3 67.4 67.5 67.5 67.5 67.5 67.6 67.6
[46] 67.6 67.8 67.9 68.0 68.0 68.0 68.0 68.2 68.2 68.2 68.2 68.3 68.3 68.3 68.4
[61] 68.4 68.4 68.5 68.6 69.1 69.4 69.4 69.5 69.7 69.8 69.8 69.9 69.9 70.0 70.0
[76] 70.1 70.2 70.3 70.3 70.3 70.3 70.3 70.4 70.5 70.5 70.6 70.7 70.7 70.7 70.8
[91] 70.8 70.8 70.8 70.8 70.9 71.0 71.0 71.1 71.2 71.3 71.5 71.5 71.6 71.6 71.8
[106] 71.9 72.1 72.1 72.2 72.2 72.2 72.2 72.3 72.3 72.4 72.4 72.5 72.6 72.7 72.7
[121] 73.0 73.1 73.2 73.2 73.2 73.4 73.8 73.9 73.9 73.9 73.9 74.0 74.0 74.2 74.7
[136] 74.8 75.1 75.4 76.0 76.3
```

(notice that the numbers in square brackets in the left hand side of the display indicate the sequential value in the series of data. For example, the notation [43] indicates that the value 67.5 is the 43-th in the series).

As  $n = 140$  is even, the median is the mean between the  $\frac{n}{2}$ -th (70-th) and the  $\frac{n}{2} + 1$ -th (71-th) observations, i.e.

$$\text{med}(y) = \frac{y_{70} + y_{71}}{2} = \frac{(69.8 + 69.8)}{2} = \frac{139.6}{2} = 69.8.$$

**Quartiles (and other quantiles).** In the same way as for the median, we may calculate the value below which some specified fraction of the observations lie. The **lower quartile**  $q_L$  is the value below which one quarter of the observations lie and the **upper quartile**  $q_U$  is the value below which three quarters of the observations lie. The lower and upper quartiles are also known as the 25 and 75 **percentiles**. Different text books may use slightly different definitions of sample quartiles. Here is a standard one: as when finding the median, first put all the  $n$  observations in increasing order. Then:

- If  $\frac{n}{4}$  is not a whole number then calculate  $a$ , the next whole number larger than  $\frac{n}{4}$ , and  $b$ , the next whole number larger than  $\frac{3n}{4}$ . The lower quartile is the  $a$ th observation and the upper quartile is the  $b$ th observation.
- If  $\frac{n}{4}$  is a whole number then the lower quartile is the mean of the  $\frac{n}{4}$ -th and  $(\frac{n}{4} + 1)$ -th observations and the upper quartile is the mean of the  $\frac{3n}{4}$ -th and  $(\frac{3n}{4} + 1)$ -th.

In Example 1.3,  $n = 140$ , so  $\frac{n}{4} = \frac{140}{4} = 35$ , which is a whole number. So the lower quartile  $q_L$  can be computed using the rule in point ii. above.

In R we can easily compute all these summaries using built-in functions, for example as in the following code.

```
# Mean
mean(height)

[1] 69.44143

# Median
median(height)

[1] 69.8
```



```
# 0.25 Quantile (=lower quartile)
quantile(height,0.25)

25%
67.075

# 0.75 Quantile (=upper quartile)
quantile(height,0.75)

75%
71.825
```

## 1.6 Measures of spread

**Range.** The range is the largest observation minus the smallest observation. In Example 1.3, the range is  $76.3 - 61.3 = 15$ .

**Interquartile Range.** The range has the disadvantage that it may be greatly affected by extreme values that are a large distance away from the main body of the data, so that it may not give an informative measure of the spread of most of the data. A more stable measure is the *interquartile* range, which is the range of the middle half of the data. Thus

$$\text{interquartile range} = \text{upper quartile} - \text{lower quartile} = q_U - q_L.$$

For the data in Example 1.3 the interquartile range is  $71.825 - 67.075 = 4.75$ . In R we can also use the built-in function `IQR(...)`, where `...` is the name of the vector of data for which we want to compute the interquartile range, to make the same computation.

**Variance and Standard deviation.** If we consider the population at large, then the **variance** is defined as

$$\text{Population variance} = \sigma^2 = \sum_{i=1}^N \frac{(y_i - \mu)^2}{N}.$$

This quantity is the sum of squares of the residuals (i.e. the difference between each observation and the overall sample), divided by the total number of observations  $N$ . The units of the variance are the square of the units of the original data, so its numerical value is not particularly useful as a measure of spread.

Thus, we usually consider the square root of the variance

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^N \frac{(y_i - \mu)^2}{N}},$$

which is called the **standard deviation**. This directly reflects how each observation deviates from the central tendency as represented by the mean — notice that  $\sigma$  is defined on the same scale as the original data  $y_i$  and their mean and as such is a more naturally interpretable quantity. In general, large values of the standard deviation indicate that the population is very variable — there are very large residuals, i.e. some of the units have values that deviate substantially from the mean. The two quantities  $\sigma$  and  $\sigma^2$  are population *parameters* — they characterise the overall population. But they are not directly measurable when we consider a small(er) sample. The sample counterparts are defined in a fairly similar way to the population parameters. The **sample variance** is

$$\text{Sample variance} = s^2 = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \dots + (y_n - \bar{y})^2}{n - 1} = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}$$

(the **sample standard deviation**, is obviously defined as the square root of the sample variance).

The only two real differences between the population and sample definitions are that

1. The sample statistics are computed using the  $n$  available data points, while the population (theoretical) parameters are computed using the whole  $N$  data points that make it up.
2. The sample statistics are scaled by  $n - 1$ , instead of by  $n$ . The reason for this is that the standard deviation (and the variance) is a function of the mean — whether you consider the population or the sample version, the numerator is made by the difference between each observation and the overall mean. And because the mean is

$$\bar{y} = \sum_{i=1}^n \frac{y_i}{n} = \frac{(y_1 + \dots + y_n)}{n}$$

it follows that

$$y_n = n\bar{y} - (y_1 + \dots + y_{n-1}). \quad (1.1)$$

Thus, if you know the mean and the first  $(n - 1)$  observations, the  $n$ -th one is automatically determined by Equation 1.1. For this reason, when we compute the standard deviation or the variance, we only have  $(n - 1)$  **degrees of freedom** (i.e. the total number of parameters that are free to vary with no restrictions); in other words, you have  $n$  data points and are trying to estimate the mean  $\mu$  using  $\bar{y}$  and the standard deviation  $\sigma$  using the sample counterpart  $s$ . But  $\bar{y}$  and  $s$  cannot both vary independently — once you have estimated  $\bar{y}$  from the  $n$  data points,  $s$  cannot vary at leisure any more. And for this, we re-scale the sample quantities by  $(n - 1)$ .

## Statistical distributions: working with probability calculus

The objective of this chapter is to present a brief introduction to the use of some *probability distributions* to model sampling variability in observed data. For now, we consider a situation very similar to the one discussed in the context of Figure 1.1: we assume that there is a **data generating process** (DGP), i.e. a way in which data can arise and become available to us. This DGP determines the way in which some units are actually sampled from the theoretical target population.

As mentioned in Chapter 1, there are many different ways in which we can obtain a sample of  $n$  individuals out of the  $N \gg n$  that make up the whole population. In a nutshell (and somewhat making a more trivial argument than it really is), the ideas we explore in this chapter assume that we can safely assume a DGP characterised by a probability distribution; often, we use the phrase “the variable  $y$  has a XXX distribution”. While this terminology is almost ubiquitous in Statistics and Probability Calculus, it is in fact slightly misleading. What we really mean is a rather handy shortcut for the much more verbose (and correct!) phrasing

“We can associate the observed data with a XXX distribution to describe our level of uncertainty, e.g. on the sampling process that has determined the actual observation we have made (or we will make in the future).”

Of course, it would be very impractical to always use this mouthful sentence — and practically nobody does. **But:** it is important to understand that probability distributions or DGPs are not physical properties of the data — that is why the data cannot *have* a probability distribution. Rather, they are mathematical idealised concepts that we use to represent complex phenomena in a convenient way.

In general, the DGP will depend on a set of **parameters** (which in general we indicate with Greek letters, e.g.  $\theta$ ,  $\mu$ ,  $\sigma$ ,  $\lambda$ , etc). The distribution is characterised by a *mathematical function*, something like

$$p(y | \theta_1, \theta_2, \theta_3) = \theta_1 \frac{\theta_2}{\sqrt{\theta_3}} \exp\left(\frac{(y - \theta_2)^2}{\log(\theta_3)}\right)$$

(the actual form of this function is irrelevant here and this specific one is only chosen to make a point!). In this case, the parameters are  $\theta = (\theta_1, \theta_2, \theta_3)$  and, given a specific value for them, we determine a certain value of the underlying probability distribution for the observable outcomes,  $y$ . In other words, we can use a probability distribution to describe the chance that a specific outcome arises, given a set values of the parameters.

For example, for the fake distribution above, if  $\theta = (2, 1, 0.6)$ , then  $p(y) = 2.582 \exp\left(\frac{(y-1)^2}{-0.5108}\right)$ . The graph in Figure 2.1 shows a graphical representation of the distribution  $p(y)$  for the set values of  $\theta$ . In this case, we are implying that the probability is distributed around a central part (more or less in correspondence of the value 1 along the  $x$ -axis) and that values increasingly further away are associated with very small probability weight (and thus are deemed unlikely by this model).

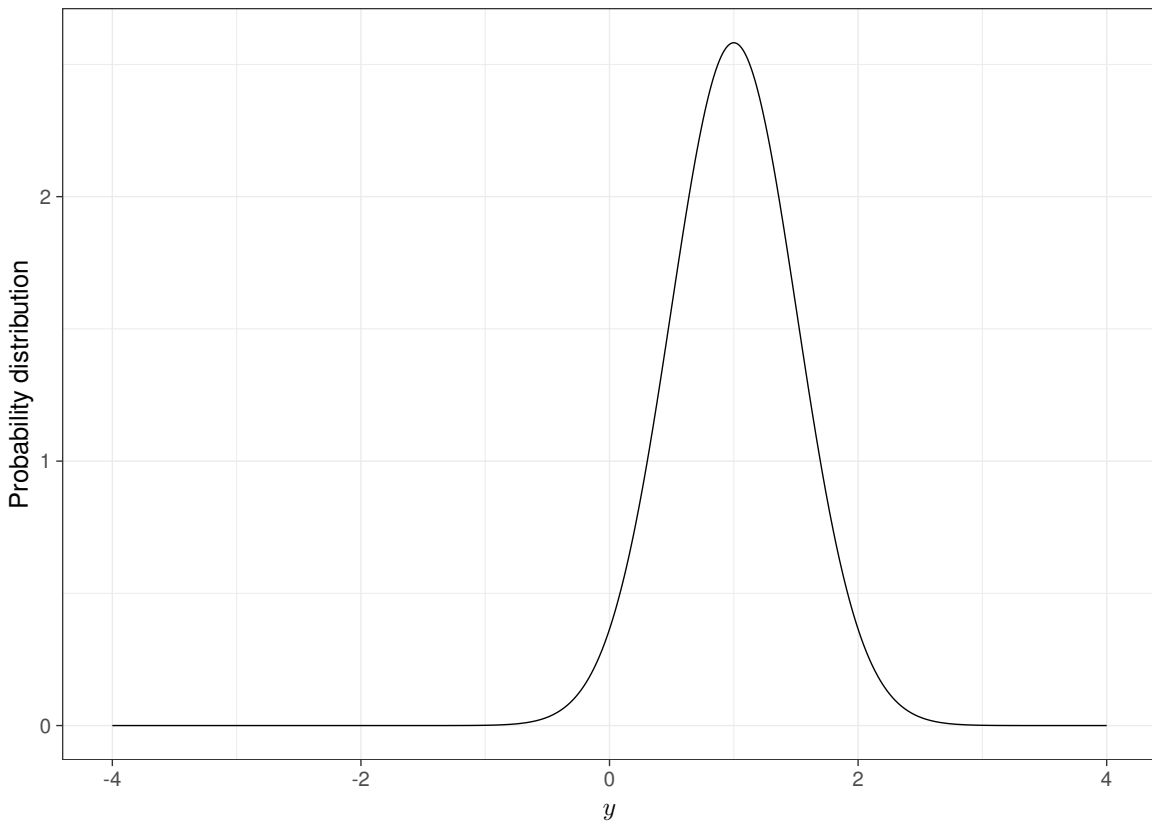


Figure 2.1: A graphical example of a probability distribution for a given set of value for the parameters  $\theta$

Of course, the assumption that we *know* the value of the parameters is certainly a big one and, in general, we are not in a position of having such certainty. And that is the point made in Figure 1.2 — instead of going from left to right, in reality we will try and go from right (i.e. using the one and only sample that we have indeed observed) to make assumptions and learn something about the DGP. This is the process of statistical analysis/estimation (which we will consider in Chapter 3).

In the rest of this chapter we present some of the most important and frequently use probability distributions.

## 2.1 Binomial and Bernoulli

The **Binomial** distribution can be used to characterise the sample distribution associated with a discrete variable  $R$  describing the total number of “successes” in  $n$  independent binary trials (e.g. the outcome is either 0 or 1, dead or alive, male or female, etc.), each with probability  $\theta$  of success.

You may think of this as a case where an individual randomly selected from the target population is associated with a probability  $\theta$  of experiencing an event of interest (e.g. having cardiovascular disease). Then, if you select  $n$  individuals randomly and you can assume that the event happens more or less independently on the individuals (e.g. if I have cardiovascular disease, this does not modify the chance that you do), then the sampling process can be described by the following equation

$$p(r \mid \theta, n) = \binom{n}{r} \theta^r (1 - \theta)^{n-r}; \quad r = 0, 1, \dots, n, \quad (2.1)$$

where the term

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n(n-1)(n-2)\cdots 1}{[r(r-1)(r-2)\cdots 1][(n-r)(n-r-1)(n-r-2)\cdots 1]}$$

is known as the *binomial coefficient*.

### ! Independence

The assumption of *independence* is most useful from the mathematical point of view. In a nutshell, this comes from the fact that if we have  $n$  observed data points and we can assume that they are independent, then their **joint** probability distribution (i.e. a function describing their joint variability) can be factorised into the product of the single probability distributions:

$$p(y_1, y_2, \dots, y_n | \theta) = \prod_{i=1}^n p(y_i | \theta). \quad (2.2)$$

The main advantage of this factorisation is that the elements of the product on the right-hand side of Equation 2.2 are of lower dimension than the full joint distribution on the left-hand side. For example, full independence implies that instead of modelling an  $n$ -dimensional probability distribution, we can simply model  $n$  1-dimensional distributions, which is much simpler, both intuitively and computationally.

Of course, there are instances where the assumption of independence clearly does not hold. For example, you may think of  $n$  observations taking value 1 if the  $i$ -th individual has some infectious disease (e.g. sexually transmitted) and 0 otherwise. Then, if I do have the disease, this may well affect your chance of becoming infected (depending on what the transmission mechanism is). Or in a slightly simpler context, it may be that we cannot claim *marginal* independence (i.e. that, without reference to any other feature, the two variables  $X$  and  $Y$  are independent). But we may be able to claim a *conditional* version — for instance if, given the value of a third variable  $Z$ , then  $X$  and  $Y$  may be reasonably assumed to not influence each other.

Essentially, the terms  $\theta^r$  and  $(1-\theta)^{n-r}$  are used to quantify the fact that out of the  $n$  individuals,  $r$  experience the event, each with probability  $\theta$ . Thus, because we are assuming that they are independent on one another, this happens for each with probability  $\theta$  or overall by multiplying this by itself for  $r$  times, i.e.  $\theta^r$ . Conversely,  $(n-r)$  do not experience the event, which again assuming independence is computed as  $(1-\theta)$  multiplied by itself for  $(n-r)$  times — or  $(1-\theta)^{(n-r)}$ .

The binomial coefficient is considered to account for the fact that we do not know which  $r$  of the  $n$  actually experience the event — only that  $r$  do and  $(n-r)$  do not. The binomial coefficient quantifies all the possible ways in which we can choose  $r$  individuals out of  $n$ . For example, if we consider of population of size 5 and we want to sample 3 people from it, we could enumerate all the possible combinations. In R we can do this using the following command

```
combn(5, 3)
```

```

      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    1    1    1    1    1    2    2    2    3
[2,]    2    2    2    3    3    4    3    3    4    4
[3,]    3    4    5    4    5    5    4    5    5    5

```

which returns a matrix where each column represents one of the possible  $\binom{5}{3} = \frac{5 \times 4 \times 3 \times 2 \times 1}{[3 \times 2 \times 1][2 \times 1]} = 10$  samples. The units in the population are labelled as 1,2,...,5 and so the first possible sample (the first

column of the output) would be made by the first three units, while the tenth (the last column) would be made by units 3, 4 and 5.

A Binomial with  $n = 1$  is simply a **Bernoulli** distribution, denoted  $Y \sim \text{Bernoulli}(\theta)$ . As is obvious,  $Y$  can only take on the values 0 (if the individual does not experience the event) or 1 (otherwise); in addition, because by definition  $\binom{1}{1} = \binom{1}{0} = 1$ , the probability distribution for a Bernoulli variable is simply

$$p(y | \theta) = \theta^y (1 - \theta)^{1-y}; \quad y = 0, 1.$$

The Bernoulli model essentially describes the sampling process for a binary outcome applied to a single individual. So if you have  $n$  individuals and you record whether each has an event or not, you can either describe the sampling process as  $n$  independent Bernoulli variables  $y_1, \dots, y_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$ , where the notation  $\stackrel{iid}{\sim}$  indicates *independent and identically distributed* variables; or by considering the total number of people who have had the event  $r = \sum_{i=1}^n y_i = y_1 + \dots + y_n$  and modelling it using a single Binomial variable  $r \sim \text{Binomial}(\theta, n)$ . When the only information available is about either the individual outcomes ( $y_i$ ) or the aggregated summary ( $r$ ), the two models are equivalent. If we have access to the individual level data (ILD), we can use the  $n$  Bernoulli variables directly. Often, however, we will not be able to access the ILD and will only know the observed value of the summary ( $r$ ) — in this case we cannot use the Bernoulli model and need to work with the Binomial sampling distribution.

Equation 2.1 can be used to compute the probability of observing exactly  $r$  individuals experiencing the event in a sample of  $n$ , given that the underlying probability is  $\theta$ . For example, if we set  $r = 12$ ,  $n = 33$  and  $\theta = 0.25$ , then

$$\begin{aligned} p(r = 12 | \theta = 0.25, n = 33) &= \binom{33}{12} 0.25^{12} (1 - 0.25)^{33-12} \\ &= 354817320 \times 0.00000005960464 \times 0.002378409 \\ &= 0.0503004. \end{aligned}$$

In R we can simply use the built-in command `dbinom` to compute Binomial probabilities, for instance

```
dbinom(x=12, size=33, prob=0.25)
```

```
[1] 0.0503004
```

would return the same output as the manual computation above.

We can also use the built-in command `rbinom` to simulate values from a given Binomial distribution. For example, the following code can be used to produce the graph in Figure 2.2 that illustrates a histogram from a random sample of 10000 observations from a Binomial(0.25,33) distribution.

```
tibble(r=rbinom(n=10000, size=33, prob=0.25)) %>%
  ggplot(aes(r)) + geom_histogram(
    breaks=seq(0, 33), color="black", fill="grey"
  ) + theme_bw() + xlab("Number of successes") +
  ylab("Absolute frequency") +
  scale_x_continuous(breaks = seq(0, 33, 1), lim = c(0, 33))
```

The R code here is probably unnecessarily complicated to show some of the features in terms of customisation of the graph, using the **tidyverse** and **ggplot2** approach. We first define a **tibble**, a special R object containing data, which we fill with a vector `r` that contains 10000 simulated values from the relevant Binomial distribution. We then apply `ggplot` to construct and style the histogram.

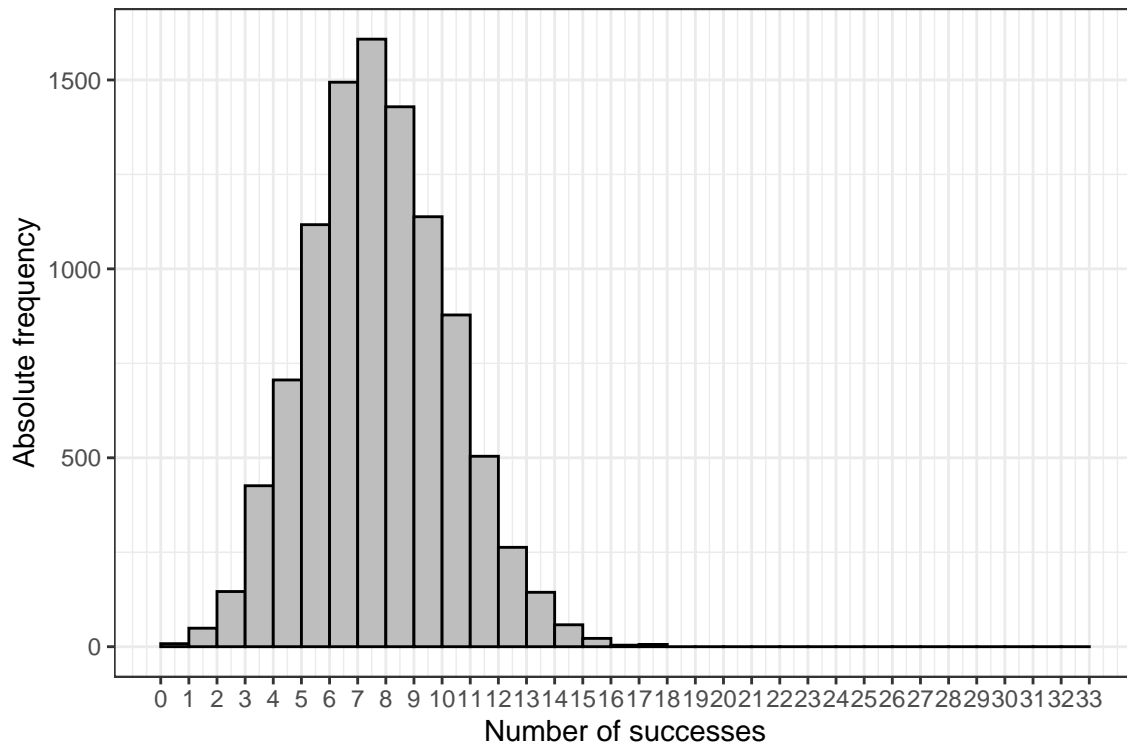


Figure 2.2: Histogram of a random sample of 10000 observations from a Binomial( $\theta = 0.25, n = 33$ ) distribution

The graph in Figure 2.3 shows three examples of Binomial distributions with  $\theta = 0.3$  but upon varying the sample size. As the sample size increases, the histogram becomes more symmetrical around the mean ( $\theta = 0.3$ ).

Because of the mathematical definition of the Binomial distribution, it can be proved that if  $R \sim \text{Binomial}(\theta, n)$ , then

- The mean (or “expected value”) is  $E[R] = \mu = n\theta$ ;
- The variance is  $\text{Var}[R] = \sigma^2 = n\theta(1 - \theta)$ .

Consequently, for a single Bernoulli variable, the mean is simply  $\theta$  and the variance is simply  $\theta(1 - \theta)$ . Both the Bernoulli and Binomial distributions were derived by the Swiss mathematician [Jacob Bernoulli](#), in the late 17th century. Bernoulli was part of a large family of academics and mathematicians, who have contributed to much of the early development of probability calculus and statistics.

## 2.2 Poisson

Suppose there are a large number of opportunities for an event to occur, but the chance of any particular event occurring is very low. Then the total number of events occurring may often be represented by a discrete variable  $Y$ . The sampling process that can be used to describe this situation is based on the **Poisson** distribution, named after the French mathematician [Siméon Denis Poisson](#). Poisson used this model in the 19th century in his “*Research on the Probability of Judgments in Criminal and Civil Matters*”, in which he modelled the number of wrongful convictions.

Mathematically, if  $y \sim \text{Poisson}(\theta)$  then we have that

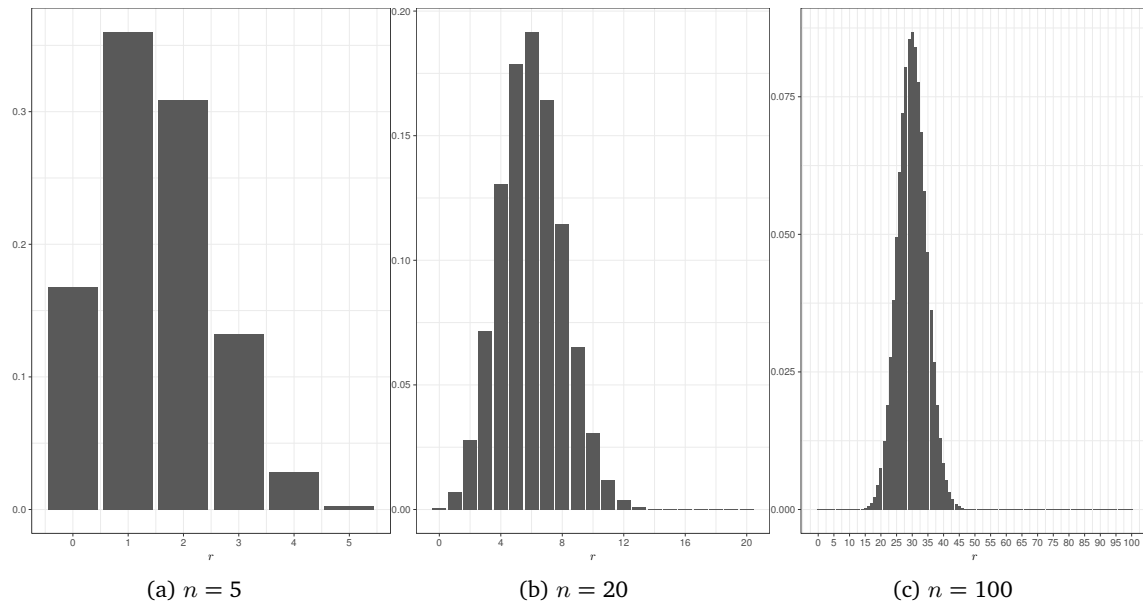


Figure 2.3: Binomial distributions for the number of successes in  $n = 5, 20, 100$  Bernoulli trials, each with probability  $\theta = 0.3$  of success

$$p(y | \theta) = \frac{\theta^y e^{-\theta}}{y!} \quad y = 0, 1, 2, 3, \dots \quad (2.3)$$

For a Poisson distribution, the parameter  $\theta$  represents both the mean *and* the variance:  $E[Y] = \mu = \text{Var}[Y] = \sigma^2 = \theta$ . This may at times be a limitation because often empirical data tend to violate this assumption — they show larger variance than the mean, a phenomenon often referred to as *overdispersion*. Suitable models can be used to expand the standard Poisson set up to account for this feature. In addition, in many applications, the Poisson sampling distribution will arise as a total number of events occurring in a period of time  $T$ , where the events occur at an unknown rate  $\lambda$  per unit of time, in which case the expected value for  $Y$  is  $\theta = \lambda T$ .

Generally speaking, the Poisson distribution is used to model sampling variability in observed counts, e.g. the number of cases of an observed disease in a given area. The examples in Figure 2.4 show that if events happen with a constant rate, observing for longer periods of time leads to smaller relative variability and a tendency towards a symmetrical shape. Comparison of Figure 2.4 with Figure 2.3 shows that, when sample size increases, a Binomial might be approximated by a Poisson with the same mean.

The main distinction between the Poisson and the Binomial models is that in the latter case we consider the number of events out of a fixed and known total number of possible occurrences (the sample size,  $n$ ). In the case of the Poisson, we generally consider the overall number of events (counts), without formally restricting the total number of occurrences. For this reason, the Poisson distribution may be used to model the probability of observing a certain number of goals in a football match (which, *theoretically* is unbounded), while the Binomial distribution can be used to model the number of Gold medals won by Italy at the next Olympic Games (which *physically* is bounded by the total number of sporting events).

If you consider this, it becomes perhaps more intuitive why the Poisson is the *limiting distribution* for the Binomial, as  $n$  increases to  $\infty$ . Figure 2.5 shows two histograms summarising 1,000,000 simulations from: a)  $y_1 \sim \text{Binomial}(\theta = 0.05, n = 200)$ ; and b)  $y_2 \sim \text{Poisson}(\mu = n\theta = 7.5)$ . As is possible to see, because the underlying probability  $\theta$  is fairly small (i.e. the event of interest is rare), for all intents and purposes,  $n = 200$  is effectively as large as  $n \rightarrow \infty$  and the two distributions overlap almost completely. If  $\theta$  is



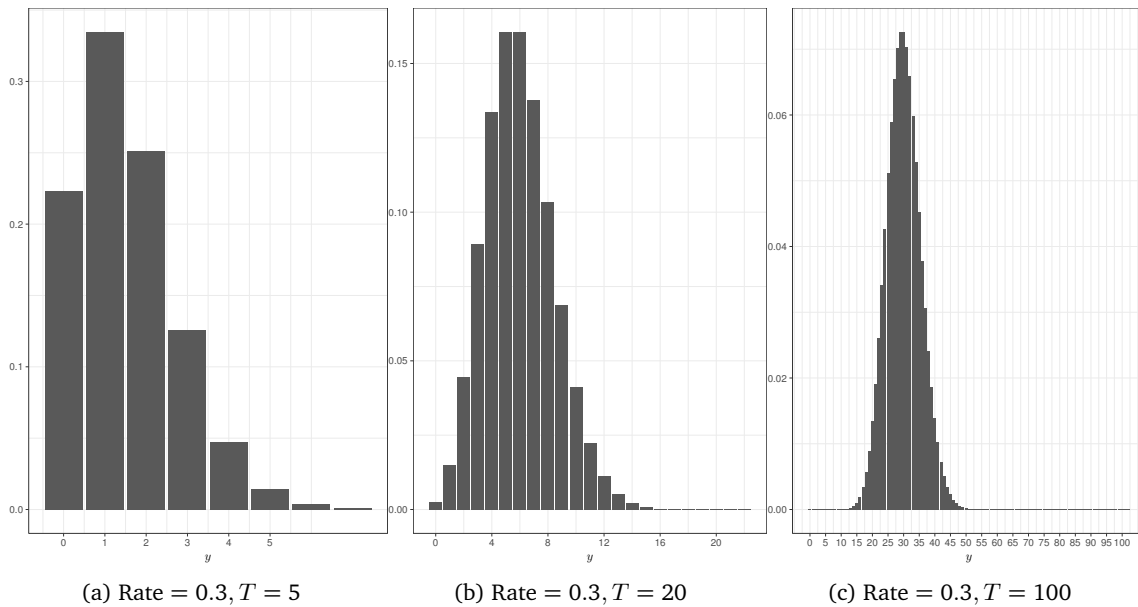


Figure 2.4: Poisson distributions representing the number of events occurring in time  $T = 5, 20, 100$ , when the rate at which an event occurs in a unit of time is  $\lambda = 0.3$ : the Poisson distributions therefore correspond to  $\theta = 1.5, 6$  and  $30$ .

larger, then we need a much bigger sample size  $n$  before the Binomial distribution is fully approximated by a  $\text{Poisson}(\lambda = n\theta)$ .

We can use R to compute probabilities or simulate random numbers from a Poisson distribution in a similar fashion as to what shown above. For instance, if we set  $\theta = 2$  we can compute the probability of observing  $y = 8$  events using the following command (note that, somewhat confusingly, R calls the parameter we have indicated as  $\theta$  with the name `lambda`).

```
dpois(x=8, lambda=2)
```

```
[1] 0.0008592716
```

The answer could be determined also by a simulation approach, as follows.

```
set.seed(10230)
# Vector of number of simulations
n=c(100,1000,10000,100000,1000000,10000000)
# Initialise (an empty) vector of (numeric) probabilities
prob=numeric()
# Simulates n[i] observations from a Poisson(lambda=2) and then
# counts the proportion of simulations with value 8
for (i in 1:length(n)) {
  y=rpois(n=n[i], lambda=2)
  prob[i]=sum(y==8)/n[i]
}
# Formats and shows the output
nlab=format(n, scientific=FALSE, big.mark=",")
```

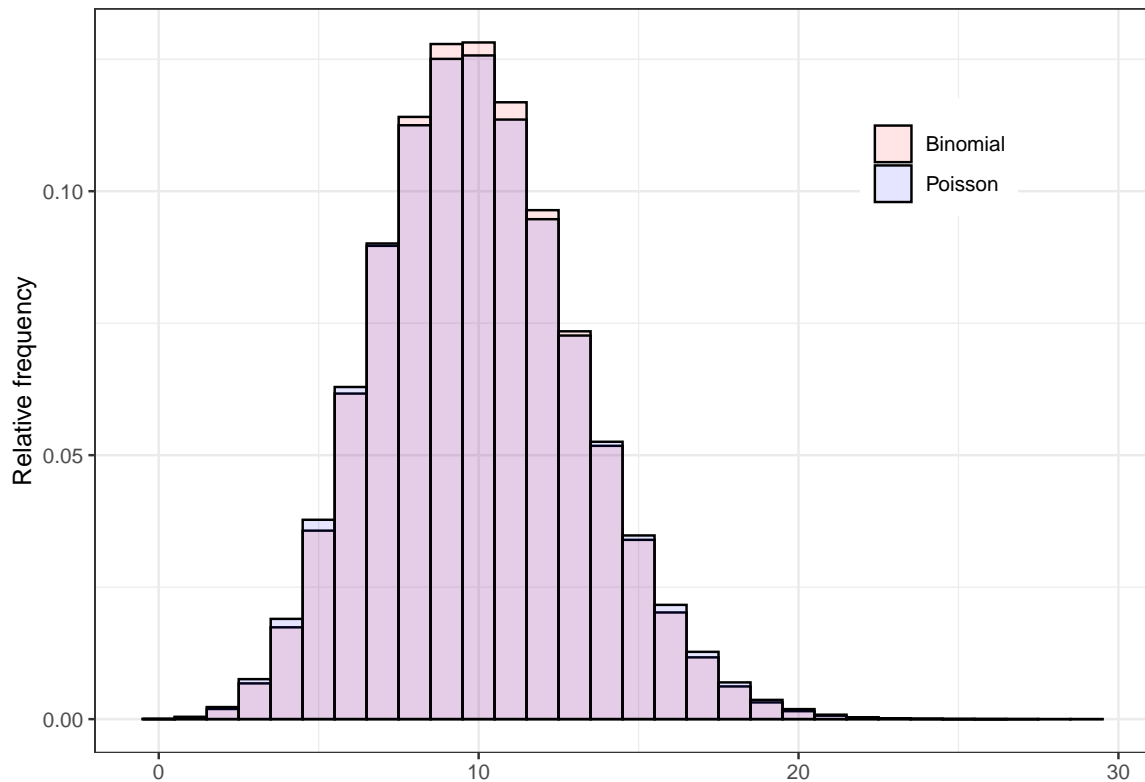


Figure 2.5: Histogram for two samples of 1000000 observations from a Binomial(0.05,200) (in red) and a Poisson( $7.5=0.05 \times 200$ ) (in blue). When the two histograms overlap, the resulting colour is shaded to purple

```
names(prob)=nlab
prob
```

```
      100      1,000      10,000      100,000      1,000,000      10,000,000
0.0000000  0.0000000  0.0015000  0.0008100  0.0007820  0.0008638
```

If we inspect the output of this process, we see that as we increase the size of the simulation to 10,000,000, then the numerical answer (0.0008638) becomes very close to the analytic one (0.0008593). When we consider a small number of simulations, our numerical estimate of the “true” analytic value is not very precise at all.

This process of numerical approximation of a quantity through simulation is often termed **Monte Carlo** analysis (more on this in STAT0019, if you take it).

### 2.3 Normal

The **Normal** distribution is fundamental to much of statistical analysis. Often it is referred to as **Gaussian**, from the name of its inventor, the German mathematician [Carl Friedrich Gauss](#), who in the late 18th century used it to describe “normally distributed errors”.

A *continuous* variable is associated with a Normal distribution if we can assume that the underlying sampling process has the following properties:

1. Symmetry: we expect to see most of the observations scattered around a central location, with “errors”, or deviation from the central location becoming increasingly smaller as we move further away.
2. Unboundness: we do not place any *physical* restriction on the size of the values that the variable can take on. Technically speaking, we say that the *range* of the variable is the set  $(-\infty; \infty)$ . Notice however that, in reality, we will never observe a variable to take on the value  $\infty$  and all our observations will in fact be finite values.

When we assume that a variable  $Y$  is well described by a Normal distribution, we then write  $Y \sim \text{Normal}(\mu, \sigma)$ , with

$$p(y | \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} \frac{(y - \mu)^2}{\sigma^2}\right); \quad -\infty < y < \infty. \quad (2.4)$$

Equation 2.4 shows that the Normal distribution is characterised by two parameters:  $\mu$  is the population mean and  $\sigma$  is the population standard deviation (see Chapter 1). Mathematically, the function in Equation 2.4 is a probability *density*. This is due to the fact that the underlying variable is continuous and, as such, it can take on any real number, e.g. -1.21131765366772, 2.32279253568801, -6.52670986118001, 25.5187887120438.

As mentioned in Chapter 1, for a continuous variable it is a mathematical impossibility that two different observations have the exact same value (technically, there is 0 probability that this happens). Thus, we can think of a probability density as a histogram where each group width is arbitrarily small. Figure 2.6 shows this for a sample of values from a Normal(0,1) distribution. The histogram in panel (a) uses group width of 0.50 — this means that the base on each rectangle is exactly 0.50 and the height is indicated along the  $y$ -axis of the graph. The histograms in panels (b)–(d) have increasingly small bar widths, approaching to a value  $\epsilon \rightarrow 0$ . Each graph has superimposed a Normal(0,1) density — as is possible to see the graph in panel (d) shows essentially no distinction between the histogram approximation and the true density.

An important consequence is that, unlike the case of discrete variables, for which R commands such as `dbinom(...)` or `dpois(...)` allows us to calculate the actual *probability* of observing an exact value, for continuous variables the command `dnorm(...)` computes the *density* associated with a small interval around the value.

The Normal distribution (and the respective R code) can be used to compute *tail area* probabilities. For example, the command

```
qnorm(p=0.975, mean=0, sd=1)
```

```
[1] 1.959964
```

returns the value  $y$  such that for a variable  $Y \sim \text{Normal}(\mu = 0, \sigma = 1)$ , we obtain  $\Pr(Y \leq y) = 0.975$  — that is the 97.5% quantile of the Normal distribution. Figure 2.7 shows graphically that the area under the Normal(0,1) density between  $-\infty$  and  $1.959964 \approx 1.96$  does cover most of the probability mass — and in fact exactly 97.5%, which in turns implies that  $\Pr(Y > 1.96) = 0.25$ .

Tail area probabilities can be used to compute the probability that a Normal variable lies within a given range. For example, we could use the following code

```
# Computes y1 so that, given Y~Normal(0,1), Pr(Y<=y1)=0.975
y1=qnorm(p=0.975, mean=0, sd=1)
# Computes y2 so that, given Y~Normal(0,1), Pr(Y<=y1)=0.025
y2=qnorm(p=0.025, mean=0, sd=1)
# Now verifies that Pr(Y<=y1) - Pr(Y<=y2) = 0.975 - 0.025 = 0.95
pnorm(q=y1, mean=0, sd=1)-pnorm(q=y2, mean=0, sd=1)
```

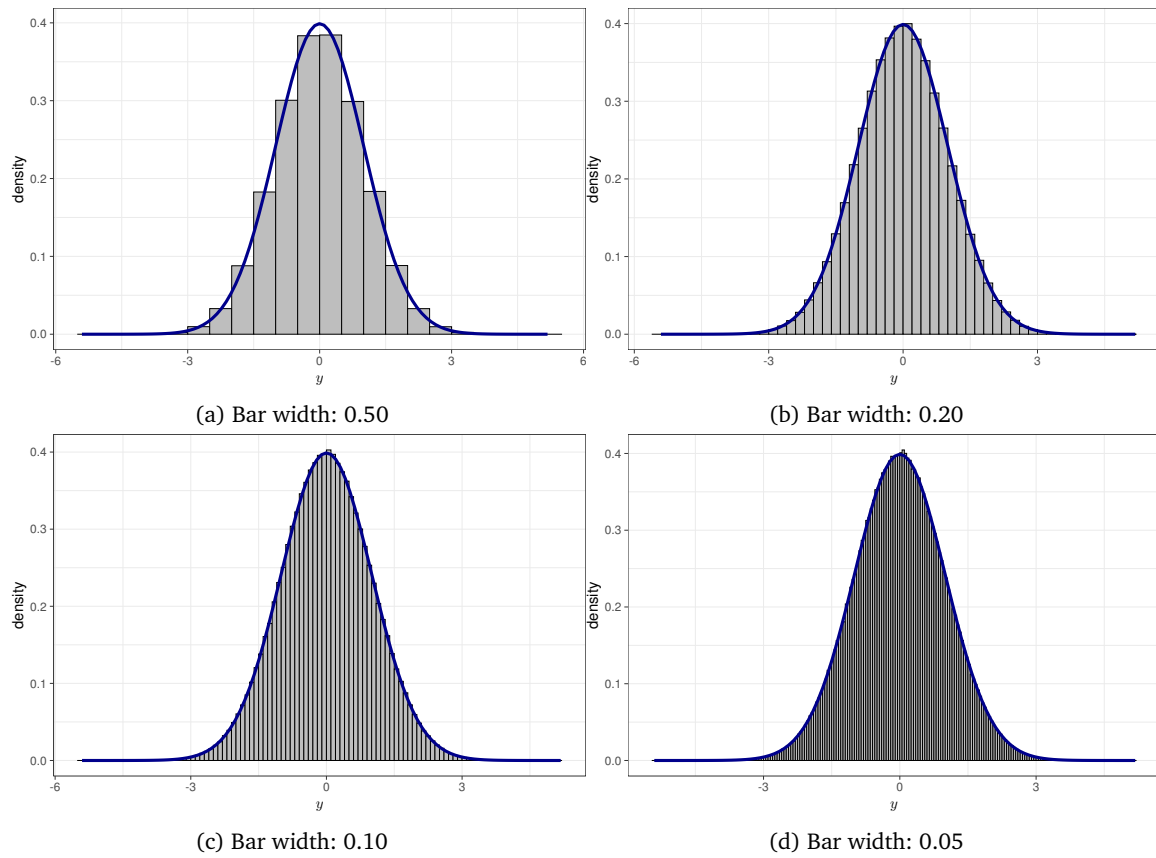


Figure 2.6: Histograms for a sample from a Normal distribution with  $\mu = 0$  and  $\sigma = 1$ , with superimposed (in blue) the density of a Normal(0,1).

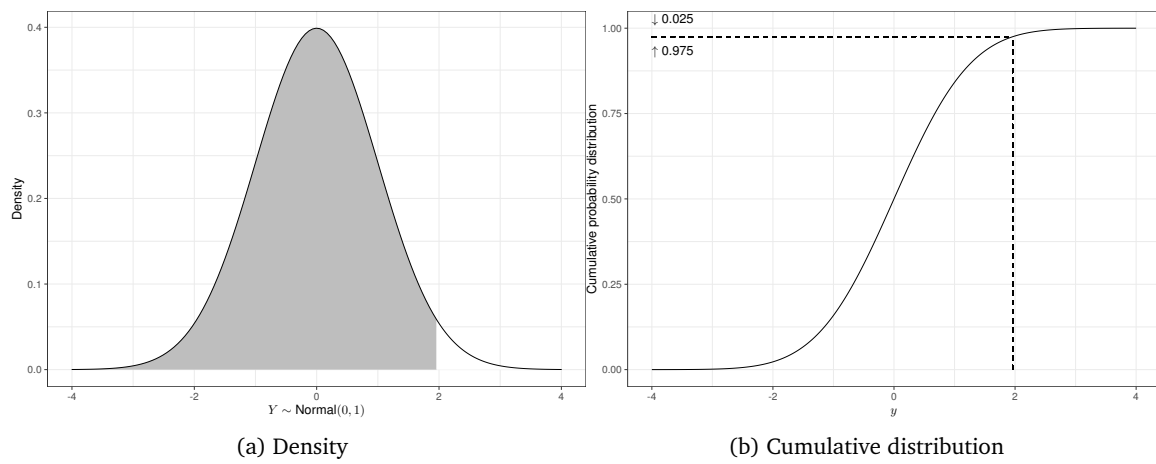


Figure 2.7: Tail area for a Normal(0,1) distribution

[1] 0.95

to check that the area comprised between the 97.5%-quantile and the 2.5%-quantile (i.e. the interval -1.96 to 1.96) does contain 95% of the probability distribution. Note the use of the built-in R functions `qnorm` and `pnorm` that compute the quantile, given a specified probability or a probability, given a specified quantile, under a Normal distribution.

## 2.4 Student's t

A standardized **Student's t** distribution has a prominent role in classical statistics as the sampling distribution of a sample mean divided by its estimated standard error.  $Y \sim t(\mu, \sigma^2, \nu)$  represents a Student's  $t$  distribution with  $\nu$  degrees of freedom:

$$p(y \mid \mu, \sigma^2, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi\nu\sigma}} \frac{1}{\left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{\frac{\nu+1}{2}}}; \quad -\infty < y < \infty, \quad (2.5)$$

where the symbol  $\Gamma(x) = (x-1)! = (x-1)(x-2)\cdots 1$  indicates the *Gamma function* (NB: not to be confused with the *Gamma distribution* presented in Section 2.5!).

For a Student's  $t$  distribution, we can prove that

- $E[Y] = \mu$ ;
- $\text{Var}[Y] = \sigma^2 \frac{\nu}{\nu-2}$ .

Because of the mathematical format of the Student's  $t$  distribution, it can be proved that the mean only exists if  $\nu > 1$ , and the variance only exists if  $\nu > 2$ .

The Student's  $t$  distribution was invented by [William Sealy Gosset](#), who published a paper with its derivation in 1908 under the pseudonym of “Student”, while being in secondment from his normal job at the Guinness brewery in Dublin, in the *Department of Statistical Science* at UCL. His work concerned the modelling of sampling variability for continuous, symmetric variables. He started by using a Normal model. However, because his problem involved samples of small sizes (for example, the chemical properties of barley where sample sizes might be as few as 3), this was not *robust*, i.e. it could not cope well with outliers, or extreme observations, which were still likely to be obtained, just because of sampling variability due to the small sample size.

The graph in Figure 2.8 shows the comparison between two Student's  $t$  distributions (with  $\mu = 0$  and  $\sigma = 1$ ) with degrees of freedom equal to  $\nu = 10$  and  $\nu = 2$ , (in red and green, respectively), against a standard Normal distribution (again with  $\mu = 0$  and  $\sigma = 1$ ).

As is possible to see, the lower the degrees of freedom, the “fatter” the tails of the Student's  $t$  distribution, in comparison to the Normal. This implies that the Student's  $t$  model assigns higher density to values that have larger deviations from the central tendency — this makes this model more “robust” to outliers or extreme observations. As  $\nu \rightarrow \infty$ , the Student's  $t$  distribution quickly converges to the standard Normal (note that  $\nu = 10$  is already producing a pretty good level of overlap between the red and blue curves).

It is possible to use R to sample values from the Student's  $t$  distribution. For example, we can create a graph similar to that of Figure 2.8 by typing the following command.

```
# Defines the range of the x-axis from -5 to 5 with increments of 0.01,
# i.e. -5.00, -4.99, -4.98, ..., 4.97, 4.98, 4.99, 5.00
# Plots the density of a t(0,1,3) distribution (in R, the default is
# to assume mean=0 and var=1)
```

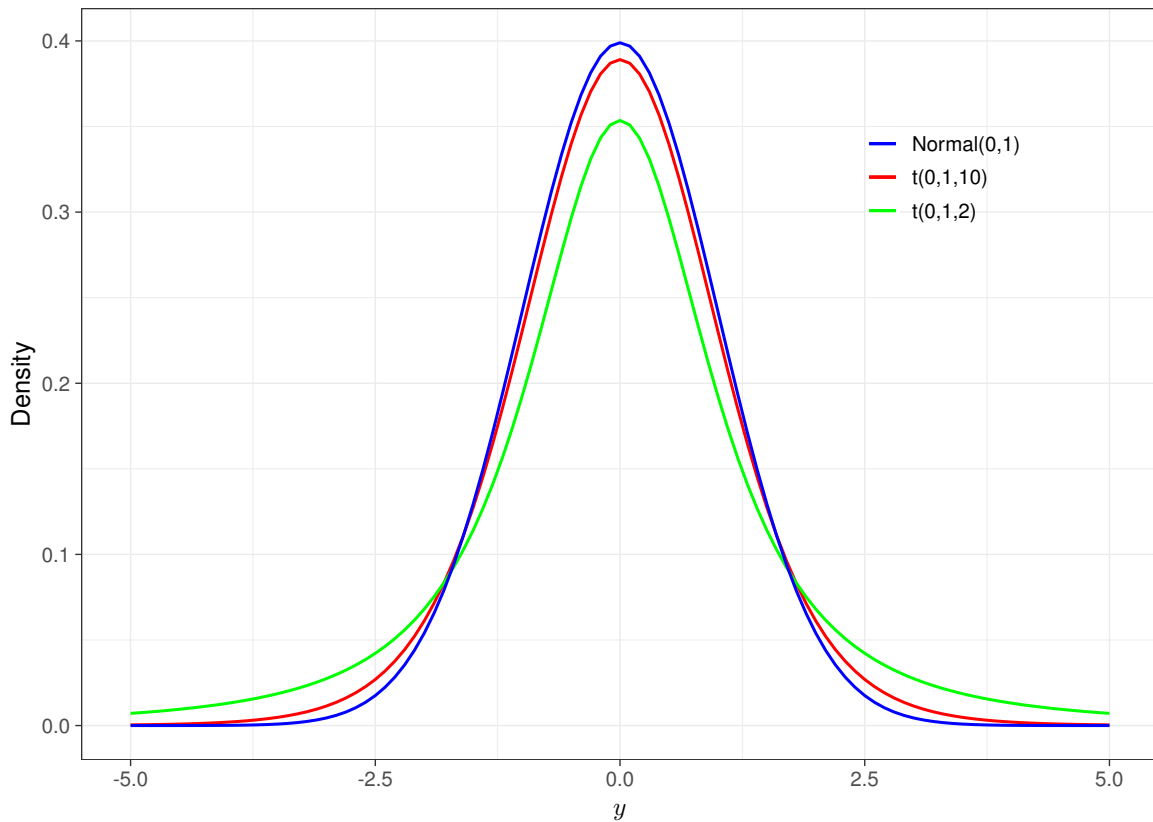


Figure 2.8: Comparison between different Student's t and Normal distributions

```
tibble(x=seq(-5,5,.01)) %>% ggplot(aes(x)) +
  stat_function(
    fun=dt, args=list(df=3), lwd=1.1
  )
```

Other built-in functions such as `rt(...)`, `pt(...)` and `qt(...)` are also available and can be used — more on this in Chapter 4.

## 2.5 Gamma and related distributions

**Gamma** distributions form a flexible and mathematically convenient class for continuous quantities constrained to be positive. Then  $Y \sim \text{Gamma}(a, b)$  represents a Gamma distribution with properties:

$$p(y | a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by}; \quad -\infty < y < \infty; a, b > 0 \quad (2.6)$$

and

$$\begin{aligned} E[Y] &= \frac{a}{b} \\ V[Y] &= \frac{a}{b^2} \end{aligned}$$

The parameter  $a$  is called the *shape*, while  $b$  is called the *rate* of the Gamma distribution. Alternative parameterisations exist in terms of the parameters  $(a, c)$ , where  $c = 1/b$  is called the *scale* — but note that, in this case, the form of the density in Equation 2.6 needs to be re-written accordingly!

One of the typical uses of the Gamma distribution is to model sampling variability in observed costs, associated with a sample of patients (you will see this extensively, if you take STAT0019). Other examples involve time-to-event models (e.g. to investigate how long before some event of interest occurs)

The general form of the Gamma distribution includes as special cases several other important distributions. Figure 2.9 shows a few examples of different Gamma distributions, upon varying the two parameters  $a$  and  $b$ . As is possible to see, for specific choices of the parameters, we retrieve other distributions, related to the Gamma. Some important examples of such cases are discussed below.

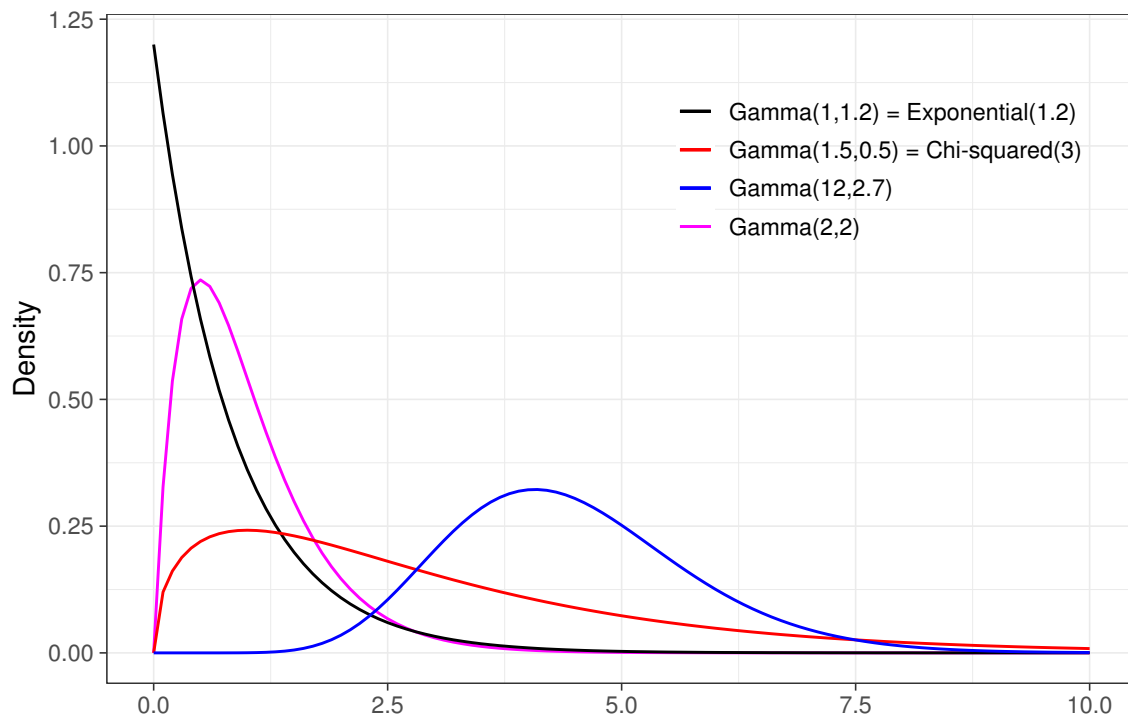


Figure 2.9: Some examples of Gamma distributions upon varying the parameters  $a$  and  $b$ . For different choices of the parameters, the Gamma distribution can give rise to other distributions, including the Exponential and the Chi-squared

### 2.5.1 Exponential

The **Exponential** distribution is obtained when we set  $a = 1$ , i.e.

$$Y \sim \text{Gamma}(1, b) \equiv \text{Exponential}(b). \quad (2.7)$$

The Exponential distribution is sometimes used as model for sampling variability in times until an event of interest happens (e.g. time until a patient dies of a given disease). One major limitation of the Exponential model is however that it is only characterised by a single parameter  $b$ , which also determines the value of the mean and variance as

- $E[Y] = \frac{1}{b}$ ;

- $\text{Var}[Y] = \frac{1}{b^2}$ .

For this reason, the Exponential model is often too rigid and cannot represent well variations across a wide range of values for the variable  $y$ .

### 2.5.2 Chi-squared

The **Chi-squared** distribution (sometimes indicated using the Greek letter notation  $\chi^2$ ) is obtained from a Gamma distribution in which  $a = \nu/2$  and  $b = 1/2$ , i.e.

$$Y \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{1}{2}\right) \equiv \text{Chi-squared}(\nu). \quad (2.8)$$

Using the properties of the Gamma distribution, it is trivial to derive that if  $Y \sim \text{Chi-squared}(\nu)$ , then

- $E[Y] = \frac{\nu/2}{1/2} = \nu$ ;
- $\text{Var}[Y] = \frac{\nu/2}{1/2^2} = 2\nu$ .

A useful piece of distribution theory is that if  $Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{Normal}(\mu, \sigma^2)$  then the sample summaries are

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad \text{and} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

(recall Section 1.6) and it can be proved that

$$(n-1) \frac{S^2}{\sigma^2} \sim \text{Chi-squared}(n-1) \quad \text{and that} \quad \frac{(\bar{Y} - \mu)}{(S/\sqrt{n})} \sim t(0, 1, n-1). \quad (2.9)$$

In addition, in a 1900 paper, [Karl Pearson](#)<sup>1</sup> also proved that if  $Y_1, \dots, Y_n$  are a set of observations and  $E_1, \dots, E_n$  are the corresponding expected values from those observations (given a particular data generating process assumed to underlie the collection of the observations), then

$$\sum_{i=1}^n \frac{(Y_i - E_i)^2}{E_i} \sim \text{Chi-squared}(n-1). \quad (2.10)$$

We will use these results extensively in Chapter 4, in the context of hypothesis testing.

## 2.6 Other distributions

Probability calculus as taught in Statistical courses often concentrates on a small number of probability distributions, e.g. those mentioned above. But there are of course many more possibilities. Which

---

<sup>1</sup> Karl Pearson was a controversial figure. He was the founder of the *Department of Applied Statistics* (later renamed to *Department of Statistical Science*) at UCL in 1911, the first ever department of Statistics in any academic institutions. During his academic career, he has provided enormous and important contributions to the development of statistical theory and was widely regarded as the leading statistician in his time. However, he was sadly also a proponent of eugenics, a discipline that aimed at improving the genetic quality of a human population by excluding certain genetic groups judged to be inferior, and promoting other genetic groups judged to be superior. UCL has recently launched an [inquiry](#) into the history of eugenics at the university, which will also deliver recommendations on how to manage its current naming of spaces and buildings after prominent eugenicists.



distribution to use to reflect assumptions about the underlying data generating process and sampling variability (or indeed epistemic uncertainty about unobservable quantities — more on this in STAT0019) is a matter of substantive knowledge of the problem. And in reality, it becomes as much an art as it is a science, which requires experience as well as expertise.

Additional distributions that you are likely to encounter in STAT0014, STAT0015 and STAT0019 include the Uniform, log-Normal, Weibull, Gompertz, Beta, Multinomial, Dirichlet and Fisher's F distributions. Good compendia of probability distributions are presented in several textbooks, including Spiegelhalter, Abrams, and Myles (2004) and Lunn et al. (2012)

The graph in Figure 2.10 shows a graphical representation of the relationships among a large number of probability distributions. As is possible to see, there are many more choices than the simple few showed above. And interestingly, in many cases, the distributions tend to show close mathematical connections (for instance one distribution may be obtained as a special case of another, just like for the Exponential/Gamma). More details are given in Leemis and McQueston (2008).

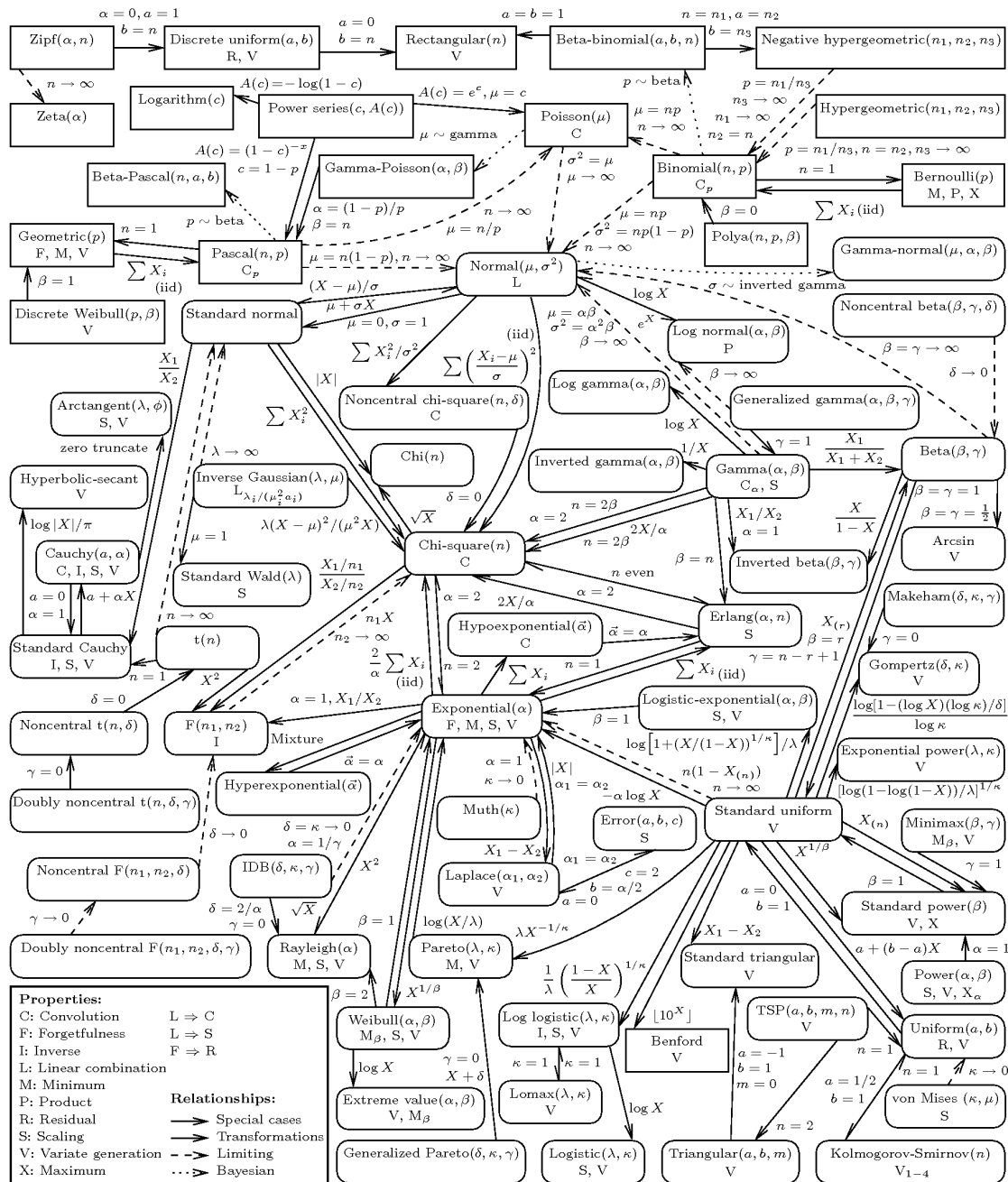


Figure 2.10: A graphical representation of the relationships among a (non-exhaustive) set of univariate probability distributions. Source: Leemis and McQueston (2008).

## Parameter estimation: doing Statistics

As mentioned earlier, in a nutshell the problem of statistical *inference* consists in

1. Obtaining a sample of observations  $\mathbf{y} = (y_1, \dots, y_n)$  from a population of interest. The process with which the data become observed is subject to **sampling variability** — you get to see only one of the many possible samples that could be obtained by extracting a number  $n$  of units that is (typically) much smaller than the total size of the population  $N$ .
2. Characterise the sampling variability using a suitable probability distribution  $p(\mathbf{y} | \boldsymbol{\theta})$ , defined as a function of a set of model **parameters**,  $\boldsymbol{\theta}$ , to describe the data generating process.
3. Use the observed data  $\mathbf{y} = (y_1, \dots, y_n)$  to **learn** about the unobservable population features described by the model parameters.

Interestingly (and somewhat confusingly) not even statisticians have agreed on a single way in which this process should be carried out. In fact, there are at least three main philosophical approaches to the problem of statistical inference.

### 3.1 Point estimates

#### 3.1.1 The Bayesian approach

The **Bayesian** approach (which is the topic of STAT0019) is historically the first to have been developed. The original ideas and the basic structure date back to the publication of an essay by Reverend [Thomas Bayes](#) (Bayes 1763), an English non-conformist minister, after whom the whole approach is named.

A Bayesian model specifies a full probability distribution to describe uncertainty. This applies to data, which are subject to sampling variability, as well as to parameters (or hypotheses), which are typically unobservable and thus are subject to epistemic uncertainty (e.g. the experimenter's imperfect knowledge about their value) and even future, yet unobserved realisations of the observable variables (Gelman et al. 2013).

As a consequence, probability is used in the Bayesian framework to assess any form of imperfect information or knowledge. Thus, before even seeing the data, the experimenter needs to identify a suitable probability distribution to describe the overall uncertainty about the data  $\mathbf{y}$  and the parameters  $\boldsymbol{\theta}$ . We generally indicate this as  $p(\mathbf{y}, \boldsymbol{\theta})$ . Using the basic rules of probability, it is always possible to factorise a joint distribution as the product of a marginal and a conditional distribution (you will see this again if you take STAT0019). For instance, we could re-write  $p(\mathbf{y}, \boldsymbol{\theta})$  as the product of the marginal distribution for the parameters  $p(\boldsymbol{\theta})$  and the conditional distribution for the data, given the parameters  $p(\mathbf{y} | \boldsymbol{\theta})$ . But in exactly the same fashion, one could also re-express the joint distribution as the product of the marginal distribution for the data  $p(\mathbf{y})$  and the conditional distribution for the parameters given the data  $p(\boldsymbol{\theta} | \mathbf{y})$ .

Consequently,

$$p(\mathbf{y}, \boldsymbol{\theta}) = p(\boldsymbol{\theta})p(\mathbf{y} | \boldsymbol{\theta}) = p(\mathbf{y})p(\boldsymbol{\theta} | \mathbf{y})$$

from which Bayes' Theorem follows in a straightforward way:

$$p(\boldsymbol{\theta} | \mathbf{y}) = \frac{p(\boldsymbol{\theta})p(\mathbf{y} | \boldsymbol{\theta})}{p(\mathbf{y})}. \quad (3.1)$$

While mathematically incontrovertible, Bayes' Theorem has deeper philosophical implications, which have led to heated debates, within and without the field of Statistics. In fact, the qualitative implications of this construction are that, if we are willing to describe our uncertainty on the parameters before seeing the current data through a probability distribution, then we can update this uncertainty by means of the evidence provided by the data into a *posterior* probability, the left hand side of Equation 3.1. This allows us to make inference in terms of direct probabilistic statements.

In all but trivial models, Equation 3.1 also presents some computational challenges because it is often very hard or even impossible to compute the ratio on the right hand side analytically. Consequently, until the 1990s the practical implementation of Bayesian models has been hampered by this problem. The widespread availability of cheap computing as well as the development of suitable clever methods for simulations (e.g. *Markov Chain Monte Carlo*, or MCMC, which you will encounter extensively if you take STAT0019) have helped overcome these problems and make Bayesian analysis very popular in several research areas, including economic evaluation of health care interventions and adaptive clinical trial designs.

Leaving all the technicalities aside (which you will encounter in more details if you take STAT0019), Bayesian inference proceeds by using the following scheme.

1. Define a “**prior**” distribution  $p(\boldsymbol{\theta})$  to describe the current uncertainty on the model parameters. This represents the knowledge *before* the data  $\mathbf{y}$  become available.
2. Observe data  $\mathbf{y}$ , whose sampling variability is modelled using  $p(\mathbf{y} | \boldsymbol{\theta})$ .
3. Apply (an approximation to the exact computation intrinsic in) Bayes' Theorem of Equation 3.1 to compute the “**posterior**” distribution  $p(\boldsymbol{\theta} | \mathbf{y})$ . This distribution represents the “revised” or “updated” level of uncertainty on the parameters  $\boldsymbol{\theta}$ , after the data  $\mathbf{y}$  have become available.

Suppose we consider a simple case where the data are  $R \sim \text{Binomial}(\theta, n)$ , with  $n = 13$  and we have observed  $r = 9$  successes. We want to make inference on the parameter  $\theta$ . Figure 3.1 shows an example of Bayesian inference in action (we will dispense with all the difficult technical points here and only concentrate on the interpretation).

Imagine that you are willing to specify the current level of uncertainty about  $\theta$  using the black curve (labelled as “Prior”). This essentially implies that, before seeing any other data, you believe reasonable to assume that the most likely value for  $\theta$  is around 0.4 and most likely it will be included in the interval (0.2 – 0.6). Values below 0.2 or above 0.6 are associated with increasingly smaller values of the probability mass as you move away from the mode (0.4) and towards the extremes (0 and 1). Technically, we could use a Beta(9.2,13.8) distribution to encode these assumptions (but, again, this is only a technicality and you will see more on this if you take STAT0019).

The red curve is a representation of the contribution brought by the observed data. In fact, this is the likelihood function (that is described in Section 3.1.2). Again, leaving all the details aside, the interpretation of the red curve is that, intuitively, because we have observed  $r = 9$  successes over  $n = 13$  individuals, the data seem to suggest that the “true” underlying probability of success may be higher than we originally thought (the red curve has a mode around 0.69231).

Finally, Bayesian inference is obtained by inspecting the blue curve, showing the posterior distribution. This is typically a compromise between the prior knowledge and the evidence provided by the data. As mentioned above, technically this can be complex, or even impossible to determine analytically — and in fact, most often we resort to simulation algorithms to obtain a suitable approximation.

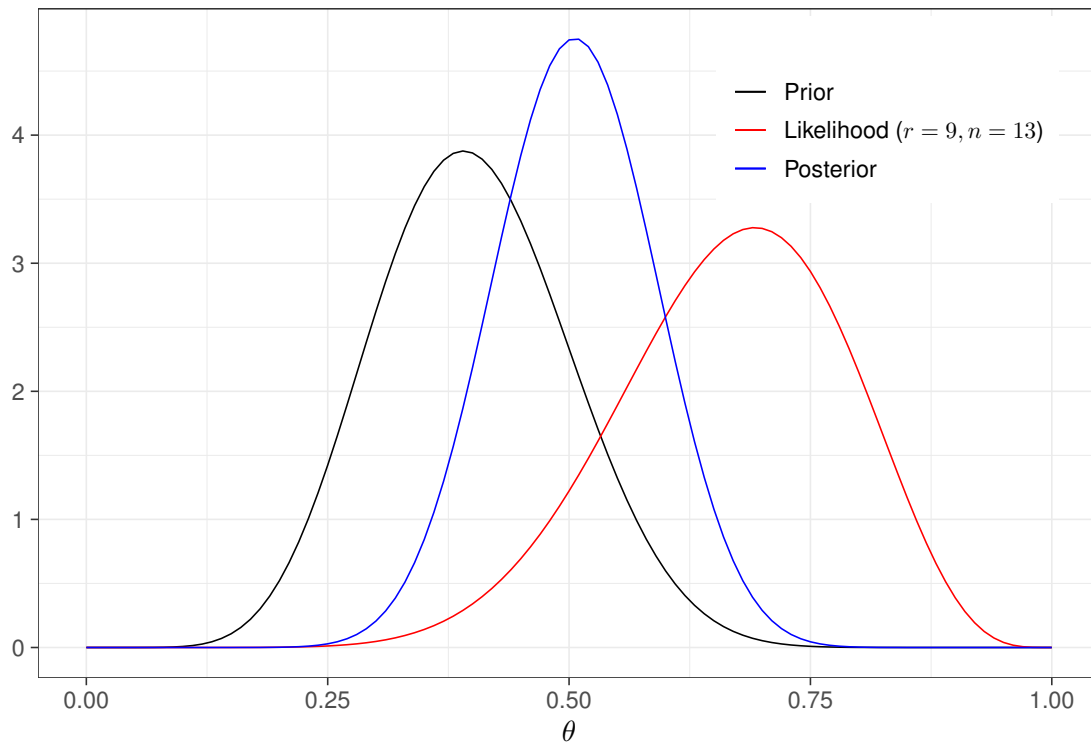


Figure 3.1: An (extremely simplified!) example of Bayesian inference on a probability  $\theta$  for a Binomial sampling distribution model, with  $r = 9$  observed successes out of  $n = 13$  individuals

In general, once the whole posterior distribution is available, it is then possible to describe it using suitable summaries. For example, the usual point estimate is the mean of the posterior distribution. In the case of Figure 3.1, because the blue curve is reasonably symmetrical, the mean corresponds with the mode (the point where the distribution is highest), which in this case can be computed as 0.5056. So we initially thought that the probability that a random individual would experience the event under study was centered around the prior mean 0.4000 and we have revised this to the posterior summary 0.5056, after observing 9 individuals in a sample of 13 experiencing the event.

### 3.1.2 The Likelihood approach

This approach to statistical inference is almost single-handedly developed by [Ronald Fisher](#)<sup>1</sup>. The main ideas (which you will encounter extensively in STAT0015 and STAT0016) underlying Fisher's theory can be somewhat roughly summarised as follows.

1. Unlike in the Bayesian approach, Fisher considers parameters as *fixed* (although unknown) quantities. As such, we cannot model our uncertainty over the true underlying value of  $\theta$  using a probability distribution  $p(\theta)$ .
2. The only randomness in a statistical analysis is generated by the sampling variability associated with the observed data. We still want to use the data to learn about the world (as described by our model,

<sup>1</sup> Much as Karl Pearson, Fisher was also a very controversial figure. His brilliance as a scientist is undisputed and he has made contributions to many disciplines, including Statistics, Biology and Genetics. On the other hand, his views were also close to eugenics — in fact, he took the post of head of the Department of eugenics at UCL, in 1933. He was also heavily criticised for his views on the link between smoking and lung cancer, which he strongly denied in favour of some underlying genetic features, despite the evidence that was already at the time becoming substantial.

indexed by the parameters  $\theta$ ) and to do so, the only thing we need to consider is the **likelihood function**.

! Known, unknown, fixed...

Notice that in regards to point 1. above, also from the Bayesian point of view parameters may be fixed quantities: it is possible that the “true” proportion of males in the world is an unmutable constant — we just do not (and cannot!) know its true value with absolute certainty. The Bayesian way to deal with this *epistemic* uncertainty is to consider  $\theta$  as a random variable and describe current uncertainty with the prior.

As shown in Chapter 2, we can model sampling variability using probability distributions. For example, in the case of the Binomial distribution, this is defined as

$$p(r | \theta, n) = \binom{n}{r} \theta^r (1 - \theta)^{n-r}$$

— recall Equation 2.1. The expression above is a function of the observed data, given the values of the parameters (as should be clear from Chapter 2).

However, what we really want to do when making inference is not so much fixing the value for  $\theta$  and computing probabilities for possible values of  $r$  — rather we want to use the observed value of  $r$  to learn what the most plausible value of  $\theta$  is in the “true”, underlying DGP. Thus, Fisher’s main idea was to create a different kind of function, which varied with the parameters, but depended on the observed (fixed)  $r$ . He called this the likelihood function and defined it as the same equation used for the sampling variability associated with the model — except that the fixed and varying arguments are flipped around. For example, in the Binomial case, the likelihood function is

$$\mathcal{L}(\theta | r) = \theta^r (1 - \theta)^{n-r}. \quad (3.2)$$

Equation 3.2 only takes the terms in the Binomial sampling distribution that depend directly on the model parameter. For instance, the Binomial coefficient  $\binom{n}{r}$  is irrelevant because it does not include  $\theta$  and thus it is discarded in forming the likelihood.

Panel (a) in Figure 3.2 shows the Binomial sampling distribution for a fixed value of  $\theta = 0.3$ . If that was the “true”, underlying value of the probability that a random individual drawn from the population of interest experiences the event of interest, then we would expect 4 successes out of  $n = 13$  individuals sampled to be the most probable outcome. Observing 9 successes would be a somewhat unlikely event: we can use R to compute this probability as `dbinom(x=9, size=13, prob=0.3)=0.00338`.

Panel (b) presents the likelihood function for three possible observed samples. The black curve corresponds to the analysis when  $r = 2$ ,  $n = 13$ , the red curve is drawn for  $r = 4$ ,  $n = 13$  and the blue curve is derived in the case where  $r = 9$  and  $n = 13$  — notice that in the interest of comparability, we present here the *rescaled* likelihood functions, i.e.  $\mathcal{L}(\theta | r) / \max[\mathcal{L}(\theta | r)]$ , so that the range on the  $y$ -axis goes from 0 to 1.

The interpretation of the likelihood function is that it describes “how likely” each possible value of the parameter is, given the observed data. For example, in the case of the black curve, the most likely value of the parameter after observing  $r = 2$  successes out of  $n = 13$  individuals is  $\hat{\theta} = 0.154$ , which is indicated by the black dashed line. For  $r = 4$ , then  $\hat{\theta} = 0.308$  (the red dashed line) and for  $r = 9$  then  $\hat{\theta} = 0.692$  (the blue dashed line).

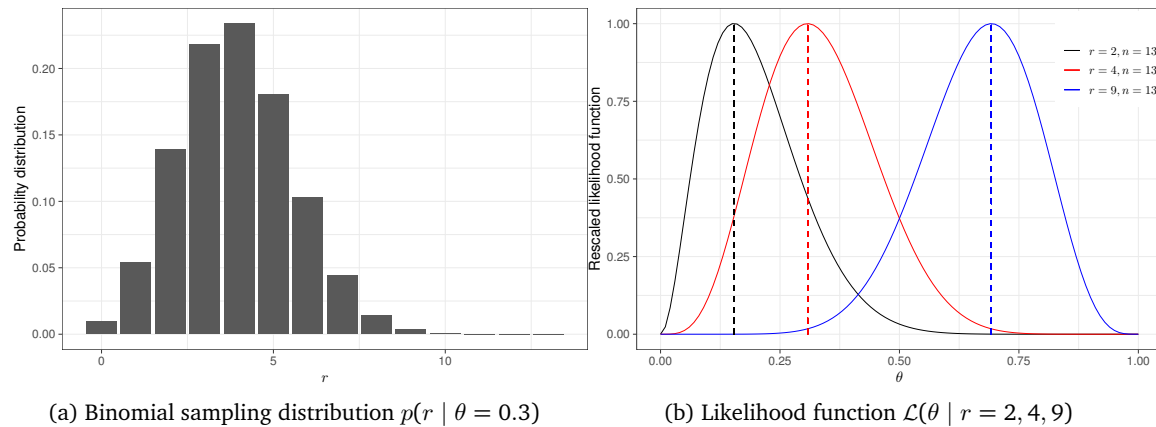


Figure 3.2: Binomial sampling distribution compared to three likelihood functions for different observed values of the data  $r$ . Each likelihood function is rescaled to have a maximum value at 1, for comparability. Panel (a) shows the sampling distribution for  $\theta = 0.3$ , while panel (b) shows the likelihood function for three possible observed values of  $r = 2, 4, 9$  (in black, red and blue respectively)

### ! Important

Notice that the language we have used here is purposely pedantic and focuses on the concept of “how *likely*”, as opposed to “how *probable*”. This is because the likelihood function is **not** a probability distribution (technically, the reason for this is that the likelihood function does not integrate to 1, i.e. the area under the curve representing the likelihood function is not equal to 1, which is a necessary condition for a mathematical function to represent a probability distribution). Fisher was very clear on this — that is why he coined the phrasing “likelihood function”. But the concepts are sometimes conflated, so it is important to be careful on the distinction.

As seems sensible, the greater the number of observed successes, the higher the most likely value of the underlying true probability of success — if  $\theta$  is indeed very large, then most people will experience the event and so we will tend to observe large values for  $r$  in any given sample. So, by and large, the simplified process presented here describes how inference is performed in this setting:

1. we define a sampling distribution to describe variability in the observed data;
2. we turn this into a likelihood function for the model parameter(s); and
3. we determine the value that is associated with the highest likelihood (again: not probability!).

This is used as the best estimate — and it is referred to as the **maximum likelihood estimator** (MLE). Mathematically, this last step amounts to maximising the likelihood function. Technically this is done by finding the value for which the first derivatives of the function is equal to 0 and then checking that the second derivative of the function is negative to ensure that that point is indeed a maximum.

In order to perform this *analytically* (i.e. for a general value of the random variable  $R$ , rather than the specific observed value  $r$ ), it is usually easier to work on the log scale, i.e. trying to maximise the log-likelihood  $\ell(\theta) = \log \mathcal{L}(\theta \mid R)$ . Figure 3.3 shows that both functions are maximised by the same point on the  $x$ -axis.

Thus, if we consider the log-likelihood for the Binomial example, we have

$$\ell(\theta) = \log \mathcal{L}(\theta \mid R) = R \log \theta + (n - R) \log(1 - \theta)$$

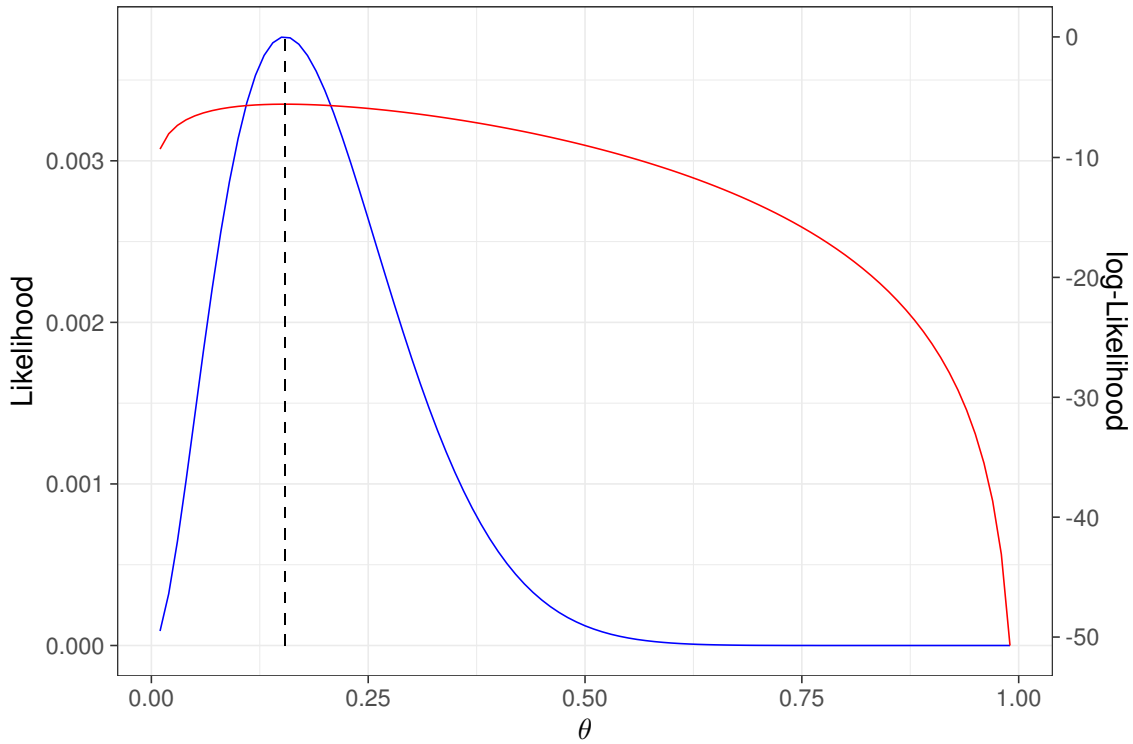


Figure 3.3: Likelihood function (in blue, with values on the  $y$ -axis on the left hand side) and Log-likelihood function (in red, with values on the  $y$ -axis on the right hand side) for the Binomial example with  $r = 2$ . Both functions are maximised at the same point along the  $x$ -axis

and so, making use of the properties that: i. the derivative of a sum is the sum of the derivatives; ii. for a variable  $x$ , the derivative of  $\log(x)$  is  $\frac{1}{x}$ ; iii. for a variable  $x$ , the derivative of  $\log(1 - x)$  is  $-\frac{1}{1-x}$ ; we can compute

$$\text{First derivative of } \ell(\theta) = \ell'(\theta) = \frac{R}{\theta} - \frac{n - R}{1 - \theta}. \quad (3.3)$$

Setting this to 0 and solving for  $\theta$ , i.e. finding the value  $\hat{\theta}$  in correspondence of which the first derivative is 0, gives

$$\begin{aligned} \frac{R}{\hat{\theta}} - \frac{n - R}{1 - \hat{\theta}} = 0 &\Rightarrow (1 - \hat{\theta})R - \hat{\theta}(n - R) = 0 \\ &\Rightarrow R - \hat{\theta}R - \hat{\theta}n + \hat{\theta}R = 0 \\ &\Rightarrow \hat{\theta} = \frac{R}{n} = \sum_{i=1}^n \frac{Y_i}{n} = \bar{Y}, \end{aligned}$$

i.e. the sample mean of the  $n$  individual Bernoulli variables (recall Section 2.1).

If we also compute the second derivative, i.e. the derivative of Equation 3.3 and check that it is negative (which in this case it is) to ensure that  $\hat{\theta}$  is indeed a maximum point, then we have obtained a *functional form* for the MLE. And we can use this to compute its value for the observed sample, which in this case is  $\hat{\theta} = \bar{y} = \frac{r}{n} = \frac{2}{13} = 0.154$ , which can be used as the point estimate for the parameter of interest.

In many cases the derivatives can be computed analytically, which means we can find the MLE as a function of the general random variable  $R$  and then compute the value in correspondence of the observed value  $r$ ,



as in the case above. However, there may be cases where the resulting likelihood function is complex and either difficult or even impossible to differentiate (i.e. compute the derivatives). In these cases, most likely, you will be using a computer to maximise the likelihood function *numerically*. For example, R has several built-in functions to perform numerical optimisation, for instance `optim` or `optimise`, which can be used to that effect. The following code implements this calculation using `optimise`.

```
# Defines the likelihood function as a R function
Lik=function(theta,r,n) {
  # The function depends on three arguments:
  # 'theta' is a vector of values specifying the range of the parameter
  # 'r' is the observed number of successes
  # 'n' is the observed sample size
  theta^r*(1-theta)^(n-r)
}
# Use 'optimise' to obtain the MLE for w=2 and n=13 in the interval (0,1),
# ie the range of theta
optimise(Lik,r=2,n=13,interval=c(0,1),maximum=TRUE)

$maximum
[1] 0.1538463

$objective
[1] 0.003768044
```

Firstly, we code up the likelihood function of Equation 3.2 in the function `Lik`, which takes as arguments the parameter `theta`, as well as the observed data `r` and `n`. Then we run `optimise` by passing as arguments the function we want to maximise (`Lik`), the values for the data (`r` and `n`), the interval over which we want to maximise the function (`interval=c(0,1)`, which represents the range of  $\theta$ ) and the option `maximum=TRUE`, which instructs R to compute the maximum (instead of the minimum of the function). The results is stored in the object `maximum`, which has a value of 0.1538463, which is essentially identical to the analytic value 0.1538462 (the differences are due to the approximation in the numerical optimisation procedure performed through `optimise`).

### 3.1.3 The Frequentist approach

The third major approach to statistical inference is termed **frequentist** and it was built on major contributions by [Jerzy Neyman](#) and [Egon Pearson](#), in the 1930s<sup>2</sup>.

As in the likelihood approach, parameters are considered to be unknown, but fixed quantities. However, the frequentist school does not attempt to make inference for a specific set of data, but rather it considers and evaluates inference *procedures* (e.g. the way in which an estimator is defined). Inference consists in the probabilistic assessment of the properties of the procedure, according to suitably defined criteria (more on this later).

In a nutshell, the frequentist approach defines a *statistic*, i.e. a function  $f(Y)$  of the observed data, based on its “optimality” according to the long-run performance. In other words, we want to use as *estimator* for a given parameter a function of the observed data that, *if we were able to repeat the experiment over and over under the same conditions* would guarantee that, in the long-run, we would be certain of some properties.

<sup>2</sup> Jerzy Neyman (a Polish mathematician and statistician) and Egon Pearson (the son of Karl Pearson — see Section 2.5.2) developed most of the theory underlying the main frequentist ideas while the former was visiting the Department of Statistical Science at UCL, where the latter had taken over his father as head.

**i** Long-run/short-life

The “long-run” argument underpins the frequentist approach and all the theory developed by Neyman and Pearson. This was very popular at the time, although not universally advocated. And criticism of the frequentist philosophy was not confined to the field of Statistics. The British economist [John Maynard Keynes](#) is quoted to have dismissed it because “... *in the long run, we are all dead*” (Keynes 1923).

For instance, let us consider a Bernoulli sample of  $n$  individuals  $\mathbf{Y} = (Y_1, \dots, Y_n) \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$  and the two possible estimators

$$f_1(\mathbf{Y}) = \sum_{i=1}^n \frac{Y_i}{n} = \bar{Y} \quad \text{and} \quad f_2(\mathbf{Y}) = \frac{\text{Med}(\mathbf{Y})}{n}.$$

Here,  $f_1(\mathbf{Y})$  is the sample mean, where  $f_2(\mathbf{Y})$  is computed as the median of the data  $\mathbf{Y}$  divided by the sample size  $n$ .

Now, imagine that we knew with no uncertainty that the true value of the underlying probability of success is  $\theta = 0.3$  — of course, in reality we cannot know this and to give our best estimate for its value is in fact the objective of the analysis. But if we did know, then we could imagine a simulation process that aims at mimicking would could happen if we were to repeat a very large number of times the experiment in which we collect data on  $n = 13$  individuals and record how many experience the event.

In R we could do this by using the following code

```
# Sets the 'seed' so that we always get the same results
set.seed(12)
# Sets the "true" value of the probability of success (assumed known)
theta=0.3
# Sets the sample size in each repeated experiment
n=13
# Sets the number of simulations
nsim=1000
# Defines a matrix "samples" with nsim rows and n columns,
# initially with "NA" (empty) values
samples=matrix(NA,nsim,n)
# Then creates a loop to fill each row of the matrix "samples" with n
# simulated values from a Binomial(theta, 1) (i.e. we simulate all the
# individual Bernoulli data Y_i)
for (i in 1:nrow(samples)) {
  samples[i,]=rbinom(n,1,theta)
}
head(samples)
```

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
[1,]    0    1    1    0    0    0    0    0    0    0    0    1    0
[2,]    0    0    0    0    0    0    0    0    1    0    1    0    0
[3,]    0    0    0    0    0    1    1    1    0    1    0    1    0
[4,]    0    0    0    1    0    1    0    0    0    0    1    0    0
[5,]    0    0    0    1    0    0    0    1    1    0    1    0    0
[6,]    0    0    1    0    1    0    1    0    0    0    1    1    1
```

(the command `head(samples)` shows the first few rows of the matrix of simulations, where 0 indicates that we have simulated a “failure” and 1 indicates a “success”).

For each simulated dataset (=replicate of the experiment), we can record the observed value of the two statistics  $f_1(\mathbf{y})$  and  $f_2(\mathbf{y})$ . Figure 3.4 shows the output for the first 20 replicates of the experiments: in each of the 20 rows in the graph, the black dots represent the  $n = 13$  simulated values (notice that, because we are simulating from a Bernoulli, then the only possible outcomes are 0=“failure” and 1=“success”); the red diamonds are the observed value of the sample mean  $f_1(\mathbf{y})$ ; and the blue squares are the observed values of the rescaled median  $f_2(\mathbf{y})$ .

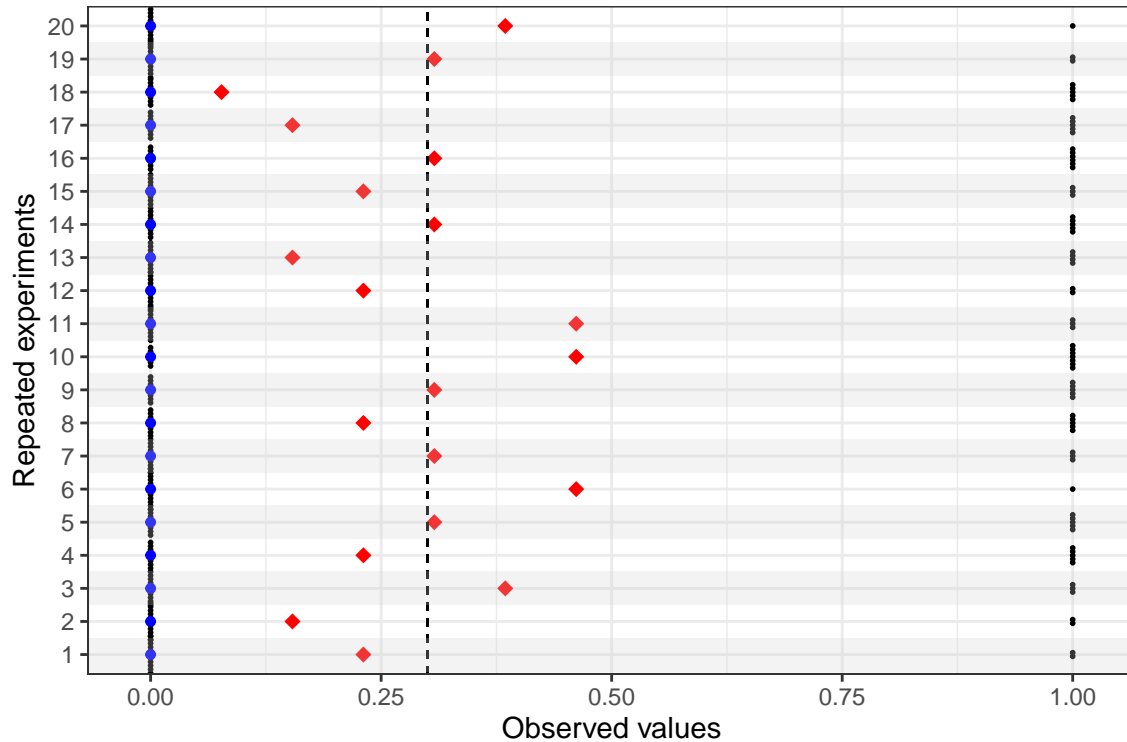


Figure 3.4: A graphical representation of a simulation exercise to describe repeated experiments from a Binomial case with  $n = 13$  and  $\theta = 0.3$ . For each of the 20 simulations presented, the black dots indicate the simulated Bernoulli outcomes  $y_1, \dots, y_n$ ; the red diamonds indicate the sample means  $\bar{y}$  and the blue dots are the rescaled sample medians  $\frac{\text{Med}(\mathbf{y})}{n}$ . The vertical dashed line indicates the true value for the parameter  $\theta = 0.3$

The frequentist approach makes use of the fact that, because they are functions of the observed data (which are subject to sampling variability), the statistics  $f_1(\mathbf{Y})$  and  $f_2(\mathbf{Y})$  also are subject to sampling variability — intuitively, this is expressed by the different values that we could observe for them, if we were able to do the experiment over and over again (as in Figure 3.4). Using the `nsim` simulations stored in the matrix `sample`, we could investigate the sampling distributions of  $f_1(\mathbf{Y})$  and  $f_2(\mathbf{Y})$ , for example using the following commands

```
tibble(x=samples %>% apply(1,mean)) %>%
  ggplot(aes(x)) + geom_histogram(bins=10,fill="grey",col="black") +
  theme_bw() + xlab("") + ylab("Density") +
```

```

geom_segment(
  aes(x=0.3,y=-Inf,xend=0.3,yend=Inf),
  linetype="dashed",lwd=0.85
)

tibble(x=samples %>% apply(1,median)/n) %>%
  ggplot(aes(x)) + geom_histogram(bins=10,fill="grey",col="black") +
  theme_bw() + xlab("") + ylab("Density")

```

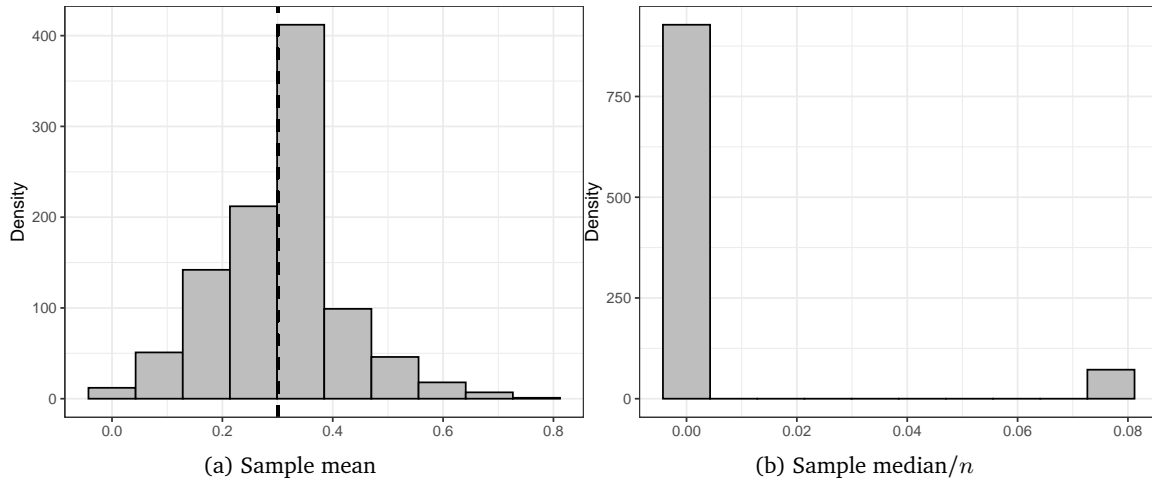


Figure 3.5: Histograms for the sampling distributions of the two statistics  $f_1(\mathbf{Y})$  and  $f_2(\mathbf{Y})$

(the built-in function `apply(matrix, 1, FUN)` takes all the rows of the first argument `matrix` and applies the function `FUN` to each of them, returning a vector of summaries, e.g. the mean or the median). Figure 3.5 shows histograms for the sampling distributions of  $f_1(\mathbf{Y})$  and  $f_2(\mathbf{Y})$ , while Table 3.1 reports some summaries, including the mean, standard deviation, 2.5%, 50% (median) and 97.5% percentiles of the sampling distributions.

Table 3.1: Summary of the sampling distributions for the two statistics

	Mean	SD	2.5%	Median	97.5%
$f_1(\mathbf{Y}) = \bar{Y}$	0.3030	0.1316	0.0769	0.3077	0.6154
$f_2(\mathbf{Y}) = \text{Med}(\mathbf{Y})/n$	0.0055	0.0199	0.0000	0.0000	0.0769

As is possible to see, the two estimators have rather different characteristics.

1. The rescaled sample median  $f_2(\mathbf{Y})$  has a much smaller standard deviation. However, this is only due the fact that its possible values are 0 or 0.0769 only and thus there is much less possible variation in the observed data, which leads to a smaller variance. If we consider the mean over the sampling distribution, the value is 0.00554, which is in fact very far from the true value  $\theta = 0.3$ .
2. Conversely, the distribution of the sample mean  $f_1(\mathbf{Y})$  is *centered* around the “true” value for  $\theta$ . This is a very important property to a frequentist — it is called *unbiasedness* and a statistic having this

property is called an *unbiased estimator* (for a given parameter). Formally, we define an unbiased estimator  $f(\mathbf{Y})$  as one for which  $E[f(\mathbf{Y})] = \theta$ .

The reason why unbiasedness is so important to a frequentist is that it encapsulate perfectly the concept of long-run optimality: for any given sample, we have no reassurance that the unbiased statistic would give us a value equal, or even close, to the true underlying parameter. In fact, looking at Figure 3.4, none of the observed values for  $f_1(\mathbf{Y})$  is equal to the true value for  $\theta = 0.3$ . However, because of its unbiasedness, we can say that in the long-run, on average we would “get it right” by using  $f_1(\mathbf{Y})$  to estimate  $\theta$ .

### ! Unbiasedness (et al)

Unbiasedness is only one of the frequentist properties — arguably, the most compelling from a frequentist perspective and possibly one of the easiest to verify empirically (and, often, analytically). There are however many others, including:

1. *Bias-variance trade-off*: we would consider as optimal an estimator with little (or no) bias; but we would also value ones with small variance (i.e. more precision in the estimate), So when choosing between two estimators, we may prefer one with very little bias and small variance to one that is unbiased but with large variance;
2. *Consistency*: we would like an estimator to become more and more precise and less and less biased as we collect more data (technically, when  $n \rightarrow \infty$ ).
3. *Efficiency*: as the sample size increases indefinitely ( $n \rightarrow \infty$ ), we expect an estimator to become increasingly precise (i.e. its variance to reduce to 0, in the limit).

#### 3.1.4 You’re my one and only (theory)...?

Interestingly, as it happens, the MLE for a given parameter does generally have all the good frequentist properties. For this reason, we can effectively be frequentist and select the MLE as our optimal estimator, which has contributed to some people presenting and using these two approaches as a combined and unified theory. In fact, Fisher on the one hand and Neyman and Pearson on the other saw them as two irreconcilable schools of thought and have spent many years arguing vehemently with each other — so much so that when Fisher moved to UCL, a new special chair was created for him, to avoid him being in the same department as Pearson.<sup>3</sup>

What the two approaches do have in common, as mentioned above, is the fact that both take a non-Bayesian stance on how to deal with the model parameters. In both cases, parameters are considered as fixed but unknown quantities, which govern the DGP. In order to learn about the true, underlying value of the parameters, we can use suitable statistics: in the case of Fisher’s theory, the MLE because it is based on the likelihood function (which contains all the information that the sample can provide on the parameters); in the case of the frequentist approach, again most often the MLE, because it upholds all the good frequentist properties. But you should be aware of the fundamental distinction in these two approaches.

## 3.2 Interval estimation

We have just seen how different approaches to statistical inference produce point estimates for a parameter of interest. This is often a very good starting point — and sometimes it is the only summary reported, e.g. in non-scientific publications (such as in the mainstream media, as shown in Figure 3.6).

<sup>3</sup> During his early career, Fisher had also heated arguments with Karl Pearson, whom he started off admiring very much, but with whom he fell out over a rejection of one of Fisher’s papers in *Biometrika*, the scientific journal edited by Karl Pearson.

ADVERTISEMENT Privacy Policy | Feedback Like 16.1M Monday, Jul 15th 2019 12PM 17°C 3PM 19°C 5-Day Forecast ADVERTISEMENT

# MailOnline News

Home News U.S. | Sport | TV&Showbiz | Australia | Femall | Health | Science | Money | Video | Travel | DailyMailTV | Discounts

Latest Headlines | Royal Family | News | World News | Arts | Headlines | France | Most read | Wires | Prime Day Login

## Obesity is the new smoking, warns NHS chief: Cancers linked to weight are set to DOUBLE and there will be 40,000 cases a year by 2035

- NHS England predicts 40,800 obesity-linked cancer cases every year by 2035
- Boss Simon Stevens has said that obesity is 'replacing' smoking as largest threat
- By 2035, one person will be diagnosed every 13 minutes with cancer that has developed because of their weight

By BEN SPENCER MEDICAL CORRESPONDENT IN CHICAGO FOR THE DAILY MAIL  
 PUBLISHED: 22:41, 31 May 2019 | UPDATED: 22:52, 31 May 2019

Share
787 shares
1.2k View comments

Cancer cases caused by obesity will double in the next two decades, shocking new figures reveal.

Britain's spiralling obesity crisis – driven by poor diets and sedentary lifestyles – means that by 2035, someone will be diagnosed every 13 minutes with cancer that has developed because of their weight.

**NHS** England predicts there will be 40,800 obesity-linked cancer cases every year by then – up from 22,800 in 2015.



Site Web Enter your search Search



ADVERTISEMENT

Like Daily Mail

Follow @DailyMail

Follow @dailymailuk

Follow Daily Mail

Follow Daily Mail

Follow Daily Mail

**DON'T MISS**

- Love Island: Amy reveals Arabella told her she was 'absolutely HATED' by the public and admits she had therapy 12 times in the villa
- Alesha Dixon shows off her baby bump in a leaf-print bikini as she poses make-up free after saying she 'finally respects' her body at 40
- A look at EastEnders star Maisie Smith's 18th birthday: Actress who played Tiffany Butcher celebrates with her first legal drink in a pub and boozy presents
- Kate Wright sends Rio Ferdinand wild as she shows off her incredible figure in sizzling selfie... after speaking out on her decision to quit

Figure 3.6: An example of inappropriate reporting of statistical quantities only as point estimates in the mainstream media

In reality, the point estimate is only likely to be indicative of what the true value of the underlying parameters could be — because of a combination of the sampling variability (that implies we can only learn so much from the observed data, unless  $n \rightarrow \infty$ , which it never is...) and the intrinsic epistemic uncertainty that we have on the parameter.

For this reason, it is a good idea to provide measures of **interval** estimate to complement the point estimate for a given (set of) parameter(s).

### 3.2.1 Bayesian approach

From a Bayesian point of view, in theory, interval estimate does not pose any additional complication. Once the full posterior probability distribution for the parameter has been computed, then we can simply present any summary we want. As seen in Section 3.1.1, the point estimate can be taken as the mean or the mode of the posterior. But we can also simply compute intervals that express directly probabilistic statements about the values of the parameters.

For example, we can prove that the posterior distribution for the Binomial example shown in Section 3.1.1 is a Beta(18.2, 17.8) — again the technical details are not important here and you will see them if you take STAT0019. We can use this information to compute analytically quantities such as the value  $\theta_U$ , the point in the parameter space (defined in this case as the interval  $[0; 1]$ , because  $\theta$  represents a probability) in correspondence of which  $\Pr(\theta < \theta_U | y) = 0.95$ . With the current specification, this is 0.6408 — that is we can estimate that the posterior probability that  $\theta$  is less than 0.6408 is exactly 95%.

Sometimes we will be able to make this or similar calculations analytically (which technically means we can solve an integral) and we can make this computations in R for instance using the command

```
q_U=qbeta(p=0.95, shape1=18.2, shape2=17.8)
```

which returns the exact value 0.640801.

A Bayesian interval is essentially computed using a similar process as the interval in correspondence of which a certain amount of probability lies:

$$\text{Bayesian 95\% interval} = [\theta_L; \theta_U] \text{ such that } \Pr(\theta_L \leq \theta \leq \theta_U | y) = 0.95.$$

Generally speaking we can use a simulation approach (which is essentially what MCMC does) and

1. Simulate a large number of values from the posterior distribution
2. Use the simulated values to compute tail-area probabilities and obtain the relevant interval.

For example, we can use the following R code to compute the 95% interval for the example above.

```
# Simulates 10000 values from the posterior distribution Beta(18.2,17.8)
theta=rbeta(n=10000, shape1=18.2, shape2=17.8)
# Computes the 2.5% quantile (that is the point leaving an area of 2.5% to its left)
q_L=quantile(theta, 0.025)
# Computes the 97.5% quantile (that is the point leaving an area of 97.5% to its left)
q_U=quantile(theta, 0.975)
# Displays the resulting interval estimate
c(q_L, q_U)
```

2.5%      97.5%

0.3436120 0.6679136



The result is that we compute that 95% of the posterior distribution lies in the interval  $[0.344; 0.668]$ , or, in other words, that the probability that the true parameter is contained between 0.344 and 0.668, *given the model assumptions and the observed data* is exactly 95%.

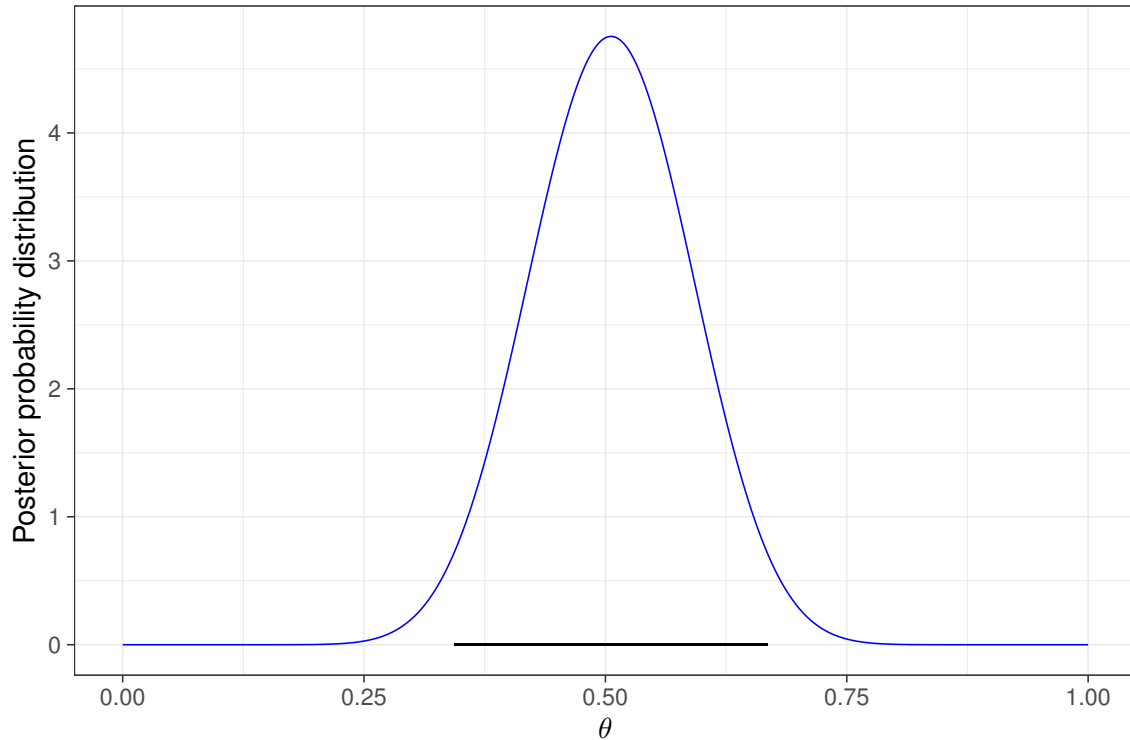


Figure 3.7: Posterior distribution for the Binomial example. The dark horizontal line below the distribution indicate the 95% interval estimate

Figure 3.7 shows again the posterior probability distribution; the dark horizontal line at the bottom of the histogram indicates the 95% interval.

#### **i** Bayesian intervals

As mentioned above, *theoretically*, summarising the posterior distribution through an interval does not pose any additional problem, once the target distribution is available. The main problem is, of course, that we need to know what the posterior is before we can manipulate it — and this is the main point about doing simulations, e.g. using MCMC algorithms. This is not trivial, though and thus requires some specific training.

In addition, the procedure shown above to compute a Bayesian interval delivers what is often called a *central* interval — that is the one that leaves equal amount of area to its right and to its left. It does work very well if the underlying distribution is reasonably symmetric, but it may not be the best option when the distribution is skewed or “multimodal” (e.g. it has several “humps”, like a camel). In those cases, we can still compute suitable Bayesian intervals, based on slightly different theory and computations (these are often called “High Posterior Density”, or HPD intervals). The details are not important here.



### 3.2.2 Likelihood and frequentist approach

As mentioned in Section 3.1.2 and Section 3.1.3, the likelihood and frequentist approaches are not, in fact, a unified theory. Nevertheless, because effectively the MLE is often the “best” frequentist estimator, we can present the methodology in a compacted way.

The basic idea is that, as mentioned above, the MLE is a statistic, i.e. a function  $f(Y)$  of the observed data and, as such, is associated with a sampling distribution. If we are able to determine what this sampling distribution is, then we can use it to compute tail-area probabilities that can be used to derive interval estimates.

For example, for a generic statistic  $f(\mathbf{Y})$ , we can prove that

$$Z = \frac{f(\mathbf{Y}) - \mathbb{E}[f(\mathbf{Y})]}{\sqrt{\text{Var}[f(\mathbf{Y})]}} \sim \text{Normal}(0, 1), \quad (3.4)$$

at least approximately, as the sample size  $n \rightarrow \infty$  (which in practical terms, simply means that  $n$  is “large enough”). A very general rule of thumb is that  $n > 30$  suffices for this result to hold).

We have seen before that the MLE for the probability of success  $\theta$  in the  $n$  independent Bernoulli case is the sample mean  $f_1(\mathbf{Y}) = \bar{Y} = \sum_{i=1}^n \frac{Y_i}{n} = \frac{R}{n}$ , where  $Y_i \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$ . We can prove that the sampling distribution for  $f_1(\mathbf{Y})$  is *approximately* a Normal with mean  $\theta$  (which makes  $f_1(\mathbf{Y})$  an unbiased estimator) and variance  $\sigma^2/n$ , where  $\sigma^2 = \theta(1 - \theta)$ . From this we can also derive that the *standardised* version of  $f_1(\mathbf{Y})$ , obtained by considering  $f_1(\mathbf{Y})$  minus its expected value and divided by its standard deviation, is

$$\frac{\bar{Y} - \theta}{\sqrt{\theta(1 - \theta)/n}} \sim \text{Normal}(0, 1). \quad (3.5)$$

We have seen in Figure 2.7 that for a Normal(0,1), the point 1.96 is the one leaving 97.5% of the distribution to its left. Similarly, we can prove that the point -1.96 leaves 2.5% to its left and thus the interval  $[-1.96; 1.96]$  includes 95% of the probability

$$\Pr\left(-1.96 \leq \frac{\bar{Y} - \theta}{\sqrt{\theta(1 - \theta)/n}} \leq 1.96\right) \approx 0.95$$

(the approximation comes about because: i. the distribution of  $f_1(\mathbf{Y})$  is only approximately Normal; and ii. we are using the rounded value 1.96 instead of the exact quantile of the Normal distribution). Re-arranging the terms inside the probability statement, we get

$$\Pr\left(\theta - 1.96\sqrt{\frac{\theta(1 - \theta)}{n}} \leq \bar{Y} \leq \theta + 1.96\sqrt{\frac{\theta(1 - \theta)}{n}}\right) \approx 0.95. \quad (3.6)$$

At this point, we need a further layer of approximation: we do not know the true value of the parameter  $\theta$  — only its estimate  $\hat{\theta}$  (e.g. the MLE) and so we can compute the 95% **confidence interval** for the original statistic  $\bar{Y}$  as

$$\left[ \hat{\theta} - 1.96\sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}}; \hat{\theta} + 1.96\sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}} \right]. \quad (3.7)$$

! Probability of what?...

There is a subtle but crucial point in this argument. Equation 3.6 computes a probability. This probability however is computed with respect to the sampling distribution of the statistic — **not** the parameter  $\theta$ . From the frequentist/likelihood point of view, this is perfectly fine — in fact neither Neyman and Pearson, nor Fisher would want to compute a probability distribution directly for the model parameters. To them, the parameters are just fixed quantities and so cannot be associated with distributions.  $\theta$  has a “true” value and so  $\Pr(\theta = \text{true value}) = 1$  and  $\Pr(\theta \neq \text{true value}) = 0$  — we just do not know what the true value is.

Again in line with the long-run philosophy, the interpretation of a confidence interval is that *if we were able to replicate the experiment over and over again under the same conditions and each time we computed a confidence interval according to the procedure in Equation 3.7, then, in the long-run, the resulting interval would cover the true value of the unknown but fixed parameter 95% of the times*. Figure 3.8 expands on Figure 3.4, by including the 95% intervals computed using Equation 3.7, depicted as the blue horizontal lines.

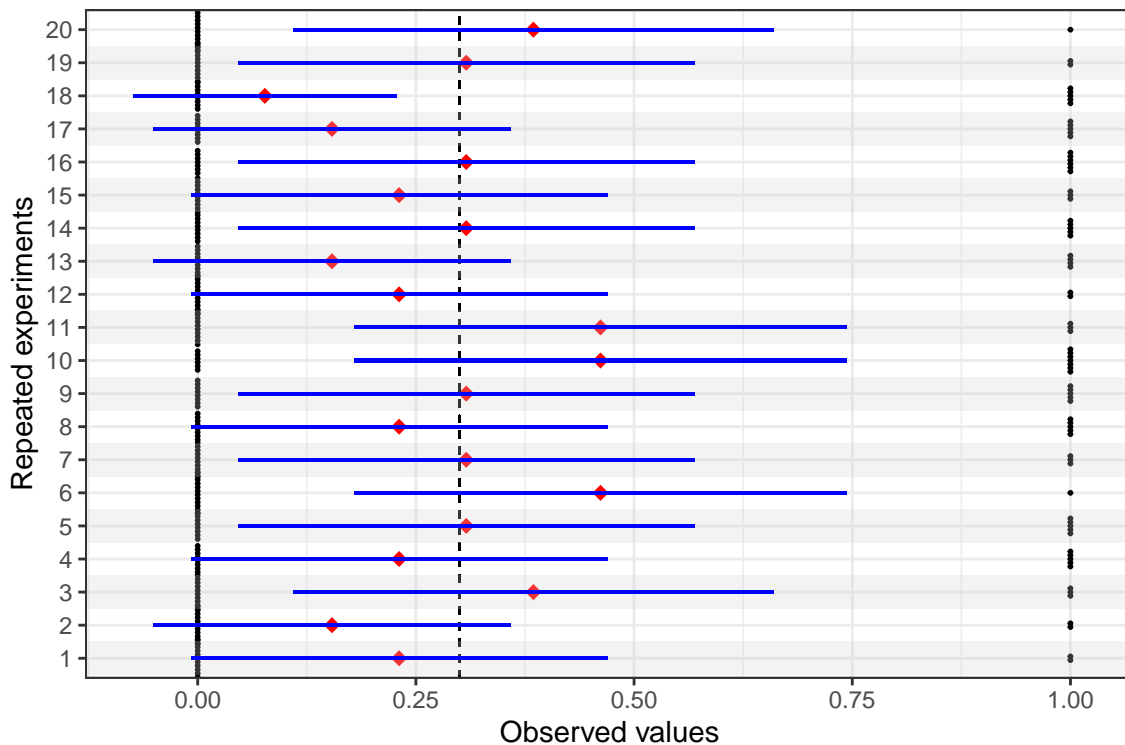


Figure 3.8: Graphical representation of the concept of confidence interval. The black dots are the simulated values for the observed Bernoulli data, in each of 20 replicates of the experiment. The red diamonds indicate the computed sample mean for each replicate. The dashed vertical line is drawn in correspondence of the “true” value of the parameter  $\theta = 0.3$ . The blue lines indicate the 95% confidence intervals computed using the procedure described above.

For 19 out of the 20 potential replicates of the experiments presented in Figure 3.8, the procedure for computing the confidence interval succeeds in covering the true underlying value of  $\theta = 0.3$ , shown as the dashed vertical line.

However, there is one potential replicate (experiment number 18) in which the procedure fails to cover the true value of the parameter. This is not entirely surprising: the procedure for the confidence interval “gets it right” 19/20 or 95% of the times.

And this is the meaning of the analysis based on the confidence interval: nothing to do to the uncertainty about the true value of the parameter — as mentioned above, there is no such thing in the frequentist paradigm. What we can evaluate is the long-run performance of the procedure.

In more general terms, building on Equation 3.4, considering a statistic  $\hat{\theta} = f(\mathbf{Y})$  used to estimate a parameter  $\theta$  and with sampling distribution described (at least approximately) by a  $\text{Normal}(\theta, \sigma^2/n)$ , we can derive a form of the 95% confidence interval as

$$\left[ \hat{\theta} - 1.96 \frac{\sigma}{\sqrt{n}}; \hat{\theta} + 1.96 \frac{\sigma}{\sqrt{n}} \right]. \quad (3.8)$$

### **i** Confidence intervals

To be precise, the idea of confidence intervals as presented above is central to the purely frequentist approach — in fact the mathematical derivation is due to the work of Neyman, who wrote a landmark paper on this topic in 1937, drawing on the work he had done with Egon Pearson at UCL. The mathematical formulation in purely likelihood terms derives the interval estimates in a fairly similar way, by considering the sampling distribution of the statistic  $f(Y)$  (possibly approximated by a Normal distribution) and then computing the interval as

$$\left[ \hat{\theta} - 1.96 \frac{1}{\sqrt{-\ell''(\hat{\theta})}}; \hat{\theta} + 1.96 \frac{1}{\sqrt{-\ell''(\hat{\theta})}} \right],$$

where  $\ell''(\hat{\theta})$  is the second derivative of the log-likelihood, evaluated at the MLE  $\hat{\theta}$ . The quantity  $-\ell''(\theta)$  is often referred to as the **observed Fisher’s Information**.

For all intents and purposes, often the two procedures tend to return the same numerical value for the confidence interval.

**Example 3.1** (Normal data; unknown mean, known variance). We now turn to a slightly more complex (but extremely useful and used) example. Consider data  $y_1, \dots, y_n \stackrel{iid}{\sim} \text{Normal}(\mu, \sigma^2)$ , e.g. the following  $n = 30$  observations.

```
[1] -3.899644 70.906300 64.799692 48.582441 42.830802 2.448935
[7] 18.350713 6.201232 -41.826326 -19.574827 7.668143 6.037137
[13] 58.216779 13.642063 25.834383 40.493779 19.744473 60.579431
[19] 22.770209 50.304467 -37.099194 14.960456 7.974603 3.912337
[25] 17.916688 87.621813 30.473627 10.823798 42.116799 13.358217
```

The parameters vector is thus  $\theta = (\mu, \sigma)$ . However, imagine for now that we have full knowledge of the **population** standard deviation, e.g.  $\sigma = 32$ . Thus, the only relevant parameter (for which we want to make point and interval estimation) is the **population** mean,  $\mu$ .

We can compute the MLE by considering the likelihood function. Recalling Equation 2.4, we can write the Normal sampling distribution for  $n$  iid variables  $\mathbf{Y} = (Y_1, \dots, Y_n)$  as the product of the individual distributions for each  $Y_i$  (this results comes about thanks to the assumption of *independence*), as

$$\begin{aligned}
p(\mathbf{Y} \mid \mu, \sigma) &= \prod_{i=1}^n p(Y_i \mid \mu, \sigma) = p(Y_1 \mid \mu, \sigma)p(Y_2 \mid \mu, \sigma) \cdots p(Y_n \mid \mu, \sigma) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^n \frac{(Y_i - \mu)^2}{\sigma^2}\right).
\end{aligned}$$

Thus, the likelihood function for  $\mu$ , given  $(\mathbf{Y}, \sigma)$  is

$$\mathcal{L}(\mu \mid \mathbf{Y}, \sigma) = \exp\left(-\sum_{i=1}^n (Y_i - \mu)^2\right), \quad (3.9)$$

from which we can derive the log-likelihood

$$\ell(\mu) = \log\left(\exp\left(-\sum_{i=1}^n (Y_i - \mu)^2\right)\right) = -\sum_{i=1}^n (Y_i - \mu)^2.$$

Making use of the facts that: *i.* for a variable  $x$  and a constant  $a$ , the derivative of  $(a - x)^2$  is  $-2(a - x)$ ; *ii.* the derivative of a sum is the sum of the derivatives; *iii.*  $n\bar{Y} = \sum_{i=1}^n Y_i$ ; and *iv.*  $\sum_{i=1}^n \mu = n\mu$ ,

we can compute the first derivative

$$\ell'(\mu) = -\left(-2 \sum_{i=1}^n (Y_i - \mu)\right) = 2(n\bar{Y} - n\mu) = 2n(\bar{Y} - \mu). \quad (3.10)$$

Setting this to 0 and solving for  $\mu$ , we get

$$\begin{aligned}
2n(\bar{Y} - \mu) &= 0 \Rightarrow 2n\bar{Y} = 2n\mu \\
&\Rightarrow \hat{\mu} = \bar{Y}.
\end{aligned}$$

Once again, the MLE for the *population* mean is the *sample* mean  $\bar{Y}$ . Numerically, given the sample above, the MLE is  $\bar{y} = 22.872$ .

In order to compute the 95% confidence interval around this point estimate, we can make use of simple probability calculus and prove that if  $n$  variables  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Normal}(\mu, \sigma^2)$ , then

$$\sum_{i=1}^n X_i = X_1 + \dots + X_n \sim \text{Normal}(n\mu, n\sigma^2).$$

In addition, we can prove that for any constant  $a$ , if  $X \sim \text{Normal}(\mu, \sigma^2)$ , then  $aX \sim \text{Normal}(a\mu, a^2\sigma^2)$ .

Putting these two results together implies that

$$\bar{Y} = \sum_{i=1}^n \frac{Y_i}{n} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right), \quad (3.11)$$

i.e. the sampling distribution of the sample mean is Normal with mean equal to the underlying population mean (which makes it unbiased) and variance equal to the population variance  $\sigma^2$ , rescaled by the sample size  $n$ . We have already used this results when computing the confidence interval for the Bernoulli case seen above.

In the present case, because we know that  $\sigma = 32$  and  $n = 30$  then  $\sqrt{\frac{\sigma^2}{n}} = \sqrt{34.13} = 5.84$  and thus  $\bar{Y} \sim \text{Normal}(\mu, 34.13)$ . Consequently, using Equation 3.8, we can compute the 95% confidence interval as

$$\begin{aligned} \left[ \bar{Y} - 1.96 \frac{\sigma}{\sqrt{n}}; \bar{Y} + 1.96 \frac{\sigma}{\sqrt{n}} \right] &= [22.872 - 1.96(5.84); 22.872 + 1.96(5.84)] \\ &= [11.42; 34.323]. \end{aligned}$$

**Example 3.2** (“Normal data; unknown mean and variance”). The case in Example 3.1 is obviously unrealistic: it is fairly difficult to imagine that we *know* with absolute certainty the value of a population parameter. More likely, we may observe data that we are willing to associate with some Normal sampling distribution, without knowing the underlying “true” value for  $\theta = (\mu, \sigma^2)$ . We may still be mainly interested in point and interval estimates for the population mean  $\mu$ , but this time we consider the case where also the population variance  $\sigma^2$  is unknown.

The first extra complexity is that now, when computing the likelihood function we have something that depends on 2 parameters:

$$\mathcal{L}(\mu, \sigma \mid \mathbf{Y}) = \sigma^{-n} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \mu)^2 \right).$$

The log-likelihood is

$$\begin{aligned} \ell(\mu, \sigma) &= \log \left( (\sigma^2)^{-n/2} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \mu)^2 \right) \right) \\ &= -\frac{n}{2} \log \sigma^2 - \sum_{i=1}^n \frac{(Y_i - \mu)^2}{2\sigma^2} \end{aligned}$$

and we now need to compute two first derivatives: one with respect to  $\mu$  and the other with respect to  $\sigma^2$ . These are

$$\ell'(\mu) = \frac{2}{2\sigma^2} \sum_{i=1}^n (Y_i - \mu) \quad (3.12)$$

$$\ell'(\sigma^2) = -\frac{n}{2\sigma^2} + \sum_{i=1}^n \frac{(Y_i - \mu)^2}{2\sigma^4}. \quad (3.13)$$

Equation 3.12 is basically identical to Equation 3.10 and so the computation of  $\ell'(\mu)$  is similar to what shown in the case for known variance. As for Equation 3.13, we make use of the facts that: i.  $\frac{1}{\sigma^n} = \sigma^{-n} = (\sigma^2)^{-n/2}$ ; ii. for a variable  $x$ , the first derivative of  $\log x$  is  $\frac{1}{x}$ ; iii. for function  $f(x)$ , the first derivative of  $\frac{a}{f(x)}$  is the ratio  $[af'(x)]/f(x)^2$ ; and iv. the derivative of a sum is the sum of the derivatives.

Now, setting Equation 3.12 to 0 and solving for  $\mu$  gives rise to

$$\hat{\mu} = \bar{Y}$$

exactly as in the case for known variance of Example 3.1. Setting Equation 3.13 to 0 and solving for  $\sigma^2$  gives

$$\begin{aligned} -\frac{n}{2\sigma^2} + \sum_{i=1}^n \frac{(Y_i - \mu)^2}{2\sigma^4} = 0 &\Rightarrow -n\sigma^2 + \sum_{i=1}^n (Y_i - \mu)^2 = 0 \\ &\Rightarrow \sigma^2 = \sum_{i=1}^n \frac{(Y_i - \mu)^2}{n}. \end{aligned}$$

Because we do not know the true value for  $\mu$ , in order to materially compute the MLE estimator  $\hat{\sigma}^2$  of  $\sigma^2$  we need to replace it with our best estimate, i.e. the MLE  $\bar{Y}$  and so the MLE estimate for  $\sigma^2$  is

$$\hat{\sigma}^2 = \sum_{i=1}^n \frac{(Y_i - \bar{Y})^2}{n}$$

The crucial part of the procedure shown in Section 3.2.2 to compute the 95% confidence interval is given in Equation 3.4, which defines the sampling distribution for the statistic of interest. More specifically, for  $f(\mathbf{Y}) = \bar{Y}$  as an estimator of the population mean  $\mu$  (which is the case of interest here), then we can re-write

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1),$$

as seen in Example 3.1.

The only difficulty here is that in order to use the Normal(0,1) sampling distribution for  $\bar{Y}$ , we would need to know the *true* value of the other parameter  $\sigma^2$ . But we do not — the best we can do is to actually plug in an estimator. As seen above, the MLE is  $\hat{\sigma}^2$ . However, it can be proved that this is a *biased* estimator, i.e.  $E[\hat{\sigma}^2] \neq \sigma^2$ . Conversely, we can prove that the estimator

$$S^2 = \frac{n}{n-1} \hat{\sigma}^2 = \sum_{i=1}^n \frac{(Y_i - \bar{Y})^2}{n-1} \quad (3.14)$$

is unbiased and so, from a frequentist point of view, is preferred. This is one of the cases where the MLE is actually not optimal (at least from a purely frequentist point of view).

Using  $S^2$  as the best proxy to the unknown (and actually unknown-able!)  $\sigma^2$ , we obtain the statistic

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}},$$

which however is **not** associated with a standard Normal sampling distribution. However, expanding on Equation 2.9, we can derive the general result that

$$\frac{f(\mathbf{Y}) - E[f(\mathbf{Y})]}{\sqrt{\hat{\text{Var}}[f(\mathbf{Y})]}} \sim t(0, 1, n-1), \quad (3.15)$$


where  $\hat{\text{Var}}[f(\mathbf{Y})]$  indicates the “best” sample estimate of the underlying variance. Finally, we can use this result to derive the sampling distribution for  $T$

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t(0, 1, n-1).$$

Thus, to compute the confidence interval in this case we need to use tail-area probabilities from a  $t(0, 1, n-1)$  distribution — instead of the standard Normal we have used so far. Using R we can compute the quantiles of the  $t(0, 1, n-1)$  distribution as

which are numerically equal to  $q_L = -2.045$  and  $q_U = 2.045$ . Considering that the numerical value in the observed sample for  $S^2$  is 897.2, we can then use these values to derive the 95% confidence interval for the sample mean as

$$\begin{aligned} \left[ \bar{Y} - 2.045 \frac{S}{\sqrt{n}}; \bar{Y} + 2.045 \frac{S}{\sqrt{n}} \right] &= [22.872 - 2.045(5.469); 22.872 + 2.045(5.469)] \\ &= [11.69; 34.06]. \end{aligned}$$

 Warning

Notice that in this case, we happen to obtain data for which the sample variance  $S^2$  is actually smaller than the true underlying value  $\sigma^2$ . Of course, in real life we would never know the true value of  $\sigma^2$  and thus we could not make this assessment. And more importantly, we ought to use the t-version of the computation of the confidence interval, to account for the extra layer of uncertainty in the estimates.





## Statistical testing

Statistical testing is a process aimed at using statistical modelling and the available evidence to falsify claims about an alleged data generating process (DGP); a typical example is to assume a working hypothesis that in a comparison between two interventions the true population difference in effect is 0 — that is the new treatment is no better (and no worse) than the standard of care. In a nutshell, the point of statistical testing is to aim at rejecting this hypothesis (which corresponds to a specific DGP). Medical research is probably one of the research areas in which testing has historically played a pivotal role, e.g. in the design and conduct of clinical trials, as we will see later.

Arguably statistical testing can be seen as central to the Frequentist and the Likelihood paradigm. In fact, some of the major arguments between Neyman (especially) and Pearson on the one hand and Fisher on the other hand have centered around the philosophy underlying the testing procedures. As for the Bayesian approach, while suitable theory and methodology exists (which we only briefly mention in Section 4.1), it is perhaps less central to the overall paradigm. One of the reasons for this is that, by its own nature, the Bayesian analysis allows direct probabilistic statements on all the unknown features of the DGP

### ! Confusion

If you found the distinction between the three main schools of statistical thought discussed in the context of estimation confusing, things are even more subtly complex in terms of testing. The philosophy underlying the mathematical construction of the testing procedure is fundamentally different under the frequentist and the likelihood approach, as we will discuss later. But too often, the two have been seamlessly conflated into a unified theory — even in the common case of design and analysis of clinical trials (Goodman 1999). For this reason, it is important to realise the main features of each approach and appreciate the intrinsic distinctions, advantages and disadvantages. In addition, recently there has been a marked shift in the scientific community, who have started to recognise and promote estimation over testing. This has also been caused by the backlash following the so-called “reproducibility crisis” linked to the (mis-)use of  $p$ -values, which has led to a position paper (Wasserstein, Lazar, et al. 2016), discussing the potential pitfalls of practices that are too heavily focussed on testing.

### 4.1 Bayesian approach

We show here a very simple example of Bayesian testing, based on a real analysis performed by Pierre Simone Laplace, a French mathematician who in the late 19th century made great contributions to several scientific disciplines, including Statistics. Despite its almost artificial simplicity, this example illustrate

how we can use the full posterior distribution to make probabilistic statements about underlying DGPs, effectively using inference to perform a variety of testing.

**Example 4.1** (Female births in Paris). We consider here the famous data analysed by Laplace on the number of female births in Paris. In 1710, [John Arbuthnot](#), a Scottish medical doctor with a passion for mathematics, analysed data on christening recorded in London between 1629 and 1710 to conclude that males were born at a “significantly” greater rate than females. This being somehow against the assumption of equal probability for the two sexes, he deduced that divine providence accounted for it, because males die young more often than females.

Laplace analysed similar data collected in Paris from 1745 to 1770. He observed a total of  $r = 241\,945$  girls born out of a total of  $n = 493\,527$  babies and was interested in estimating the probability of a female birth,  $\theta$ . Laplace based his analysis on a reasonable Binomial model for the data:  $r \mid \theta \sim \text{Binomial}(\theta, n)$  and pragmatically assigned a Uniform distribution in  $[0; 1]$  to  $\theta$ :  $p(\theta) = 1$ . This assumption is meant to encode complete lack of knowledge on the model parameter — the only thing this prior is implying is that  $\theta$  has to be between 0 and 1 (which is of course true, as it is a probability). But we are assuming that *any* value in this range is equally likely. Although this is perhaps a rather unrealistic assumption, it simplifies computation of the posterior distribution.

With a little of algebra, it is possible to show that in this case  $\theta \mid r, n \sim \text{Beta}(r + 1, n - r + 1)$ . We can use R to quantify the posterior probability that  $\theta > 0.5$  (i.e. that the true data generating process does rely on a higher chance of a newborn being a male) as

```
# Data: number of girls (r) out of the total births (n)
# in the time interval considered
r=241945
n=493527
# Computes the tail area probability under the posterior distribution using
# simulations from the posterior
nsim=10000
sims=rbeta(n=nsim,shape1=r+1,shape2=n-r+1)
# Tail-area probability
sum(sims>=0.5)/nsim
```

[1] 0

So, given the model assumptions and the observed data, the working hypothesis that  $\theta \geq 0.5$  has essentially no support whatsoever. Laplace calculated analytically that  $\Pr(\theta \geq 0.5 \mid r, n) = 1.15 \times 10^{-42}$  and thus concluded that he was “morally certain” that it was in fact less than 0.5, in accordance with Arbuthnot’s finding.

#### **i** Bayes Factors and the strength of the evidence

Bayesian testing can get much more complicated than shown here. In principle, the main idea is to enumerate an exhaustive list of competing DGPs, indexed by a specific distributional assumption for the parameters  $\theta$ . For example, we may consider  $p(\theta \mid \mathcal{M}_1) \sim \text{Beta}(3, 22)$ ;  $p\left(\log\left(\frac{\theta}{1-\theta}\right) \mid \mathcal{M}_2\right) \sim \text{Normal}(0, 10)$ ;  $p(\theta \mid \mathcal{M}_3) \sim \text{Uniform}(0, 1), \dots$  — notice that for  $\mathcal{M}_2$ , we are specifying a prior on a different scale (technically, this is the *logit* transformation, which we will see in Chapter 5). Each of these generative models  $\mathcal{M}_1, \dots, \mathcal{M}_K$  can be associated with a prior distribution  $p(\mathcal{M}_k)$ , for  $k = 1, \dots, K$ . Bayesian updating can be then applied given the data to obtain an estimate of the posterior distribution for each of the hypotheses being analysed,  $p(\mathcal{M}_k \mid \mathbf{y})$ .

Once the posterior distributions are available, we can compute the *Bayes factor* for model  $k$  versus model  $j$  (for  $k, j = 1, \dots, K$ ) as

$$\begin{aligned} \text{BF}_{kj} &= \frac{p(\mathbf{y} \mid \mathcal{M}_k)}{p(\mathbf{y} \mid \mathcal{M}_j)} \\ &= \frac{\int p(\theta \mid \mathcal{M}_k) p(\mathbf{y} \mid \theta, \mathcal{M}_k) d\theta}{\int p(\theta \mid \mathcal{M}_j) p(\mathbf{y} \mid \theta, \mathcal{M}_j) d\theta} \end{aligned}$$

A rule of thumb to interpret the observed value of the BF is provided by Jeffreys who suggested the following interpretation.

- $\text{BF} \in [1; 3.2)$ : the strength of the evidence for  $\mathcal{M}_k$  against  $\mathcal{M}_j$  is “not worth more than a bare mention”;
- $\text{BF} \in [3.2; 10)$  the strength of the evidence for  $\mathcal{M}_k$  against  $\mathcal{M}_j$  is “substantial”;
- $\text{BF} \in [10; 32)$  the strength of the evidence for  $\mathcal{M}_k$  against  $\mathcal{M}_j$  is “strong”;
- $\text{BF} \in [32; 100]$  the strength of the evidence for  $\mathcal{M}_k$  against  $\mathcal{M}_j$  is “very strong”;
- $\text{BF} > 100$  the strength of the evidence for  $\mathcal{M}_k$  against  $\mathcal{M}_j$  is “decisive”.

In reality, the computation for the BF is complicated for two reasons:

1. In order to derive the *marginal* distribution of the data given the postulated model, we need to *integrate out* the uncertainty about  $\theta$ , described by the posterior distribution. This is essentially akin to computing some weighted average of the full models for the observed data (as a function of the prior for the models as well as the priors for the parameters within each model). As shown in the equation, this involves the computation of generally very complex integrals, thus making this calculations hard to perform.
2. Even if we could easily make this computation, the underlying assumption here is that we are able to enumerate all the competing DGPs. And that one of the models described by  $\mathcal{M}_1, \dots, \mathcal{M}_K$  is indeed the *truth* — which we have no real means of ensuring.

Bayesian testing remains one of the most complex parts of the whole approach. And, as mentioned above, it is perhaps fair to say that, by nature, the Bayesian approach is more focused on a purely estimation context, given that the output of the posterior distribution can indeed be used to make direct probabilistic statements, as in Example 4.1. More details can be found for example in Spiegelhalter, Abrams, and Myles (2004), Gelman et al. (2013) and Kruschke (2014).

## 4.2 Likelihood approach: “significance testing”

Fisher’s approach to testing could arguably be seen as an extension to his estimation procedure, based on the likelihood function (you will see this extensively in both STAT0015 and STAT0016). The most basic structure of the problem is to consider a *working* hypothesis that represents some putative data generating process. Typically, this is some kind of “null” model, implying for instance that there is no meaningful difference in the effect of two competing interventions being tested. We call this the **null hypothesis** and we indicate it as  $H_0$ .

For example we may consider a “**two-sample problems**”, where we randomise  $n_0$  individuals to intervention 0 (say, standard of care) and  $n_1$  individuals to intervention 1 (say, an innovative drug). This is a very common set up and you will encounter it repeatedly in STAT0015, STAT0016 and, to some extent, in STAT0019 too.

The total number of individuals in the study is  $n = n_0 + n_1$ . The data comprise of the two variables  $\mathbf{y}_0 = (y_{10}, \dots, y_{n_00})$  and  $\mathbf{y}_1 = (y_{11}, \dots, y_{n_11})$ , where the subscript “10” indicates the first individual in the “null” intervention arm (i.e. standard of care), while the subscript “11” indicates the first individual in the active intervention arm (the new drug). We could label our data more compactly using the notation  $y_{ij}$ , where  $j = 0, 1$  indexes the treatment arm and  $i = 1, \dots, n_j$  indexes the individuals in each.

If the outcome is some continuous, symmetric quantity, we may be willing to describe the sampling variability in the two samples as

$$y_{10}, \dots, y_{n_0} \stackrel{iid}{\sim} \text{Normal}(\mu_0, \sigma_0^2) \quad \text{and} \quad y_{11}, \dots, y_{n_1} \stackrel{iid}{\sim} \text{Normal}(\mu_1, \sigma_1^2),$$

In a case such as this, we may specify the null hypothesis as  $H_0 : \mu_1 = \mu_0$ , or alternatively (and equivalently!)  $H_0 : \delta = \mu_1 - \mu_0 = 0$  — i.e. that there is “no treatment effect” at the population level.

Given the sample data, we can produce estimates for the main model parameters. For example, as shown in Chapter 3, the MLE for  $\mu_j$  is the sample mean  $\bar{Y}_j = \sum_{i=1}^{n_j} \frac{Y_{ij}}{n_j}$ , with observed value  $\bar{y}_j$ . Using the results shown in Equation 3.11, we can prove that

$$\bar{Y}_j = \sum_{i=1}^{n_j} \frac{Y_{ij}}{n_j} \sim \text{Normal} \left( \mu_j, \frac{\sigma_j^2}{n_j} \right).$$

Moreover, taking advantage of the mathematical properties of the Normal distribution, we can prove that if  $X_1 \sim \text{Normal}(\mu_1, \sigma_1^2)$  and  $X_2 \sim \text{Normal}(\mu_2, \sigma_2^2)$  *independently*, then

- $X_1 + X_2 \sim \text{Normal}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ ;
- $X_1 - X_2 \sim \text{Normal}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$ .

Thus, we can construct an estimator for the difference in the treatment effect  $\hat{\delta} = D = \bar{Y}_1 - \bar{Y}_0$  and derive that

$$\begin{aligned} D = \bar{Y}_1 - \bar{Y}_0 &\sim \text{Normal}(\mu_D, \sigma_D^2) \\ &\sim \text{Normal} \left( \mu_1 - \mu_0, \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \right) \end{aligned}$$

If we knew the underlying values for the two population variances ( $\sigma_0^2, \sigma_1^2$ ), then we could use Equation 3.4 and derive that

$$\frac{D - \delta}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}} \sim \text{Normal}(0, 1). \quad (4.1)$$

Obviously, we are not likely to have this information and thus we can first provide an estimate of  $\sigma_D^2$  using the sample variance

$$\begin{aligned} S_D^2 &= \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1} \\ &= \frac{n_0}{n_0 - 1} \sum_{i=1}^{n_0} (Y_{i0} - \bar{Y}_0)^2 + \frac{n_1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{i1} - \bar{Y}_1)^2 \end{aligned}$$

expanding on the result shown in Equation 3.14 and then derive

$$T = \frac{D - \delta}{\sqrt{S_D^2}} \sim t(0, 1, n - 1), \quad (4.2)$$

adapting the result shown in Equation 3.15.

Fisher’s idea on testing is that we can actually use this result to compute some measure of how much support the observed data give to the null hypothesis (in this case that  $\delta = 0$ ).

For instance, suppose that the data observed are as described by the histograms and summary statistics shown in Figure 4.1.

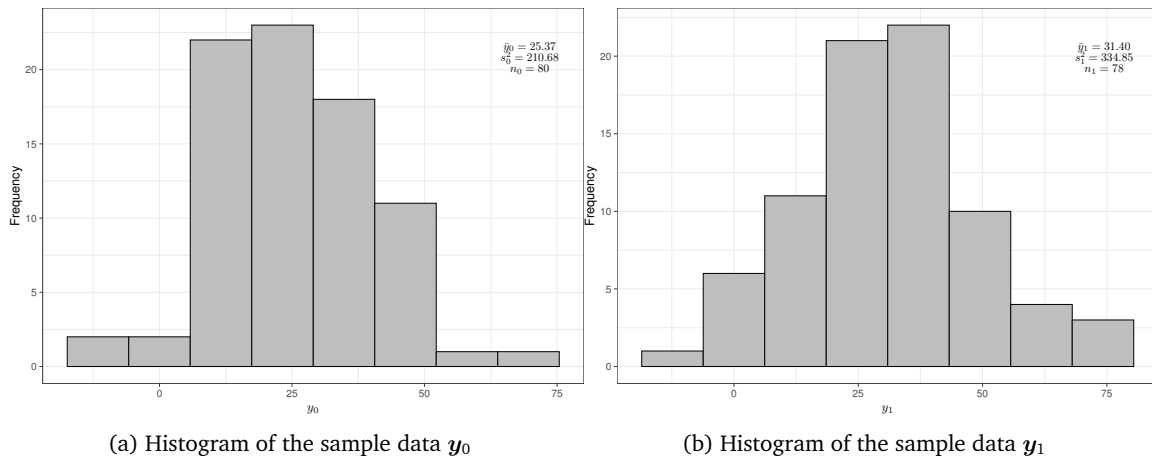


Figure 4.1: Graphical summary of the observed data

Given the sample values and estimates

$$\begin{aligned} n_0 &= 80; & n_1 &= 78; \\ \bar{y}_0 &= 25.37; & \bar{y}_1 &= 31.40; \\ S_0^2 &= 210.68; & S_1^2 &= 334.85, \end{aligned}$$

we can compute the observed value for the estimate of the true difference in the two means  $d = 6.02$  and the estimate for its variance

$$\begin{aligned} s_D^2 &= \frac{s_0^2}{n_0} + \frac{s_1^2}{n_1} = \frac{210.68}{80} + \frac{334.85}{78} \\ &= 2.63 + 4.29 = 6.93. \end{aligned}$$

**Under the null hypothesis**, the *observed* test statistic is

$$t = \frac{d - 0}{\sqrt{s_D^2}} = \frac{6.02}{2.63} = 2.29.$$

Fisher thought that *if* the null hypothesis were true, then we would expect the observed value  $t$  of the test statistic  $T$  to be fairly “central” in the range of its sampling distribution. In other words, he suggested using tail-area probabilities under the sampling distribution  $p(t \mid \theta, H_0)$  to find the chance of observing a result “*as extreme as, or even more extreme than*” the one that actually obtained in the current data. Intuitively, if  $t > \text{Med}(T)$ , i.e. the observed value of  $t$  is “larger than normal”, we would be looking at even larger values; conversely, if  $t \leq \text{Med}(T)$ , “even more extreme” values would be in fact smaller than that observed.

In the present case, we can use R to compute the *p-value*  $\Pr(T > t \mid \theta, H_0)$

```
# Computes the tail-area probability under the sampling distribution under H_0
# NB: the option 'lower.tail=FALSE' computes the area to the *right* of the
# observed value t
pt(q=t,df=n-1,lower.tail=FALSE)
```

```
[1] 0.01170658
```

**i** They didn't have a computer...

Equation 4.1 and Equation 4.2 construct the relevant test statistics using a re-scaling of the quantity of interest  $D$ . This is essentially a historical accident — the reason for this is that even if we could determine that  $D$  was associated with a Normal sampling distribution, in practical terms, without computers it was difficult (although not impossible) to compute probabilities for a non-standard Normal (i.e. one with mean different than 0 and variance different than 1). Conversely, computations with a standard Normal (or, for that matter, with a  $t$  with mean equal to 0 and variance equal to 1) are much simpler to perform by hand, which Fisher had pretty much to do. With modern computers, we do not really worry about re-scaling the relevant quantity  $D$  to the standardised test statistics  $Z$  and  $T$ . But because re-scaling does not really cost much in computation terms (in fact, hardly anything at all!), this procedure has stuck and we still use it.

The  $p$ -value is the area shaded to the right of the observed  $t = 2.29$ , in Figure 4.2 (note that the value of  $t$  is greater than the median of the distribution and so we need to look for the tail-area in the right). Because the probability of observing something *as extreme as, or even more extreme than* the data we have actually got in front of us is extremely small, we deem that the data provide very little support to the working hypothesis of no difference in the population means.

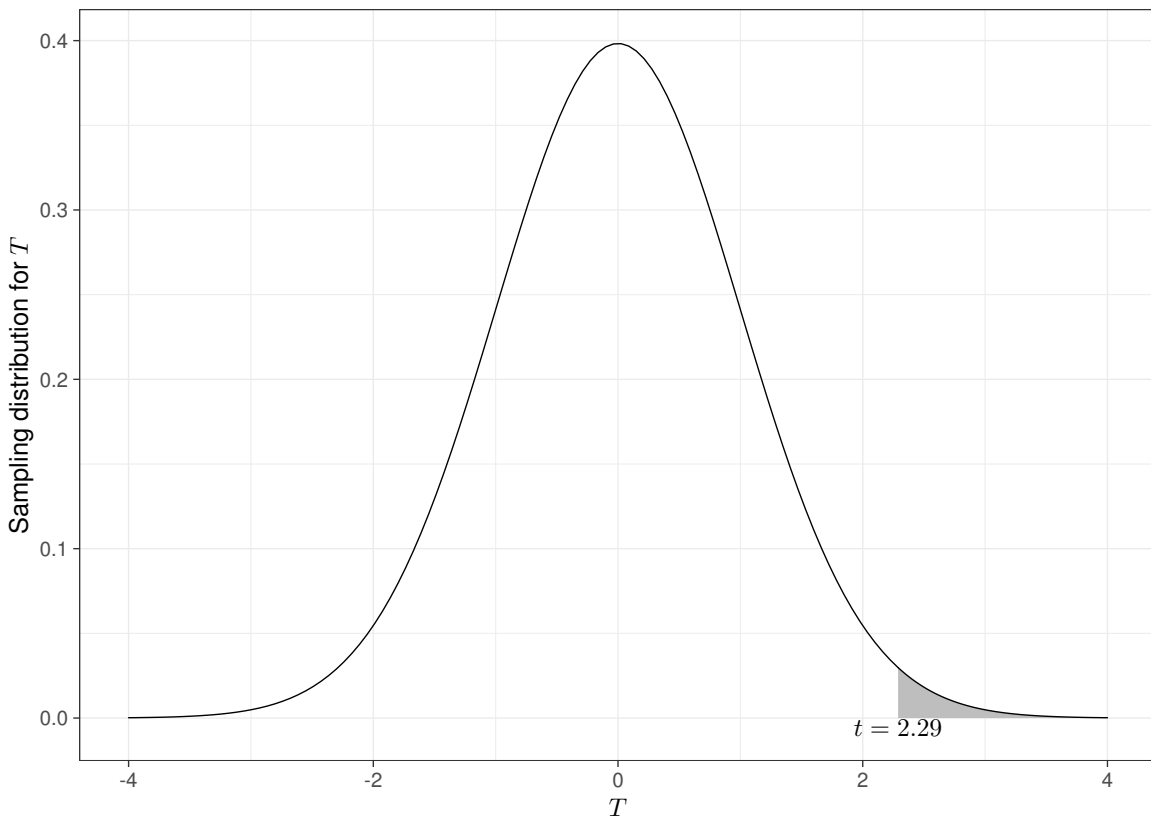


Figure 4.2: The sampling distribution for the statistic  $T$ . The shaded area indicates the  $p$ -value

### ! p-values and continuous distributions

The interpretation of the  $p$ -value in terms of assessing the probability of observing something *as extreme as, or even more extreme than* the data actually observed (or, in other words, as a tail-area probability) is clearly a mouthful and a slightly subtle concept.

The reason why Fisher had to use such a convoluted phrasing is that, when dealing with continuous data, the sampling distribution does not really represent a probability, but rather a density. So, it is impossible to quantify the strength of the evidence **just in correspondence of the observed value for the test statistic**.

One criticism associated with this (inevitable!) choice is that the strength of the evidence (or, in other words: whether the observed data are consistent with the working hypothesis) depends on the data actually observed, as well as on data (“even more extreme”) that might be — but have not been — observed.

In addition to this, by its own nature, it is possible that the same  $p$ -value is computed for a very small and a very large study; the conclusion in terms of strength of the evidence would be identical, without distinction of the underlying sample size (Goodman 1999).

In a similar way to Jeffreys, Fisher also provided some rule of thumb to interpret the  $p$ -value  $P$  observed from the data at hand.

- If  $P < 0.01$ , then conclude that there is **strong** evidence against  $H_0$ ;
- If  $0.01 < P < 0.05$ , then conclude that there is **fairly strong** evidence against  $H_0$ ;
- If  $P > 0.05$ , then conclude that there is **little or no** evidence against  $H_0$ , or alternatively, that the observed data are *consistent* with the model specified in  $H_0$ .

Of course, there is nothing special about the value 0.05 (5%), which is effectively used as the threshold for **statistical significance**. And more importantly, can a dataset giving rise to a  $p$ -value of 0.0499 really be considered as substantially different than one giving rise to a  $p$ -value of 0.0501? Yet, for a very long time, medical and psychology journals in particular have obsessed over the “quest for significance”, at times refusing to publish studies with results indicating a  $p$ -value above 0.05 as irrelevant.

### i Confidence intervals and p-values

Suppose we test a null hypothesis  $H_0 : \mu = \mu_0$  and find that the  $p$ -value is greater than 0.05. Then the 95% confidence interval for  $\mu$  will include the hypothesised value  $\mu_0$ . If  $P < 0.05$  then the 95% confidence interval will *not* include  $\mu_0$ . In other words:

*The 95% confidence interval for  $\mu$  consists of all hypothesised values  $\mu_0$  for which the  $p$ -value is greater than 0.05.*

Thus, if you calculate a 95% confidence interval that does not include a  $\mu_0$  of interest, then you can infer that the  $p$ -value will be less than 0.05. Likewise, if a 99% confidence interval does not include  $\mu_0$ , the  $p$ -value will be less than 0.01. Note, though, that there is a *logical* distinction between estimation and testing.

## 4.3 Frequentist approach: “hypothesis testing”

The main idea underlying Neyman and Pearson’s (NP’s) approach to testing is that, in fact, this is not an inferential problem, but rather a decision-making one.

The rationale in NP’s approach is that the researcher does not really believe in  $H_0$  — if we thought that a new drug did not really have any difference over something that already exists, what would be the point in investing money, time and research in developing it?  $H_0$  is just a working hypothesis that we would

like to discard, or in technical parlance *reject*, given the observed data (and the modelling assumptions we are making!).

For instance, what the researchers would probably truly believe is that in fact intervention 1 is more effective than intervention 2 (i.e., assuming that the higher the population mean, the better the health condition, that  $\mu_1 > \mu_0$ ). Thus, we can also specify an **alternative hypothesis**  $H_1 : \mu_1 > \mu_0$  or, equivalently  $H_1 : \delta > 0$ , to indicate that the new intervention is better.

### ! Unlikely events

One important and perhaps subtle feature of the testing procedure is the distinction between the *idealised* (null) hypothesis and the *empirical* evidence we want to use to disprove it.

- The null and the alternative hypotheses are defined in terms of the **population** parameters. So by considering  $H_0 : \mu_1 = \mu_0$  we are assuming that the “true” average intervention effects are identical.
- However, we will never be in a position of observing the population parameters. All we can do is get some sample data and then use suitable statistics to estimate the underlying parameters and then say something about whether these are equal or not. So it is perfectly possible that, just by chance, we observe data that *look like* they could be drawn by a common generating process where the two population means are the same. But in fact the underlying, true DGP may be characterised by different means (perhaps where the difference is only little).

For this reason, NP have framed their hypothesis testing problem in terms of the decision made about whether or not the null hypothesis is the underlying truth, based on the current data. Table 4.1 shows this schematically (and you will see more on this in STAT0015 and STAT0016).

Table 4.1: The decision problem underlying NP’s theory of hypothesis testing

	$H_0$ true	$H_0$ false
Reject $H_0$	Type I error $\alpha$ (False positive)	Correct inference (True negative)
Fail to reject $H_0$	Correct inference (True positive)	Type II error $\beta$ (False negative)

If the true “state of nature” was that there is in fact no difference in the intervention effects (labelled as “ $H_0$  true”), if the data make us *reject*  $H_0$ , then we would be making an error. NP call this “Type I error”, indicated as  $\alpha$  — this is essentially a “false positive”, because we would be erroneously claiming that there is a difference in the intervention effects, when in fact there is not. Conversely, if the data make us fail to reject  $H_0$ , then we would be making the correct decision. The Type I error is also usually (and rather confusingly!) referred to as “*significance level*”.

Similarly, if the true “state of nature” was that the new intervention is more effective (i.e. there is a difference in the population means), then the situation is reversed: if we have enough evidence to reject the null hypothesis, then we would have made the correct decision. But if the data were not indicating that we should reject  $H_0$ , then we would be making an error. NP call this the “Type II error”  $\beta$ , which can be thought of as a “false negative” result, because we would conclude that there is no difference, when in fact one is present.

In particular, in a typically frequentist fashion, we can reason along the following lines: *if we were able to replicate the experiment (data collection) over and over again, under the same experimental conditions* we would have the following situation.

- If  $H_0$  is indeed the true DGP, then we would make the wrong decision in a proportion of times equal to  $\alpha$ , while we would make the correct decision in the complementary proportion of times  $1 - \alpha$ .



- If  $H_1$  is indeed the true DGP then we would make the wrong decision in a proportion of times equal to  $\beta$ , while we would make the correct decision in the complementary proportion of times  $1 - \beta$ .

The researcher is given the choice to tune the two unknown probabilities: NP’s suggestion (which has essentially become some kind of dogma in many areas of applied research) is that it is good to keep the probability of making a Type I error  $\alpha$  to a very low value, typically 1 in 20, or 0.05. This means that *if we were able to do the experiment a very large number of times under the same conditions* if  $H_0$  is the true state of nature, on average we would correctly fail to reject the null hypothesis 95% of the times (approximately 1 every 20 replicates).

As for the Type II error, NP suggested that, somewhat arbitrarily, the researcher could live with a less stringent requirement and typically  $\beta$  is fixed at either 10% or 20%. This means that if there truly is an intervention effect (i.e.  $H_1$  is the “truth” and thus the two population means are different), we are happy to mistakenly claim the opposite result in 10-20% of the times. Intuitively, the rationale for this imbalance in the values of  $\alpha$  and  $\beta$ , can be explained as follows. Most likely, the “status quo” intervention (with population mean  $\mu_0$ ) will be established and probably a “safe option”. The new intervention may be very good and improve health by a large amount. But of course we are not sure, because perhaps the data are limited in scope and follow up. Thus, we want to safeguard against making claims that are too enthusiastic about the potential benefits of the new intervention — that is why we keep the Type I error probability to a low value. Conversely, although it is bad to miss out on claiming that the new intervention is in fact beneficial, we are more prepared to run this risk — and that is why  $\beta$  is typically higher than  $\alpha$ .

The graph in Figure 4.3 visualises the ideas underlying the procedure for hypothesis testing. For the sake of argument, imagine that the true DGP is characterised by a Normal distribution where the population mean and variance are  $\mu = 0.01$  and  $\sigma^2 = 0.85^2$ .

For simplicity (and unrealistically!), we assume that we, the researchers performing the analysis, do know the true value of the population variance, while the population mean is unknown *to us*. We then define the null hypothesis as  $H_0 : \mu = \mu_0 = 0$  in contrast with the alternative  $H_1 : \mu = \mu_1 = 0.05$ . As is almost invariably the case, the alternative hypothesis does **not** represent the “truth”, but simply a proposed model in which the treatment does have a (clinically meaningful) effect — and in this particular case, the true treatment effect  $\mu$  happens to be closer to  $H_0$  than it is to our posited alternative DGP indexed by  $H_1$ . Also, we assume that data are observed for  $n = 1250$  individuals and these are modelled as  $y_1, \dots, y_n \stackrel{iid}{\sim} \text{Normal}(\mu, \sigma)$ .

Suppose that we decide to consider as test statistic the sample mean  $\bar{Y}$ . Given the model assumptions specified above and recalling Equation 3.11, we know that

$$\bar{Y} \sim \text{Normal} \left( \mu = 0.01, \frac{\sigma}{\sqrt{n}} = \frac{0.85}{35.36} = 0.024 \right) \quad (4.3)$$

(of course, in reality, we would not know the true values for the parameters, but remember that in this example, we are pretending to be some kind of Mother Nature figure, who knows all...).

By plugging in the values for the hypothesised means  $\mu_0$  and  $\mu_1$ , we can derive the sampling distributions under the two competing hypotheses

$$p(\bar{y} \mid \theta, H_0) = \text{Normal}(0, 0.024) \quad \text{and} \quad p(\bar{y} \mid \theta, H_1) = \text{Normal}(0.05, 0.024).$$

Notice that, unlike the model in Equation 4.3, even without the gift of being Mother Nature, we are in general able to determine these two distributions, because they depend directly on our model assumptions and the observed data, which allow us to estimate the relevant parameters.

The two sampling distributions are shown in Figure 4.3 as the blue and red curve, respectively. The dark grey area under the blue curve represents the tail-area probability under the null sampling distribution, which we have constructed to be equal to  $\alpha$ . Similarly, the light grey area under the red curve represents

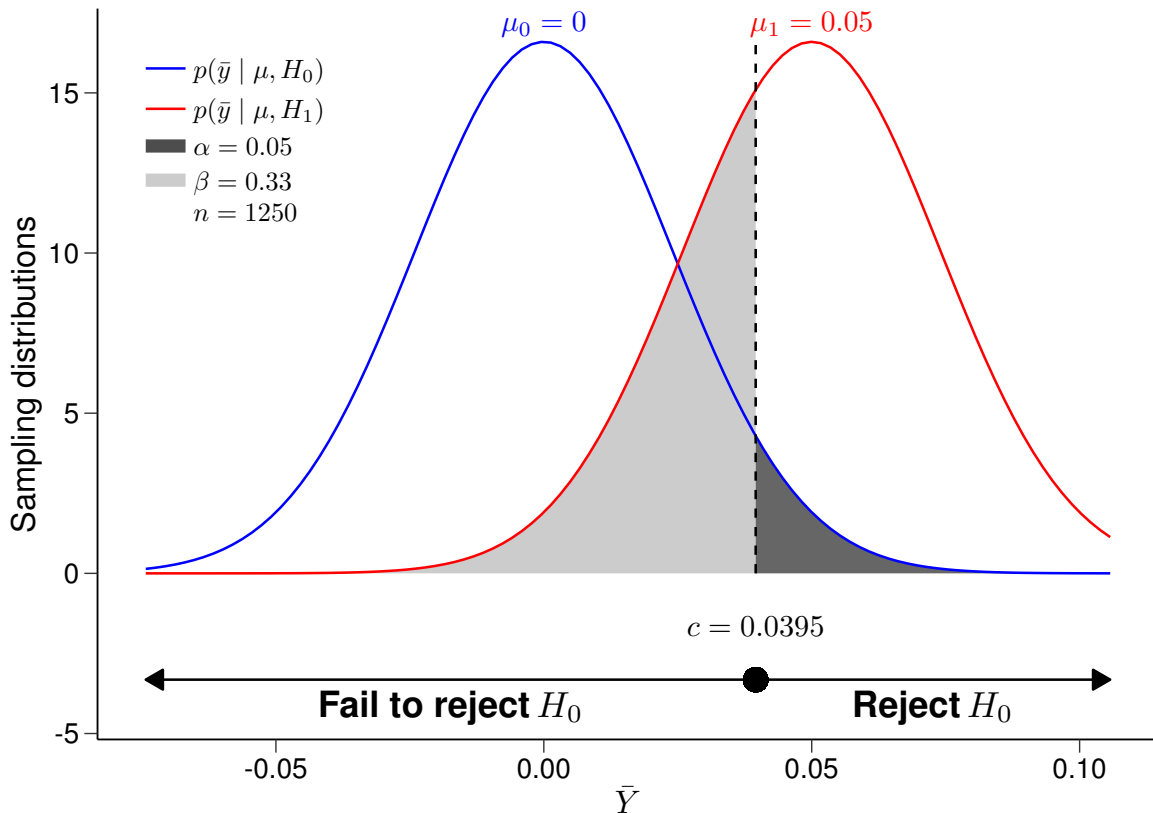


Figure 4.3: A graphical depiction of the decision-making problem in hypothesis testing

the tail-area probability under the alternative sampling distribution, equal to  $\beta$  ( $= 0.33$ , given the model assumptions specified in this case).

The 95% quantile of the sampling distribution under  $H_0$ , computed as  $c = 0.0395$  in the present case, is indicated as the large dot above the  $x$ -axis in Figure 4.3. NP refer to this as the “critical value”, because it determines the “critical” (or “rejection”) region: if the observed value of the test statistic determined by the data lies in the rejection region, then the data give more support to the sampling distribution under  $H_1$ . Intuitively, we can see this by considering that for each point in the region labelled as “Reject  $H_0$ ” in Figure 4.3 (i.e. the dark grey area), the density is higher under the red, than under the blue curve. If, on the contrary, the observed value of the test statistic does not lie in the rejection region (i.e., in this case is less than the critical value), then we do not have enough evidence to reject the null hypothesis. This is the decision rule underpinning the procedure of hypothesis testing:

*If the observed value of the test statistic is in the critical region (which is determined by the model assumptions and the observed data), then reject  $H_0$ . If not, we fail to reject  $H_0$ .*

#### ! Fisher vs Neyman-Pearson

Superficially, Fisher’s procedure of significance testing and NP’s hypothesis testing, look rather similar. They both depend on defining some null hypothesis (typically indicating the lack of treatment effect, i.e. meaningful difference between two mean responses), which we do not really believe in but would like to be able to falsify; then they both involve the definition of some test statistic, for

which we could determine a sampling distribution, based on the model assumptions; then they both involve computing the value of the test statistic.

Where they differ (substantially!) however, is that significance testing outputs a  $p$ -value, which is a numerical summary of the “strength of the evidence” against  $H_0$ . We can make a decision based on the  $p$ -value, as outlined above — so if it is very small, we can safely reject  $H_0$ , while if it is borderline the set significance level (e.g. 0.05), our assessment will be much less certain.

On the contrary, NP’s procedure has a strictly binary outcome: whether the observed value of the test statistic falls *just above* the critical value, or it is much, much larger, it is actually irrelevant for the decision problem: in both cases, we would reject  $H_0$ .

These two interpretations are often conflated — and  $p$ -values tend to be evaluated under this strict binary decision rule.

Another important distinction between significance and hypothesis testing is that in the former case, we do not take explicit notice of the alternative DGP described by  $H_1$ , which is a central part to NP’s theory. The main advantage of the fact that we do consider an alternative mechanism to generate the data is that, given the significance level  $\alpha$  and the sample size of the dataset we wish to use to compute the test statistic, we can also assess the chance of making a Type II error, which we have indicated as  $\beta$ , above. For example, for  $\alpha = 0.05$  and  $n = 1250$  as shown in Figure 4.3, the resulting Type II error is  $\beta = 0.33$ . This automatically allows us to compute the “**power**” of the test statistic,  $1 - \beta$ , which indicates the probability of rejecting  $H_0$ , when it is false.

We can use a simulation approach to better understand this concept. Given the model assumptions above, we can use R to simulate a large number of potential replicates of the experiment (i.e. the data collection), assuming that  $H_0$  is false. For example, we could use the code below.

Figure 4.4 shows a graphical summary of the results for this simulation, when we use a sample size of  $n = 1250$  and consider  $n_{\text{sims}} = 5000$  replicates. The  $x$ -axis indexes the value of the simulated test statistic  $\bar{Y}$  that is obtained for each replicate of the experiment — for better visualisation, these have been sorted from the lowest (-0.036) to the highest value (0.15). Each dot in the plot is the computed value of  $\bar{Y}$ . The dashed vertical line indicates the critical value  $c = 0.0395$ , which in turns determines the rejection region.

As is possible to see, when we set  $\alpha = 0.05$  and  $n = 1250$ , the percentage of simulations for which we *correctly* manage to reject  $H_0$  (remember that we are simulating data from the alleged DGP under  $H_1$ , which means that  $H_0$  is false!) is about 68%, in line with the theoretical outcome shown in Figure 4.3.

### **i** Likelihood ratio tests

As is customary within the frequentist approach, the choice of a possible test statistic depends on a set of long-run properties that define its optimality. There are of course very many potential choices that can be made — the discussion of these properties is beyond the scope of these notes.

In many practical cases, it can be proved that among all tests for a specified level  $\alpha$  and for a set sample size  $n$ , the **likelihood ratio test** statistic, defined as

$$\Lambda(\mathbf{Y}) = \frac{\mathcal{L}(\theta | \mathbf{Y}, H_0)}{\mathcal{L}(\theta | \mathbf{Y}, H_1)}$$

is the one with the highest power and this is often regarded as an optimal property, thus justifying the use of this test statistic in many applied cases. Intuitively, the likelihood ratio is small if the alternative model is better than the null model.

In particular, it can be proved that, at least approximately,

$$-2 \log \Lambda(\mathbf{Y}) \sim \text{Chi-squared}(\nu),$$

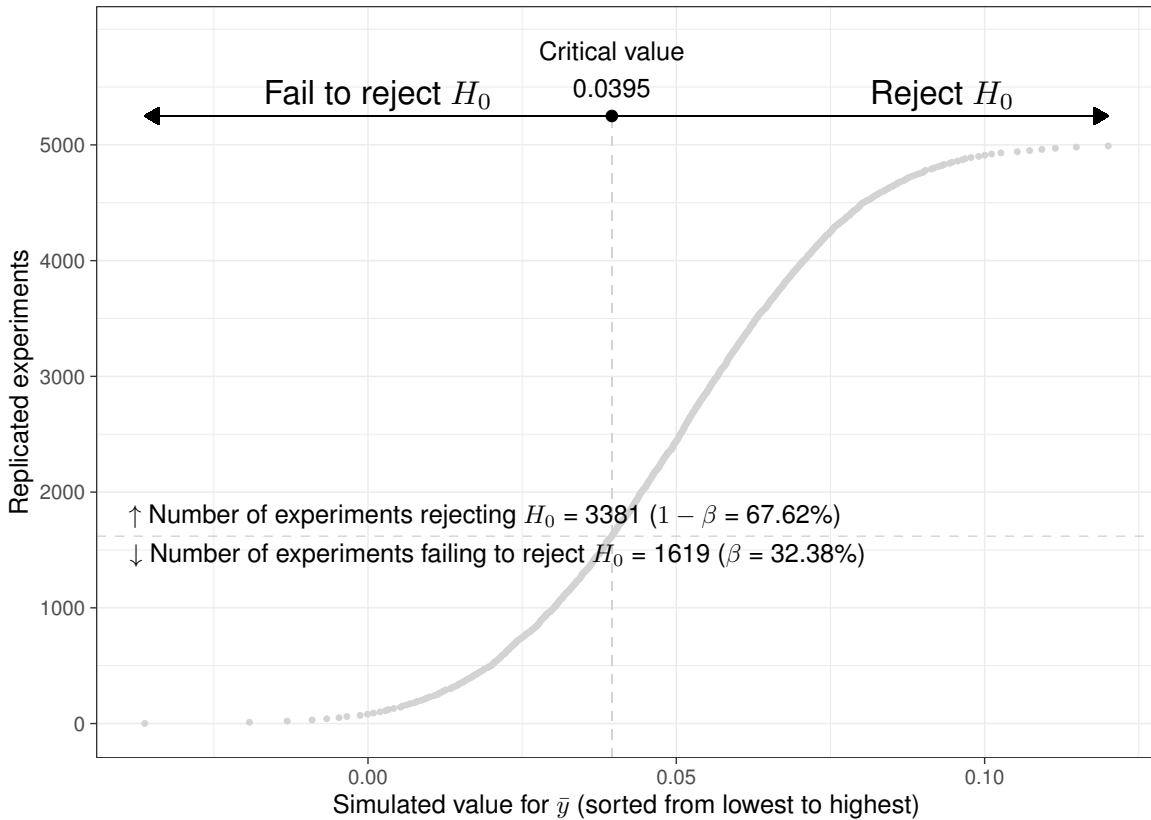


Figure 4.4: Output of 5000 simulations for data of sample size  $n = 1250$ , generated from  $H_1$

where the degrees of freedom are determined by the type of the hypotheses being tested.

If we define  $\chi_{(1-\alpha)}(\nu)$  as the  $100(1 - \alpha)\%$  quantile of the Chi-squared( $\nu$ ) distribution, then the (2-log)likelihood-ratio test provides the decision rule as follows:

- If  $-2 \log \Lambda(\mathbf{Y}) > \chi_{(1-\alpha)}(\nu)$ , then reject  $H_0$
- If  $-2 \log \Lambda(\mathbf{Y}) \leq \chi_{(1-\alpha)}(\nu)$ , then fail to reject  $H_0$ .

More details can be found for example in Casella and Berger (2002).

Because of the obvious relationship between the sample size  $n$  and the level of precision with which we can estimate the model parameters, which essentially determines the variance of the observed data  $\mathbf{Y}$  and therefore of the test statistic  $f(\mathbf{Y})$ , we can then use the power for two important purposes:

1. As a tool to determine the “best” test, for a given significance level and set sample size.
2. As a tool to determine the “optimal” sample size we need to observe to be able to constrain the Type II error, given a fixed significance level.

Figure 4.5 shows an example of the analysis shown above for several choices of the underlying sample size, in this case  $n = 20, 100$  and  $2500$ . As is possible to see, the standard deviation for the sampling distribution of the test statistic  $\bar{Y}$ , which in this case is given by  $\frac{\sigma}{\sqrt{n}}$ , does decrease for increasingly large values of the sample size  $n$ . This implies that a different critical value  $c$  would be computed in each of the three cases depicted in panels (a)—(c).

When the sample size is very small, then the two sampling distributions under  $H_0$  and  $H_1$  are very close together, because the means  $\mu_0$  and  $\mu_1$  are also assumed to be fairly close and the standard deviation is

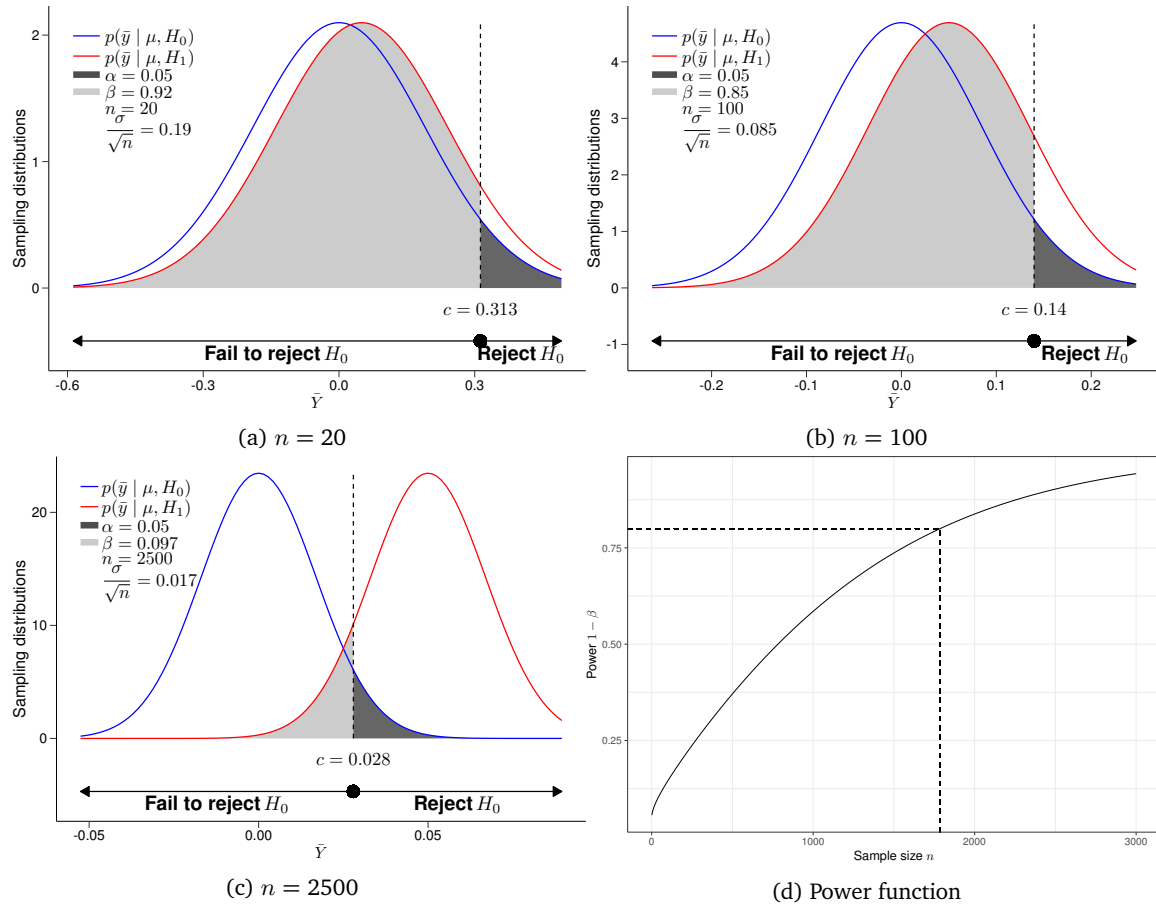


Figure 4.5: The relationship between power and sample size

relatively large. Intuitively, this means that it will be difficult to tell them apart, which in turns means that we are very much prone to mistakes in the decision-making (and so we would fail to correctly reject  $H_0$  a large proportion of times — in fact about 92% of the times). Thus, for such a small sample size, the power is very low.

But when we increase the sample size, we also tend to increase the power of the test, i.e. its ability to correctly detect a signal, while keeping fixed its ability to safeguard from claiming a false significant results. By plotting the power resulting for a range of possible sample sizes, we get the graph in panel (d), which is often referred to as the **power function** of the test.

At the stage of *designing* a study, we could use the power function to determine the sample size that is necessary so that the resulting test has a significance level (i.e. Type I error)  $\alpha$  and a power  $1 - \beta$ , for fixed values of  $\alpha$  (i.e. almost invariably 0.05) and  $\beta$  (e.g. 0.2). In the current case, given all the model assumptions, we would need to observe at least  $n = 1787$  individuals to be able to detect a difference of  $\mu_1 - \mu_0 = 0.05$  with 80% power — this is indicated by the vertical dashed line in panel (d) of Figure 4.5.

**!  $\alpha$  vs  $P$**

The sample size calculation, or power analysis, shown above highlights another subtle idiosyncrasy of the “combined” Fisher/NP approach (which, as we have mentioned above, neither of the original

proponents would have condoned!). Medical research is designed according to NP's precepts, which aim at controlling the Type I and II errors by setting a fixed significance level  $\alpha = 0.05$  and then optimising the sample size  $n$  in order to obtain a power of  $1 - \beta$  (typically 0.8). But then, once the data are actually collected, the analysis is performed under a Fisherian approach, which is based on the calculation of the  $p$ -value.

And despite the fact that the common threshold value for both  $\alpha$  and  $P$  are usually set at 0.05, these are two *fundamentally* different quantities:  $\alpha$  is set from the outset and it is immutable and independent on the data that are actually observed.  $P$ , on the other hand, is determined by the actual data and is meant to provide a graded measure of the strength of the evidence against the null hypothesis (and, notably, without any formal regards to what is the alternative hypothesis).

## 4.4 Commonly used statistical tests

In this section, we present some of the most common statistical tests (with specific reference to the Likelihood approach and the use of  $p$ -value). You will see these in many of the different topics discussed in STAT0015 and STAT0016.

### 4.4.1 Chi-squared test

The Chi-squared test is encountered in many applied cases. One of the most important cases is the comparison between groups of categorical data, e.g. grouped in a *contingency table* such as the one presented below.

Table 4.2: An example of categorical data grouped by treatment arm and clinical output

	Disease cured	Disease not cured	Total
Control arm	13	40	53
Treatment arm	18	29	47
Total	31	69	100

We can rescale the data in Table 4.2 to compute the probability that a random patient is cured from the disease under the two treatment arms  $p_C = \frac{13}{53} = 0.245$  and  $p_T = \frac{18}{47} = 0.383$ . A reasonably relevant null hypothesis would be that there is in fact no association between the disease status at the end of the study period and the treatment arm.

If this null hypothesis is true, then we would expect that the total number of patients who are cured from the disease (31) would be re-proportioned in the two groups simply according to the overall sample size observed in them (i.e. 53 and 47, respectively), without any differential impact of the treatment. Thus, under  $H_0$ , we would expect to see

$$E_{CC} = 53 \times \frac{31}{100} = 16.43$$

patients who are cured from the disease in the control arm and

$$E_{CT} = 47 \times \frac{31}{100} = 14.57$$

in the treatment arm — the notation  $E_{CC}$  and  $E_{CT}$  is meant to convey the concept of *expected* outcome among those Cured by the disease in the Control and Treatment arm, respectively, under  $H_0$ . Using a similar reasoning we can compute

$$E_{DC} = 53 \times \frac{69}{100} = 36.57$$

and

$$E_{DT} = 47 \times \frac{69}{100} = 16.43,$$

which indicate the *expected* number of patients who are still Diseased in the Control or Treatment arm, respectively, under  $H_0$ .

We can construct a test statistic that aims at comparing the *observed* data to the *expected* ones, under the null model as

$$T = \sum_{i=(C,D)} \sum_{j=(C,T)} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}},$$

where  $Y_{ij}$  are the values in the cells of the contingency table and for which the observed sample value is

$$\begin{aligned} t &= \frac{(13 - 16.43)^2}{16.43} + \frac{(40 - 36.57)^2}{36.57} + \frac{(18 - 14.57)^2}{14.57} + \frac{(29 - 32.43)^2}{36.57} \\ &= 0.716 + 0.322 + 0.807 + 0.363 \\ &= 2.21. \end{aligned}$$

Recalling Equation 2.10, we can prove that (at least approximately)  $T \sim \text{Chi-squared}(\nu)$ , where the degrees of freedom  $\nu$  are computed as  $(J - 1) \times (I - 1)$ , where  $J$  and  $I$  are respectively the number of rows and columns in the contingency table. Thus, in this case we have that  $\nu = (2 - 1) \times (2 - 1) = 1$ . We can use this information to compute a  $p$ -value to measure the strength of the evidence against the null hypothesis of no association between the treatment and the probability of curing the disease as the tail-area probability under a Chi-squared distribution with 1 degree of freedom, for example using the following R commands.

```
# Defines the number of rows and columns of the contingency table
I=J=2
# Constructs the contingency table
Y=matrix(c(13,40,18,29),byrow=T,nrow=I,ncol=J)
# Constructs the matrix of expected counts
E=matrix(NA,J,I)
for (i in 1:I) {
  for (j in 1:J) {
    E[j,i]=sum(Y[j,])*sum(Y[,i])/sum(Y)
  }
}
# Computes the test statistic
t=sum(((Y-E)^2)/E)
# Computes the p-value based on the Chi-squared((J-1)(I-1)) distribution
pchisq(q=t,df=((J-1)*(I-1)),lower.tail=FALSE)
```

```
[1] 0.1372945
```

In this case, the  $p$ -value is substantially greater than 0.05 and therefore we cannot reject the null hypothesis of no association. An alternative (more compact) way of performing the Chi-squared test is to use the built-in function `chisq.test`.

**i** One vs two sided tests

As mentioned above, the  $p$ -value computed using the  $T$  statistic and the tail-area of the Chi-squared distribution does not account for what is the alternative — it is just the probability of getting something *as extreme as, or even more extreme* than the observed value  $t$ , if  $H_0$  is true.

Often, however, available software programmes also conflate Fisher’s with NP’s interpretation and, while returning a  $p$ -value as the main output of the calculation, they will also offer the user the option to select one of three possible alternative hypotheses. For instance, there is another built-in function named `prop.test` that can be used to perform a Chi-squared test on proportions. This includes the option `alternative` that takes by default the value `"two.sided"`, implying that the alternative hypothesis  $H_1$  assumes that the true parameter  $\theta$  simply is different than the value specified under  $H_0$ ,  $\theta_0$ . In addition to that, `prop.test` also allows the user to specify the values `"less"` or `"greater"`, which imply alternatives in the form  $\theta < \theta_0$  or  $\theta > \theta_0$ , respectively.

This is essentially against the very nature of the  $p$ -value approach: because, to be pedantic, it is defined as the probability of getting a result *as extreme as, or even more extreme than* the one observed, a  $p$ -value is by definition only meant to deal with an implicitly “one-sided” alternative hypothesis. If  $t$  is “large”, then this implies that “more extreme” possible results should be measured on the right tail-area of the sampling distribution under  $H_0$ . If it is “small”, then the left tail-area is to be used for the computation.

In practice, often  $p$ -values are used to report the results of a hypothesis tests, but it is important to understand that this implies some sort of hybrid methodology.

The following example clarifies the issue. Given the data in Table 4.2, we can use the R function `prop.test` to investigate the evidence against the null hypothesis of equality of the population proportions. This is simply done as

```
prop.test(Y,alternative="two.sided",conf.level=0.95,correct=FALSE)
```

2-sample test for equality of proportions without continuity correction

```
data: Y
X-squared = 2.208, df = 1, p-value = 0.1373
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.31861433  0.04322292
sample estimates:
  prop 1  prop 2
0.2452830 0.3829787
```

This configuration assumes the default implicit “two-sided” alternative (i.e. that the population proportions under the treatment and control arms are just different, without specifying a direction for this difference). The option `correct=FALSE` does not correct for the fact that the actual data are discrete counts, while the Chi-squared distribution is continuous (this is a technicality and only relevant when the observed counts are very small — as a rule of thumb, if there is any cell with a value less than 5, the continuity correction should be applied).

As is possible to see, this command returns the same  $p$ -value shown above,  $P_{\text{two.side}} = 0.137$ , based on the same value of the test statistic (indicated as  $\chi$ -squared in the computer output above). The graph in Figure 4.6 shows a histogram for the distribution of the resulting test statistic in panel (a), as well as the *cumulative distribution function* in panel (b). This shows, for each value on the  $x$ -axis the probability that the test statistic is less than or equal to that value. So, in correspondence of the observed value 2.21,



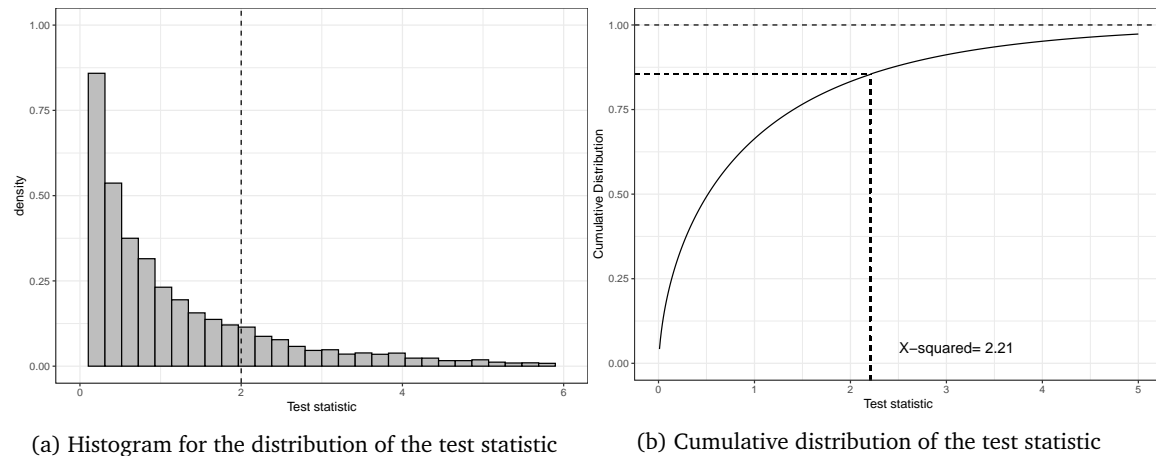


Figure 4.6: Different representations of the test statistic, highlighting the  $p$ -value as the tail-area probability. The histogram shows that most of the density is distributed before the observed value of the test statistic (indicated by the vertical dashed line), but the cumulative distribution plot in panel (b) highlights this more clearly

we can see that the cumulative distribution function is around 0.863 (i.e. the probability that the test statistic is *greater* than this observed value is  $1 - 0.863 = 0.137$ , as reported by the computer outcome).

However, if we use the option `alternative="greater"`, with the idea that in fact we want to consider the alternative hypothesis that the treatment arm is associated with a larger probability of being cured, we can run the command

```
prop.test(Y,alternative="greater",conf.level=0.95,correct=FALSE)
```

2-sample test for equality of proportions without continuity correction

```
data: Y
X-squared = 2.208, df = 1, p-value = 0.9314
alternative hypothesis: greater
95 percent confidence interval:
-0.2895274  1.0000000
sample estimates:
prop 1 prop 2
0.2452830 0.3829787
```

and obtain a different  $p$ -value:  $P_{\text{great}} = 1 - P_{\text{two.side}}/2 = 0.931$ . The rationale here is to essentially split the tail-area probability to account that when the implicit alternative hypothesis is that the parameter is different than the null value, then it may be either smaller, or greater. Assuming that these two are (at least approximately) equally likely, then we can obtain the  $p$ -value associated with the implicit alternative that the parameter is greater than the null as  $= P_{\text{two.side}}/2 (= 0.0686$  in this case).

Again, this interpretation of  $p$ -values is only possible in a combined approach where the design is based on NP's explicit set up of null and alternative hypotheses, but the analysis is based on the graded summary of the evidence provided by the  $p$ -value, rather than the binary reject/fail to reject decision-making approach advocated by NP.

#### 4.4.2 Fisher's exact test

The analysis of Table 4.2 based on the Chi-squared distribution is valid but based on an approximation of the sampling distribution of the test statistic. In fact, Fisher was able to determine this distribution more precisely, by investigating the make-up of the contingency table. He was able to prove that for a generic table of counts such as the one shown in Table 4.3:

Table 4.3: A general example of a contingency table

	Disease	Healthy	Total
Control	$a$	$b$	$a + b$
Treatment	$c$	$d$	$c + d$
Total	$a + c$	$b + d$	$n = a + b + c + d$

the exact distribution of the configuration of values  $(a, b, c, d)$  was:

$$p(a, b, c, d) = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b} \binom{c+d}{d}}{\binom{n}{b+d}}$$

(this is technically a Hypergeometric distribution). We can use this to compute an exact  $p$ -value, by enumerating all the possible configurations in which, for a fixed overall total sample size  $n$ , we would observe a result *as extreme as, or even more extreme than* the one we have actually encountered. The difficulty in the computation is not so much in obtaining the actual probability associated with a given configuration (that Fisher proved how to compute analytically), as much as in the enumeration of all the relevant cases.

For example, for the data in Table 4.2, having fixed the *margins* of the table to  $(53, 47)$  along the rows and  $(31, 69)$  along the columns, “more extreme cases” include the tables where, instead of the observed  $a = 13$ , the first cell has values  $12, 11, \dots, 0$  (notice that by modifying the first cell of the tables implies that we also modify the second cell to respect the row margin, but also the third cell of the table so that the first column total is respected. And by doing this, we are also modifying the fourth cell of the table so that the second row total as well as the second column total are also respected).


We can make this computation using the R built-in function

```
fisher.test(Y)
```

```
Fisher's Exact Test for Count Data
```

```
data: Y
p-value = 0.1935
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2018786 1.3434546
sample estimates:
odds ratio
0.5270614
```

The function outputs several results. The most important one is the  $p$ -value  $P = 0.194$ . This is slightly larger than the one computed using the Chi-squared test. This is a common feature of Fisher's exact test.

 Warning

Despite using the exact sampling distribution under  $H_0$  for the relevant test statistic (as opposed to an approximation based on the Chi-squared distribution), the Fisher's exact test is often criticised as being "too conservative", i.e. as producing  $p$ -values that are artificially too high, over and above the actual strength of the observed evidence against the null hypothesis.

The reason for this feature is that Fisher's exact test is rather technical and has to do with the underlying discrete nature of Fisher's procedure (the Hypergeometric distribution that underpins the test is used to describe the sampling variability of a discrete DGP), which essentially clashes with the significance testing machinery. More technical details are available in Casella and Berger (2002), but the main take-home message is that almost invariably the technically less precise (because it is based on an approximation) Chi-squared test is preferred.

#### 4.4.3 Difference between two proportions

We can expand the idea shown in the analysis of the data for Table 4.2 to the more general case in which we are interested in evaluating formally whether the difference between two proportions is significant. To do this, it is helpful to realise that the data in Table 4.2 could be modelled equivalently by considering

$$r_C \sim \text{Binomial}(\pi_C, n_C = 53) \quad \text{and} \quad r_T \sim \text{Binomial}(\pi_T, n_T = 47),$$

where  $\pi_C$  and  $\pi_T$  are the underlying population proportion of individuals who are cured from the disease, if given the control or the treatment, respectively. A reasonable null hypothesis is  $H_0 : \pi_C = \pi_T$ , to indicate the absence of a treatment effect. This can be described equivalently as  $H_0 : \pi = \pi_T - \pi_C = 0$ .

The MLE for the difference in population proportions is

$$\begin{aligned} \hat{\pi} &= \frac{r_T}{n_T} - \frac{r_C}{n_C} = \hat{\pi}_T - \hat{\pi}_C \\ &= \frac{13}{53} - \frac{18}{47} \\ &= 0.383 - 0.245 = 0.138. \end{aligned}$$

Replicating the argument made for Equation 4.1, the estimate for the variance of the MLE for the proportion difference can be computed as

$$\begin{aligned} \text{Var}[\hat{\pi}] &= \frac{\hat{\pi}_T(1 - \hat{\pi}_T)}{n_T} + \frac{\hat{\pi}_C(1 - \hat{\pi}_C)}{n_C} \\ &= \frac{0.383(1 - 0.383)}{47} + \frac{0.245(1 - 0.245)}{53} \\ &= 0.00503 + 0.00349 = 0.00852 \end{aligned}$$

and so, recalling Equation 3.5, we can then derive that under the null hypothesis the test statistic is associated (at least approximately) with a standard Normal distribution:

$$Z = \frac{\hat{\pi} - 0}{\sqrt{\text{Var}[\hat{\pi}]}} \sim \text{Normal}(0, 1),$$

with an observed value

$$z = \frac{0.138}{0.0923} = 1.49.$$

Thus, the  $p$ -value for the null hypothesis is computed in R using the following code.

```

# Observed data in the two treatment arms
rc=13; nc=53
rt=18; nt=47
# Computes the estimate for the difference in population proportions
pi.hat=(rt/nt)-(rc/nc)
# Estimates the pooled sd
pooled.sd=sqrt(((rt/nt)*(1-rt/nt)/nt)+((rc/nc)*(1-rc/nc)/nc))
# Computes the test statistics
z=(pi.hat-0)/pooled.sd
# And then computes the p-value based on approximate Normality
pnorm(z,lower.tail=FALSE)

```

```
[1] 0.06788723
```

Based on the Normal approximation for the test statistic, we can also compute a 95% confidence interval for the difference in the population proportions as

$$\begin{aligned}\hat{\pi} \pm 1.64\sqrt{\text{Var}[\hat{\pi}]} &= (0.138 - 1.64 \times 0.0923; 0.138 + 1.64 \times 0.0923) \\ &= (-0.0141; 0.29).\end{aligned}$$

The two results are of course consistent with one another. The  $p$ -value is marginally larger than the usual threshold for significance (0.05) and so the 95% confidence interval includes the null value (i.e. a difference of 0).

#### 4.4.4 Wald test

The Wald test is based on defining a test statistic

$$W = \frac{(\hat{\theta} - \theta_0)^2}{\text{Var}[\hat{\theta}]} \sim \text{Chi-squared}(1),$$

where  $\hat{\theta}$  is the best estimate (e.g. MLE) for the parameter of interest,  $\theta$ ;  $\theta_0$  is the value posited under the null hypothesis  $H_0$ ; and  $\text{Var}[\hat{\theta}]$  is the estimate of the variance of the estimator  $\hat{\theta}$ .

We can also prove that

$$\sqrt{W} = \frac{\hat{\theta} - \theta_0}{\sqrt{\text{Var}[\hat{\theta}]}} \approx \text{Normal}(0, 1),$$

in line with the results described above. In fact, there are more complex forms for the Wald statistic for cases where we want to test multiple parameters at once — but you are unlikely to encounter these in the modules you will take during your MSc programme.

#### **i** Wald and Likelihood ratio tests

The Wald test can be seen as an approximation to the Likelihood Ratio test — but one of the advantages is that you do not require two competing hypotheses to compute its value (as you do for the Likelihood ratio test).

The Wald test is often used in a regression analysis context (see Chapter 5), to determine whether a specific covariate (predictor) should be included in the model.

## Regression analysis

Regression analysis is one of the most important tools available to a statistician (which you will see extensively in all the statistical modules in the HEDS Programme). Gelman and Hill (2007) provide an excellent introduction to the main techniques associated with regression analysis.

The main idea of regression is to link a function of some observed *outcome* (often called the “response” variable),  $y$ , to a set of *predictors* (often referred to as “covariates”),  $\mathbf{X} = (X_1, \dots, X_K)$ . Examples include modelling the relationship between some clinical outcome  $y$ , e.g. a measurement for blood pressure, and a set of prognostic factors, e.g. sex, age, comorbidities, ethnic background, etc, which are used to predict the outcome.

### ! Helpful language

Sometimes the terminology “dependent” and “independent” variables is used to describe the outcome and the predictors — this is probably not very helpful though, because it somehow conflicts with the concept of statistical *independence*: if two variables  $X$  and  $Y$  are statistically independent on one another, then learning something about  $X$  does not tell us *anything* about  $Y$ , which is in fact the opposite assumption underlying regression analysis! We therefore do not use this terminology in the rest of the notes.

Assuming we have observed data for  $n$  individuals

$$(\mathbf{y}, \mathbf{X}) = (y_1, X_{11}, \dots, X_{1K}), \dots, (y_n, X_{n1}, \dots, X_{nK}),$$

in its simplest form, we can indicate a regression model as

$$f(y_i | \mathbf{X}_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_K X_{iK}.$$

Even more specifically, the function  $f(\cdot)$  is almost invariably chosen as  $E[Y | \mathbf{X}] = \mu$  and so the **linear regression model** can be written as

$$E[Y_i | \mathbf{X}_i] = \mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_K X_{iK}. \quad (5.1)$$

### i Helpful algebra

The expression in Equation 5.1 can be also defined more compactly and equivalently, using matrix algebra as

$$E[Y | \mathbf{X}] = \mathbf{X}\boldsymbol{\beta},$$

where  $\beta = (\beta_0, \dots, \beta_K)$  is the vector of model **coefficients**, each describing the impact (or **effect**) of the predictors on the outcome — in this case we assume that the matrix of covariates is built as

$$\begin{pmatrix} X_{10} & X_{11} & X_{12} & \cdots & X_{1K} \\ X_{20} & X_{21} & X_{22} & \cdots & X_{2K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{n0} & X_{n1} & X_{n2} & \cdots & X_{nK} \end{pmatrix}$$

and the first column of the matrix  $\mathbf{X}$  is made by a vector of ones, i.e.

$$\begin{pmatrix} X_{10} \\ X_{11} \\ \vdots \\ X_{n0} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

This is needed to ensure that when performing the matrix multiplication

$$\mathbf{X}\beta = \begin{pmatrix} X_{10}\beta_0 + X_{11}\beta_1 + \dots + X_{1K}\beta_K \\ X_{20}\beta_0 + X_{21}\beta_1 + \dots + X_{2K}\beta_K \\ \vdots \\ X_{n0}\beta_0 + X_{n1}\beta_1 + \dots + X_{nK}\beta_K \end{pmatrix}$$

the first element in each row (i.e. for each of the  $n$  individuals) returns  $\beta_0$  as in Equation 5.1.

## 5.1 Regression to the mean

The term “regression” was introduced by [Francis Galton](#)<sup>1</sup>. Among many other topics, Galton worked to study hereditary traits. In particular, he collected data on  $n = 898$  children from 197 families.

The data comprise the height of each child  $y_i$ , as well as the height of the father  $X_{1i}$  and the mother  $X_{2i}$ , for  $i = 1, \dots, n$ , all measured in inches. An excerpt of the data is presented in Table 5.1.

Table 5.1: The first few rows of Galton’s dataset on height of parents and children

Family	Father	Mother	Gender	Height	Kids
1	78.5	67.0	M	73.2	4
1	78.5	67.0	F	69.2	4
1	78.5	67.0	F	69.0	4
1	78.5	67.0	F	69.0	4
2	75.5	66.5	M	73.5	4
2	75.5	66.5	M	72.5	4

<sup>1</sup> Galton was another very controversial character. He was a polyscientist, who made contributions to Statistics, Psychology, Sociology, Anthropology and many other sciences. He was the half-cousin of Charles Darwin and was inspired by his work on the origin of species to study hereditary traits in humans, including height, which he used to essentially invent regression. He also established and then financed the “Galton Laboratory” at UCL, which Karl Pearson went on to lead. Alas, Galton was also a major proponent of eugenics (in fact, he is credited with the invention of the term) and has thus left a troubling legacy behind him.

In its basic format, Galton’s analysis aimed at measuring the level of correlation between the children and (say, for simplicity) the father’s heights.

**Warning**

There are several nuances in Galton’s data structure; for example, many families are observed to have had several children, which intuitively implies some form of correlation within groups of data — in other words, we may believe that, **over and above** their parent’s height, siblings may show some level of correlation in their observed characteristics. Or, to put it another way, that children are “clustered” or “nested” within families. There are suitable models to account for this feature in the data — some of which will be encountered in STAT0016 and STAT0019.

The data can be visualised by drawing a scatterplot, where the predictor (father’s height) is along the  $x$ -axis and the outcome (child’s height) is along the  $y$ -axis. This is presented in Figure 5.1, where each family is labelled by a different colour.

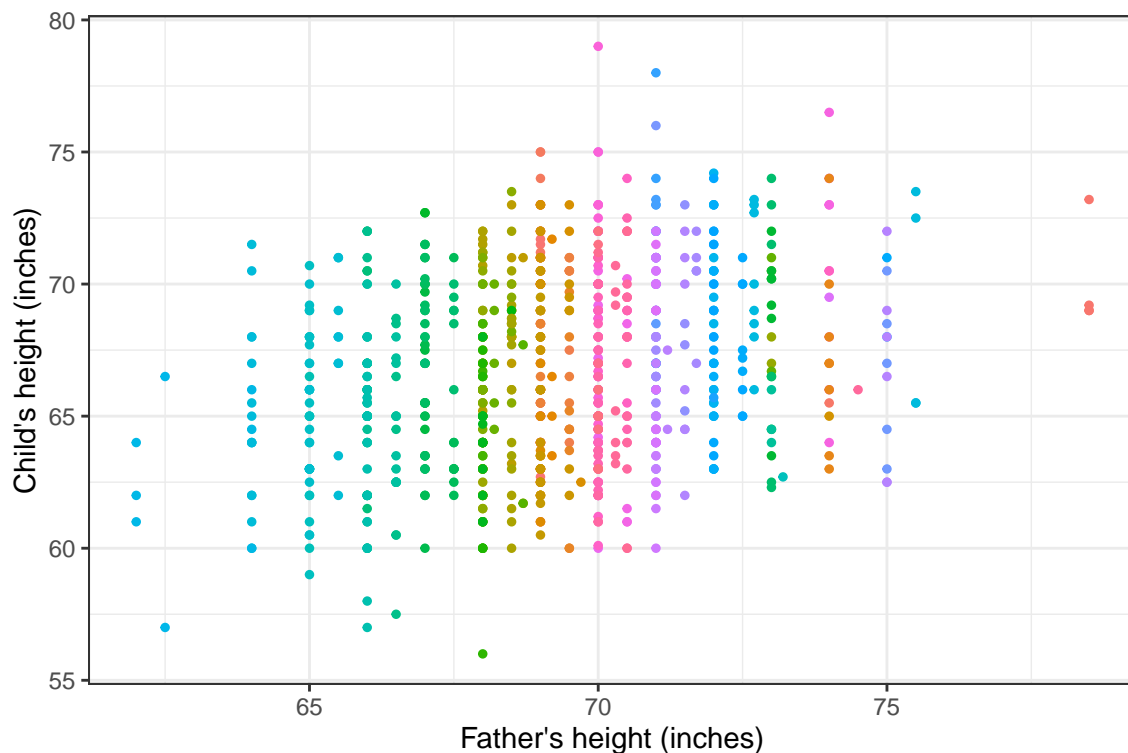


Figure 5.1: A graphical depiction of Galton’s data on father’s and children’s height

Galton’s objective was to find out whether there was some consistent relationship between the outcome and the predictor and, if so, to quantify the strength of this relationship. In developing his original work, he used a technique that was common at the time, called “*least squares fitting*”. The main idea underlying least squares fitting is that we would like to summarise a set of numbers  $y_1, \dots, y_n$ , by using a single value  $a$  and we would like to choose such a value  $a$  so that it is *as close as possible* to all the observed data points. One way to ensure this is to determine  $a$  as the solution of the equation

$$\min_a \left( \sum_{i=1}^n (y_i - a)^2 \right) = \min_a ((y_1 - a)^2 + (y_2 - a)^2 + \dots + (y_n - a)^2). \quad (5.2)$$

The intuition behind the least squares ideas is that by minimising the sum of the square distances between each data point and the value  $a$ , we ensure that the “prediction error” (i.e. the error we make in using the summary  $a$  instead of each true value  $y_i$ ) is as small as possible, overall. If we consider the linear model in Equation 5.1, Equation 5.2 becomes

$$\min_{\beta} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 X_{1i})]^2$$

and, as it happens, we can prove that the optimal values for  $\beta = (\beta_0, \beta_1)$  are

$$\hat{\beta}_1 = \frac{\text{Cov}(y, X_1)}{\text{Var}[X_1]} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}_1,$$

where

$$\text{Cov}(y, X_1) = \frac{1}{n} \sum_{i=1}^n (X_{1i} - \bar{X}_1)(y_i - \bar{y}) \quad (5.3)$$

is the **covariance** between the outcome  $y$  and the covariate  $X_1$ , indicating a measure of joint spread around the means for the pair  $(y, X_1)$ .

If we use the observed data (which we assumed stored in a data frame, named `galton`), we can compute the relevant quantities, e.g. in R using the following commands

```
y.bar=mean(galton$Height)
X1.bar=mean(galton$Father)
Cxy=cov(galton$Height,galton$Father)
Vx=var(galton$Father)
beta1.hat=Cxy/Vx
beta0.hat=y.bar-beta1.hat*X1.bar
```

which return the least square estimates for  $\hat{\beta}_0$  and  $\hat{\beta}_1$  as

$$\hat{\beta}_1 = \frac{2.44}{6.10} = 0.399 \quad \text{and} \quad \hat{\beta}_0 = 66.76 - 0.399 \times 69.23 = 39.11.$$

We can use these estimates to superimpose the **regression line** over the observed data, as shown in Figure 5.2.

We have purposely chosen to show the  $x$ -axis going all the way to the origin ( $x = 0$ ), even though the mass of points is actually very far from this point. This is not surprising — it is not very meaningful to imagine that we may have observed fathers whose height is equal to 0 inches! This highlights two important features when presenting data and, specifically, regression analyses:

1. Plot the underlying data and/or the resulting regression curve on a suitable range. This should be dictated mostly by the scale or range of the observed data;
2. It is often a good idea to re-scale the covariates. A useful re-scaling is to “centre” the covariates, by considering  $X_i^* = (X_i - \bar{X})$ .

This is also related to the interpretation of the regression coefficients (at least for a linear regression), which is why we have used this sub-optimal pictorial representation. Figure 5.2 shows that the coefficient  $\hat{\beta}_0$  identifies the value of the line in correspondence with a covariate set to the value of 0. In other words,



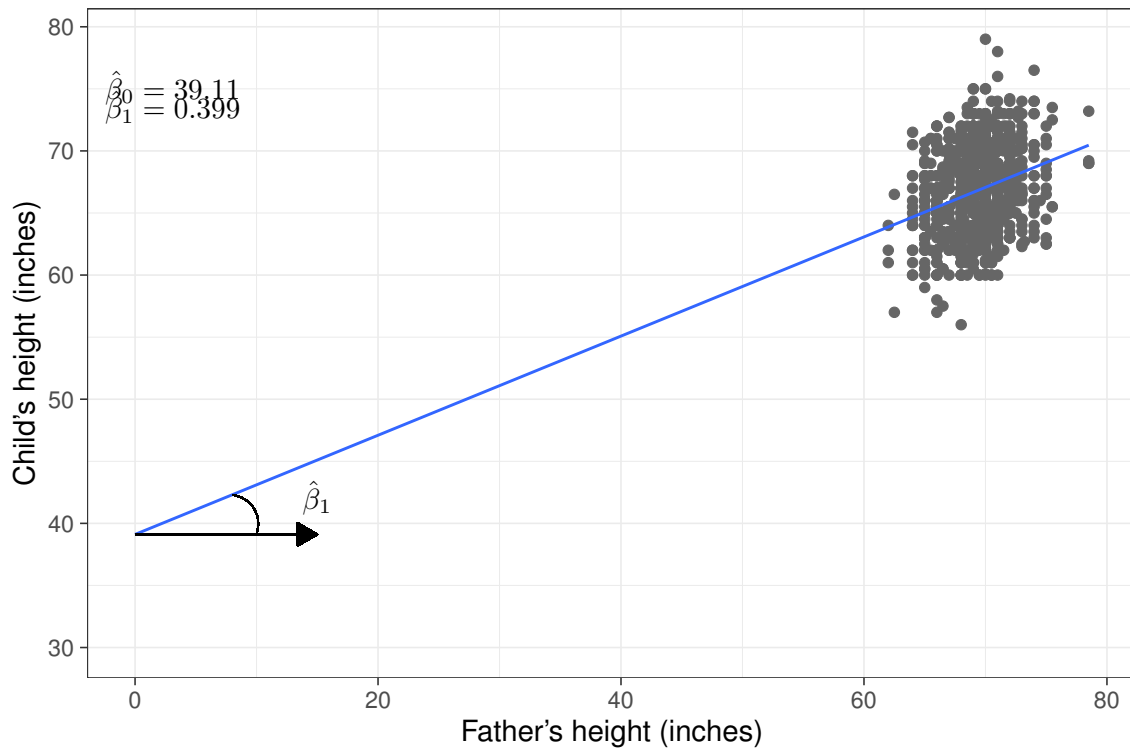


Figure 5.2: Galton's original data with the regression line superimposed

if we considered a father whose height is 0 inches, we would be predicting that his child's height would be, on average,  $\hat{\beta}_0 = 39.11$  inches — i.e. the point along the  $y$ -axis in which the regression line crosses it. This is referred to as the *intercept*.

As for the second regression coefficient  $\hat{\beta}_1$ , the graph in Figure 5.2 shows that this can be interpreted as the *slope* of the regression curve. That is the inclination of the line, as described by the angle shown in the left-hand side of the graph (the arc below the line). The interpretation of this feature is that  $\hat{\beta}_1$  is responsible for how much the line “tilts” — if the estimated value is high, then the line becomes steeper, while if it is low, it becomes more shallow. When  $\hat{\beta}_1 = 0$ , then the regression line is parallel to the  $x$ -axis, indicating the, irrespective of the value of the covariate  $X$ , the expected value for the outcome  $y$  remains unchanged. This essentially indicates that if  $\hat{\beta}_1 = 0$  then  $X$  has no effect on  $y$ .

If we compare two individuals whose  $X$  value differs by 1 unit (e.g. two fathers whose heights differ by 1 inch), the slope indicates the increase in the expected outcome (e.g. the expected height of their respective child). This can be easily seen by considering the linear predictor for two different fathers, say  $F_1$  and  $F_2$ , for whom the recorded heights are, say, 70 and 71 inches, respectively. The expected heights for the children of  $F_1$  and  $F_2$  are then

$$E[Y | X = 71] = \beta_0 + \beta_1 \times 71 \quad \text{and} \quad E[Y | X = 72] = \beta_0 + \beta_1 \times 72.$$

Thus, the difference between these two expected outcomes (i.e. the effect of a unit change in the father's height) is

$$\begin{aligned}
 \Delta_X &= E[Y \mid X = 72] - E[Y \mid X = 71] \\
 &= (\hat{\beta}_0 + \hat{\beta}_1 \times 72) - (\hat{\beta}_0 + \hat{\beta}_1 \times 71) \\
 &= \hat{\beta}_0 + \hat{\beta}_1 \times 72 - \hat{\beta}_0 - \hat{\beta}_1 \times 71 \\
 &= \hat{\beta}_1(72 - 71) \\
 &= \hat{\beta}_1 = 0.399.
 \end{aligned}$$

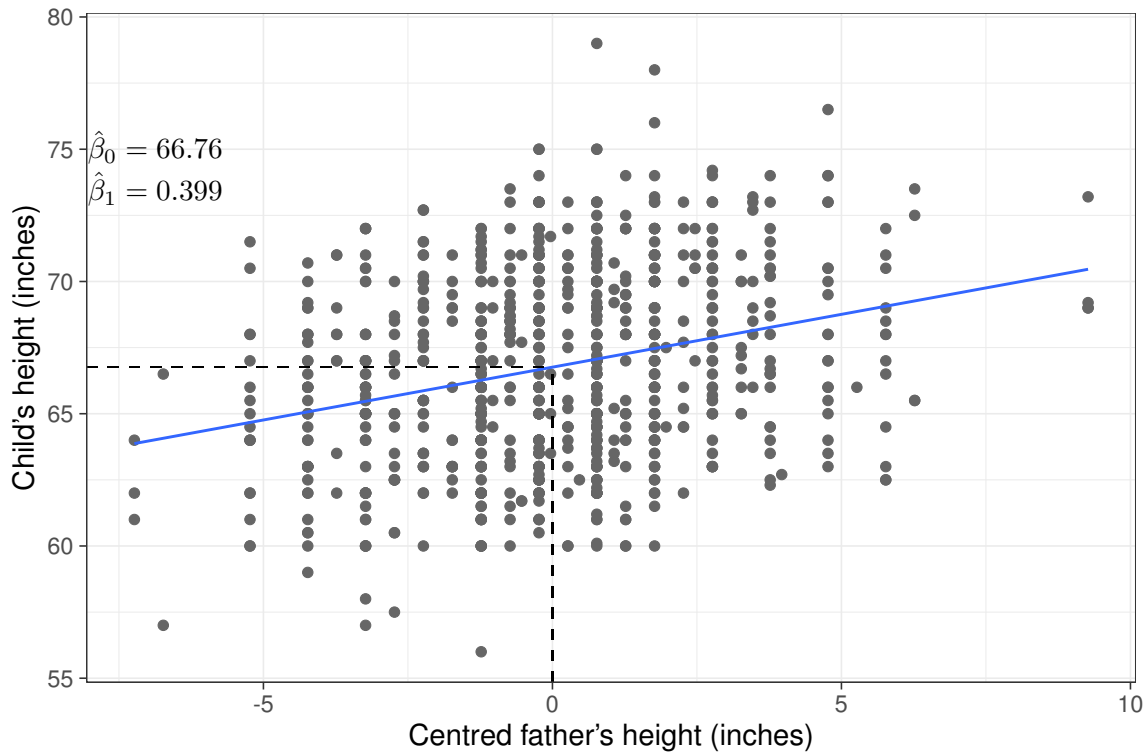


Figure 5.3: Galton’s original data with the regression line superimposed, using the centred version of the covariate on the  $x$ -axis

Figure 5.3 shows the original data on children’s height on the  $y$ -axis, along the **centred** version of their father’s data on the  $x$ -axis and with the new regression line superimposed. As is indicated in the graph, the slope  $\hat{\beta}_1$  is unchanged, while the intercept is indeed changed. That is because the change in the scale of the  $X$  covariate (which as is possible to see now goes from  $-7.23$  to  $9.27$ ) is modified. The “effect” of the covariate on the outcome is not affected by this change of scale; however, we are now in a position of providing a better interpretation of the intercept: this is the value of the expected outcome in correspondence of a centred value for the covariate equal to 0. It is easy to see that if  $X_i^* = (X_i - \bar{X}) = 0$ , then the original value  $X_i = \bar{X}$ . So the intercept is the expected outcome for a father with the average height in the observed population — and this is something that may exist and certainly makes sense!

#### **i** Centering covariates

As you will see in more details if you take STAT0019, centring covariates is particularly important within a Bayesian setting, because this has the added benefit of improving convergence of simulation

algorithms (e.g. Markov Chain Monte Carlo), which underpin the actual performance of Bayesian modelling.

The terminology “regression” comes from Galton’s original conclusion from his analysis — when plotting the original data with superimposed both the least square regression line (in blue in Figure 5.4) and the line of “equality” (the black line). This is constructed by using a slope equal to 0 and an intercept equal to 1, essentially implying that on average we expect children and their father to have the same height.

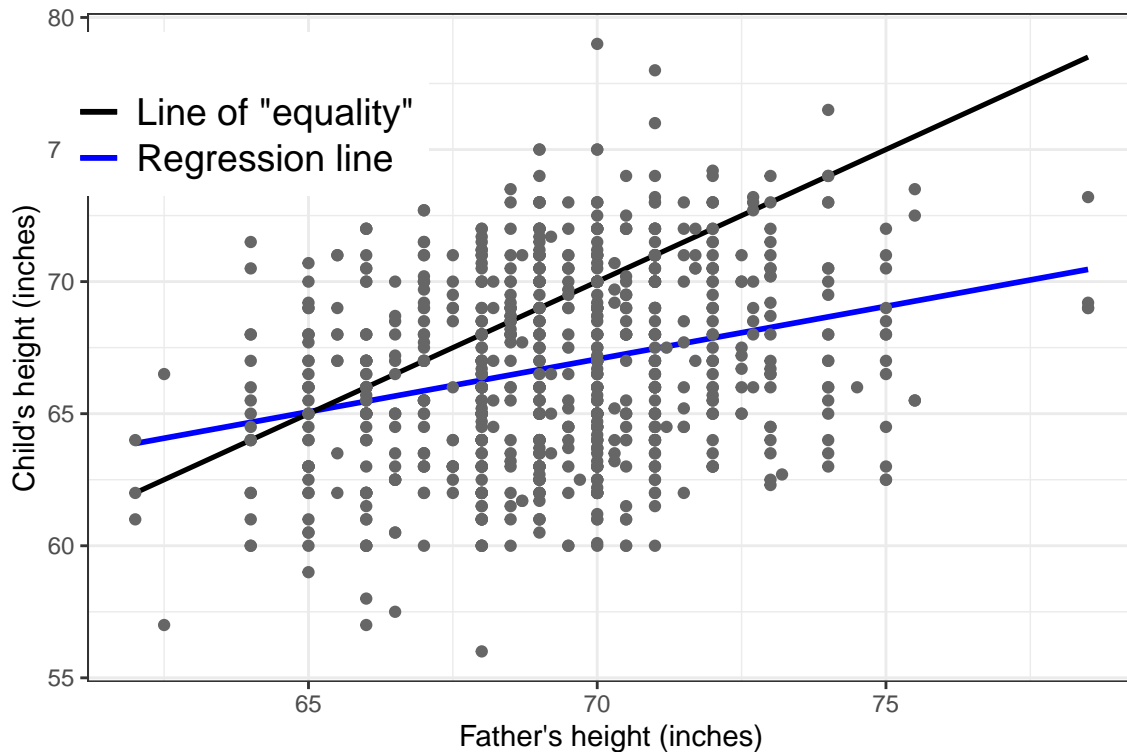


Figure 5.4: Galton’s original data with the regression and equality lines superimposed

What Galton noted is that shorter fathers tended to be associated with slightly taller children. This is noticeable because at the left-hand of the  $x$ -axis, the blue curve (the estimated least squares regression) is higher than the line of equality. Conversely, if a father was taller, then on average his child(ren) tended to be shorter than him (because at the other extreme the black line is increasingly higher than the blue line). With his eugenist hat on, he found this rather disappointing, because it meant that the species could not be improved (e.g. by selecting only taller parents to breed). For this reason, he gave this phenomenon the rather demeaning name “regression to mediocrity” or “to the mean”.

## 5.2 Regression as a statistical model

Galton’s original analysis is not really framed as a full statistical model, because it is rather based on the idea of *mathematical* optimisation provided by the least squares. In fact, he did not specify explicitly any distributional assumption behind the data observed. From the statistical perspective, this is a limiting factor, because, for example, it is impossible to go beyond the point estimate of the regression coefficients, unless we are willing to provide some more expanded model, based on probability distributions.

In practice, this extension is fairly easy, as we will show in the following. In its simplest form, we can assume that the sampling variability underlying the observed data can be described by a Normal distribution. This amounts to assuming that

$$y_i \sim \text{Normal}(\mu_i, \sigma^2) \quad (5.4)$$

$$\mu_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_K X_{Ki}. \quad (5.5)$$

for  $i = 1, \dots, n$ . This notation highlights the probabilistic nature of the assumed regression relationship between  $y$  and  $\mathbf{X}$ : we are assuming that the **linear predictor** of Equation 5.1 is *on average* the best estimate of the outcome, given a specific “profile”, i.e. a set of values for the observed covariates. But we do not impose determinism on this relationship: there is a variance, which is determined by the sampling variability in the observed data and expressed by the population parameter  $\sigma^2$ .

### **i** Statistics vs Econometrics

An alternative way of writing the regression model (which is perhaps more common in Econometrics than it is in Statistics) is to consider

$$y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + \varepsilon_i, \quad \text{with } \varepsilon_i \sim \text{Normal}(0, \sigma^2). \quad (5.6)$$

Equation 5.6 explicitly describes the assumption mentioned above. The quantity  $\varepsilon_i$  represents some kind of “white noise”, or, in other words, a random error, that is centered on 0 and whose variance basically depends on the population variability.

## 5.2.1 Bayesian approach

The model parameters for the specification in Equation 5.4 and Equation 5.5 are the coefficients  $\beta$  and the variance  $\sigma^2$ . Thus, in order to perform the Bayesian analysis, we need to specify a suitable prior distribution for  $\theta = (\beta, \sigma^2)$ . Ideally, we would specify a *joint*, multivariate prior  $p(\theta)$ , which would be used to encode any knowledge on the uncertainty about each parameter, as well as the correlation among them.

In practice, we often assume some form of (conditional) independence, where the  $(K + 2)$  dimensional distribution  $p(\theta)$  is factorised into a product of simpler (lower-dimensional) distributions, exploiting some (alleged!) independence conditions among the parameters. For instance, we may model the regression coefficients independently on one another and on the population variance

$$p(\theta) = p(\beta, \sigma^2) = p(\sigma^2) \prod_{k=0}^K p(\beta_k).$$

### **!** Prior independence vs posterior dependence

This is of course just a convenient specification and care should be taken in encoding the actual prior knowledge into the definition of the prior distribution. Notice however that even if we are assuming some form of independence in the prior distribution, it is possible that in the posterior (i.e. after observing the evidence provided by the data), we have some level of correlation among (some of) the parameters.

There are of course many possible models we can define for the prior, but a convenient (if often relatively unrealistic) choice is to assume vague Normal priors on the coefficients:  $(\beta_0, \beta_1, \dots, \beta_K) \stackrel{iid}{\sim} \text{Normal}(0, v)$

with a fixed and large variance  $v$ ; and a Gamma distribution for the *precision*  $\tau = 1/\sigma^2 \sim \text{Gamma}(a, b)$ , for some fixed, small values  $(a, b)$  — see for example Gelman et al. (2013).

As mentioned above, it is convenient, particularly within the Bayesian framework to consider a *centred* version of the covariates, where  $X_{0i}^* = X_{0i}$  and  $X_{ki}^* = (X_{ki} - \bar{X}_k)$ , for  $k = 1, \dots, K$  (notice that we should not rescale the column of the predictors matrix corresponding to the intercept — in fact we want to keep it as a vector of ones, to ensure that the matrix multiplication returns the correct values). For example, using again Galton’s data, if we wanted to include in the model also the mothers’ heights (indicated as  $X_{2i}$ ) and use a centred version of the covariates, then the linear predictor would be

$$\begin{aligned}\mu_i &= \beta_0 X_{0i}^* + \beta_1 X_{1i}^* + \beta_K X_{2i}^* \\ &= \mathbf{X}_i^* \boldsymbol{\beta}\end{aligned}$$

where  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$  and the predictors matrix  $\mathbf{X}^*$  is

$$\mathbf{X}^* = \begin{pmatrix} X_{01}^* & X_{11}^* & X_{21}^* \\ X_{02}^* & X_{12}^* & X_{22}^* \\ X_{03}^* & X_{13}^* & X_{23}^* \\ X_{04}^* & X_{14}^* & X_{24}^* \\ \vdots & \vdots & \vdots \\ X_{0n}^* & X_{1n}^* & X_{2n}^* \end{pmatrix} = \begin{pmatrix} 1 & 9.27 & 2.92 \\ 1 & 9.27 & 2.92 \\ 1 & 9.27 & 2.92 \\ 1 & 9.27 & 2.92 \\ \vdots & \vdots & \vdots \\ 1 & -0.733 & 0.92 \end{pmatrix}.$$

We may specify a prior for the father’s and mother’s effect that is fairly skeptical, by setting a Normal distribution, centred around 0 (indicating that on average we are not expecting an impact of these covariates on the predicted value of their child’s height), with a relatively large variance. For instance we could set  $\beta_1, \beta_2 \stackrel{iid}{\sim} \text{Normal}(0, \text{sd} = 10)$ . Note that in this case, looking at the context, we may argue that a standard deviation of 10 is already “large enough” to avoid including too much information in the prior. The blue curve in Figure 5.5 shows a graphical representation of this prior.

As for the intercept  $\beta_0$ , given the centring in our model, this represents the expected height of a child whose mother and father’s heights are at the average in the population. We can use some general knowledge about people’s heights in Victorian times and, assuming no particular association between the outcome and the covariate, we may set a Normal prior with mean equal to 65 inches (approximately 165 cm) and standard deviation equal to 20. Again, we are not imposing a particular value for the intercept, in the prior, but while using a reasonable choice, we still maintain some substantial uncertainty before seeing the data. Notice that essentially this prior (correctly!) assigns 0 probability of negative heights — in fact heights below 20 inches are very unlikely under the model assumed.

Unfortunately, this model is not possible to compute in closed form and so in order to estimate the posterior distributions for the parameters, we need to resort to a simulation approach (e.g. Markov Chain Monte Carlo, MCMC — the details are not important here and you will see much more on this if you take STAT0019). The output of the model is presented in Table 5.2.

Table 5.2: A summary of the posterior distributions for the model parameters

	Mean	SD	2.5%	97.5%
$\beta_0$ (intercept)	66.7545	0.1129	66.5296	66.9671
$\beta_1$ (slope for father’s height)	0.3783	0.0458	0.2862	0.4690
$\beta_2$ (slope for mothers’s height)	0.2832	0.0489	0.1888	0.3818
$\sigma$ (population variance)	3.3906	0.0829	3.2328	3.5506

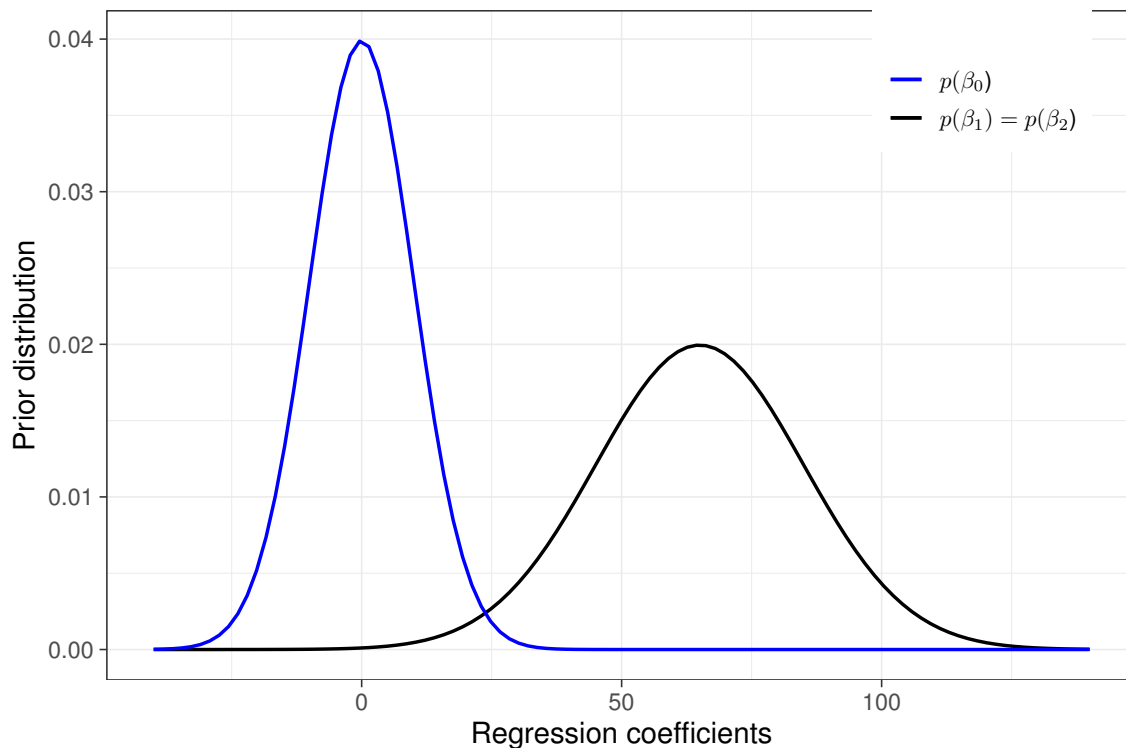


Figure 5.5: The assumed prior distribution for the regression coefficients. The distribution in black indicates the prior for the intercept  $\beta_0$ . The distribution in blue is the same for  $\beta_1$  and  $\beta_2$ . As is possible to see, most of the probability is included in a relatively large range, approximately between -20 and 20, indicating large uncertainty in the prior father's and the mother's effect

The computer output shows some summary statistics for the estimated posterior distributions of the model parameters. The posterior mean for the intercept is very close to our prior guess, but the uncertainty in the overall distribution is massively reduced by the observed data — the 95% interval estimate is indeed very narrow, ranging from 66.53 to 66.97.

As for the slopes  $\beta_1$  and  $\beta_2$ , both have a positive mean and the entire 95% interval estimate is also above 0. This indicates that there seems to be a truly positive relationship between the heights of the parents and those of their offspring. Nevertheless, the *magnitude* of the effect is not very large, which explains the phenomenon so disconcerting for Galton.

We can use a similar reasoning to that shown in Section 4.1 to determine, using the full posterior distributions, the probability that either  $\beta_1$  or  $\beta_2$  are negative (which would indicate the opposite relationship). These can be obtained numerically, using the output of the MCMC analysis. As is obvious from Figure 5.6, there is essentially no probability of either the two slopes being negative.

#### Warning

Finally, notice that the coefficients for the father's height has now changed from the least square analysis given above. This is possibly due to the influence of the prior distribution, in the Bayesian analysis. Nevertheless, because we are now including an additional covariate, it is extremely likely that the effect of the father's height be indeed modified in comparison to the simpler analysis,

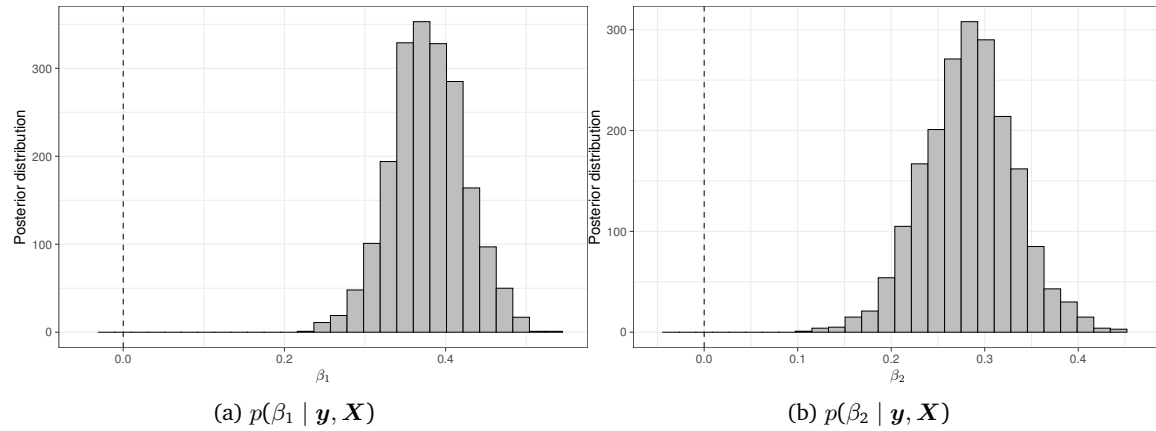


Figure 5.6: Histograms from the simulated values for the posterior distributions of  $\beta_1$  and  $\beta_2$

simply because of the joint variation brought about by the formal consideration of the mother's height.

### 5.2.2 Likelihood approach

As shown in Section 3.1.2, the likelihood approach proceeds by computing the maximum likelihood estimator for the model parameters. In this case, using relatively simple algebra, we can prove that the MLE for the coefficients  $\beta = (\beta_0, \beta_1, \dots, \beta_K)$  are equivalent to the least square solutions (as seen in Section 3.1.3, the MLE has usually all the good frequentist properties and thus it is often selected as the best estimator in that framework too).

Expanding on the result shown above, in the most general case for a linear regression, where we consider  $K$  predictors, the MLEs are

$$\hat{\beta}_0 = \bar{y} - (\hat{\beta}_1 \bar{X}_1 + \dots + \hat{\beta}_K \bar{X}_K) = \bar{y} - \sum_{k=1}^K \beta_k \bar{X}_k \quad (5.7)$$

$$\hat{\beta}_1 = \frac{\text{Cov}(y, X_1)}{\text{Var}[X_1]} \quad (5.8)$$

⋮

$$\hat{\beta}_K = \frac{\text{Cov}(y, X_K)}{\text{Var}[X_K]}. \quad (5.9)$$

In addition, the MLE for the population (sometimes referred to as “residual”) variance  $\sigma^2$  is given by

$$\hat{\sigma}^2 = \frac{\text{RSS}}{(n - K - 1)}, \quad (5.10)$$

where:

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - [\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki}])^2 \end{aligned}$$

is the *residual sum of squares*, i.e. a measure of the “error” we make by estimating the outcome using the regression line (i.e. the *residuals*  $\hat{y}_i$ ), instead of the actual observed points ( $y_i$ ); and the denominator of Equation 5.10 is the degrees of freedom, which in the general case are equal to the number of data points ( $n$ ) minus the number of regression coefficients ( $K + 1$ , in this case).

In practical terms, matrix algebra can be used to programme these equations more efficiently and compactly. For example, including again in the model both the fathers’ and the mothers’ heights on the original scale (i.e. without centring the covariates), then the linear predictor would be

$$\begin{aligned}\mu_i &= \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_K X_{2i} \\ &= \mathbf{X}_i \boldsymbol{\beta}\end{aligned}$$

where  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$  and the predictors matrix  $\mathbf{X}$  is

$$\mathbf{X} = \begin{pmatrix} X_{01} & X_{11} & X_{21} \\ X_{02} & X_{12} & X_{22} \\ X_{03} & X_{13} & X_{23} \\ X_{04} & X_{14} & X_{24} \\ \vdots & \vdots & \vdots \\ X_{0n} & X_{1n} & X_{2n} \end{pmatrix} = \begin{pmatrix} 1 & 78.5 & 67.0 \\ 1 & 78.5 & 67.0 \\ 1 & 78.5 & 67.0 \\ 1 & 78.5 & 67.0 \\ \vdots & \vdots & \vdots \\ 1 & 68.5 & 65.0 \end{pmatrix}.$$

Equations 5.7 — 5.9 can be written compactly in matrix algebra using the notation

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \quad (5.11)$$

### **i** Transpose and inverse matrices

The operator  $^\top$  indicates the *transpose* of a matrix. So if you have a matrix

$$\mathbf{X} = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

then

$$\mathbf{X}^\top = \begin{pmatrix} a & c \\ b & d \end{pmatrix},$$

i.e. the transpose matrix is constructed by flipping around the rows and the columns (the first column becomes the first row, the second column becomes the second row, etc.).

Multiplying the transpose of a matrix by the original matrix is equivalent to summing the cross-products of the values in the matrix

$$\begin{aligned}\mathbf{X}^\top \mathbf{X} &= \begin{pmatrix} a & c \\ b & d \end{pmatrix} \times \begin{pmatrix} a & b \\ c & d \end{pmatrix} \\ &= \begin{pmatrix} (a \times a + c \times c) & (a \times b + c \times d) \\ (b \times a + d \times c) & (b \times b + d \times d) \end{pmatrix} \\ &= \begin{pmatrix} a^2 + c^2 & ab + cd \\ ba + dc & b^2 + d^2 \end{pmatrix}\end{aligned}$$

So, if  $\mathbf{X}$  is the matrix of predictors



$$\begin{aligned} \mathbf{X}^\top \mathbf{X} &= \begin{pmatrix} X_{01} & X_{02} & \dots & X_{0n} \\ X_{11} & X_{12} & \dots & X_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{K1} & X_{K2} & \dots & X_{Kn} \end{pmatrix} \times \begin{pmatrix} X_{01} & X_{11} & \dots & X_{K1} \\ X_{02} & X_{12} & \dots & X_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{0n} & X_{1n} & \dots & X_{Kn} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^n X_{0i}^2 & \sum_{i=1}^n X_{0i}X_{1i} & \dots & \sum_{i=1}^n X_{0i}X_{Ki} \\ \sum_{i=1}^n X_{1i}X_{0i} & \sum_{i=1}^n X_{1i}^2 & \dots & \sum_{i=1}^n X_{1i}X_{Ki} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n X_{Ki}X_{0i} & \sum_{i=1}^n X_{Ki}X_{1i} & \dots & \sum_{i=1}^n X_{Ki}^2 \end{pmatrix}, \end{aligned}$$

which is equivalent to the computation made for the covariance in Equation 5.3. Note that, irrespective of the original dimension of a matrix, multiplying a transpose by the original matrix always produces a *square* matrix (i.e. one with the same number of rows and columns).

The matrix operator  $^{-1}$  is the generalisation of the division operation for numbers. So, much like for a number  $x$  the product  $xx^{-1} = \frac{x}{x} = 1$ , for matrices the *inverse* is such that for a square matrix  $\mathbf{X}$ , pre-multiplying by the inverse matrix produces the *identity* matrix (i.e. one with ones on the diagonal and zeros everywhere else).

$$\mathbf{X}^{-1} \mathbf{X} = \mathbf{1} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

Intuitively, Equation 5.11 is made by multiplying the inverse of the sum of squares for the matrix  $\mathbf{X}$  (which is proportional to the variance) by the sum of cross-squared between  $\mathbf{X}$  and  $\mathbf{y}$  (which is proportional to the covariance). This is basically the same as constructing the ratio between the covariance of  $\mathbf{X}$  and  $\mathbf{y}$  and the variance of  $\mathbf{X}$ , exactly as in Equation 5.3.

This matrix algebra can be programmed in R using the commands

```
# Constructs the matrix of predictors, including the first column of ones
# the second column with the fathers' heights and the third column with
# the mothers' heights, from the original dataset
X=cbind(rep(1,nrow(galton)),galton$Father,galton$Mother)

# Assigns a label 1,2,...,n to each row of the matrix X
rownames(X)=1:nrow(X)
# And then visualises X as in the equation above, e.g. using
X[c(1:4,nrow(X)),]
```

```
  [,1] [,2] [,3]
1     1  78.5  67
2     1  78.5  67
3     1  78.5  67
4     1  78.5  67
898   1  68.5  65
```

```
# Constructs the vector of outcomes with the children's heights
y=galton$Height
# Now computes the MLE for all the regression coefficients
beta=solve(t(X)%*%X)%*%t(X)%*%y
```

(in R the built-in command `solve(X)` is used to compute the inverse of a square matrix  $X$ , while the function `t(X)` is used to transpose its matrix argument).

The code above returns the following values for the coefficients.

```
MLE estimate
beta0  22.3097055
beta1   0.3798970
beta2   0.2832145
```

We can also prove that these are *unbiased* for the underlying regression coefficients  $\beta_0, \dots, \beta_K$ , i.e.

$$E[\hat{\beta}_0] = \beta_0, E[\hat{\beta}_1] = \beta_1, \dots, E[\hat{\beta}_K] = \beta_K$$

and, using the theory shown in Chapter 3, that the sampling distribution for the estimators of each  $\hat{\beta}_k$  (for  $k = 0, \dots, K$ ) is given by Normal distributions where the mean is of course the underlying “true” coefficient  $\beta_k$  and the variance is given by

$$\text{Var}[\hat{\beta}_k] = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

— the intuition behind this formula is that we rescale the estimate of the variance of the error  $\varepsilon_i$  by the variance in the covariates to provide the variance of the estimate of the effects (coefficients).

Using matrix notation to compute  $\text{RSS} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ , in R we can compute these variances using the following commands

```
# Computes the Residual Sums of Squares
RSS=t(y-X%*%beta)%*%(y-X%*%beta)
# Computes the estimate of the standard deviation sigma
# NB: "nrow(X)"=number of rows in the matrix X, while
#      "ncol(X)"=number of columns in the matrix X
sigma2.hat=as.numeric(RSS/(nrow(X)-ncol(X)))

# Now computes the variance of the coefficients using the formula above
sigma2.beta=sigma2.hat*solve(t(X)%*%X)
# Now squares the elements on the diagonal (i.e. the variances of the three
# coefficients), to obtain the standard deviations
sigma.beta=sqrt(diag(sigma2.beta))
```

to produce the estimates for the three coefficients in  $\boldsymbol{\beta}$  as below.

```
MLE estimate      sd
beta0  22.3097055  4.30689678
beta1   0.3798970  0.04589120
beta2   0.2832145  0.04913817
```

At this point, using Equation 3.8, we can substitute the MLE  $\hat{\beta}_k$  for  $\hat{\theta}$  and the estimate of the standard deviation of  $\hat{\beta}_k$  for  $\sigma/\sqrt{n}$  and compute a 95% interval for  $\hat{\beta}_k$  as

$$\left[ \hat{\beta}_k - 1.96\sqrt{\text{Var}[\hat{\beta}_k]}; \hat{\beta}_k + 1.96\sqrt{\text{Var}[\hat{\beta}_k]}, \right]$$

which in the current case gives

$$95\% \text{ interval for } \beta_0 = [22.31 - 1.96 \times 4.31; 22.31 + 1.96 \times 4.31] = [13.87; 30.75]$$

$$95\% \text{ interval for } \beta_1 = [0.38 - 1.960.04590.38 + 1.96 \times 0.0459] = [0.29; 0.47]$$

$$95\% \text{ interval for } \beta_2 = [0.283 - 1.960.04910.283 + 1.96 \times 0.0491] = [0.187; 0.38].$$

Similarly, because we know what the sampling distribution for each of the three estimates is, we can compute  $p$ -values, for instance against the null hypothesis  $H_0 : \hat{\beta}_k = 0$ . Notice that this is really relevant just for the slopes (i.e. the effects of the covariates), because if the resulting  $p$ -value is small, then we would have determined some evidence against the hypothesis of “no effect” (e.g. of the fathers’ heights on their children’s height).

Recalling Equation 4.2, we can prove that, under  $H_0$ ,

$$T = \frac{\hat{\beta}_k - 0}{\text{Var}[\hat{\beta}_k]} \sim t(0, 1, (n - K - 1)). \quad (5.12)$$

Because all the  $\hat{\beta}_k > 0$ , then the relevant tail-area probability is the one to the right of the underlying  $t$  sampling distribution and so the  $p$ -values are computed in R as the following

```

p-value
beta0 0.0000001371088497591075
beta1 0.0000000000000002260111
beta2 0.000000056616890636972

```

(recall that the option `lower.tail=FALSE` computes the area to the right of a given distribution).

Because the  $p$ -values are all very small (and certainly much lower than the common threshold of 0.05), we can claim very strong evidence against  $H_0$  and so, in a purely Fisherian interpretation, we would advocate the presence of some effect of both fathers’ and mothers’ heights on the children’s height. In common parlance, the three coefficients are deemed to be “*highly significant*” at the 0.05 level.

This results is also consistent with the analysis of the 95% confidence intervals (cfr. Section 4.4.3). The  $p$ -values are all very small — and at the same time, all the 95% confidence intervals *exclude* the value 0, indicating that the results are “significant” (and in this case, positive).

Of course, in practice, you will never need to make matrix calculations by hand — or even programme them directly. Most likely you will use routines and programmes available in statistical software to do the regression analysis. For example, in R we can compute the regression coefficients using the built-in function `lm`, as in the following code.

```

# Runs the function "lm" to run the model including "Height" as the reponse
# (that is the variable to the left of the twiddle symbol "~"), while "Father"
# and "Mother" (the variables to the right of the twiddle) are the covariates.
# These are recorded in the dataset called "galton"
m1=lm(formula=Height~Father+Mother,data=galton)
# Now displays the output
summary(m1)

```

Call:

```
lm(formula = Height ~ Father + Mother, data = galton)
```

Residuals:

```

      Min      1Q Median      3Q      Max
-9.136 -2.700 -0.181  2.768 11.689

```

Coefficients:

```

      Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.30971    4.30690   5.180 2.74e-07 ***
Father      0.37990     0.04589   8.278 4.52e-16 ***
Mother      0.28321     0.04914   5.764 1.13e-08 ***
---

```

```

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 3.386 on 895 degrees of freedom

Multiple R-squared: 0.1089, Adjusted R-squared: 0.1069

F-statistic: 54.69 on 2 and 895 DF, p-value: &lt; 2.2e-16

The standard output reported by R includes the estimates for the coefficients (the column labelled as Estimate), their standard error based on the suitable sampling distributions (the column labelled Std. Error) and the value of the test statistic  $T$  of Equation 5.12 in the column labelled as t value.

Notice that by default, `lm` reports the  $p$ -value computed for an implicit alternative hypothesis  $H_1 : \beta_k \neq 0$  (see the discussion in Section 4.4.1). For this reason, the  $p$ -values reported by `lm` in the column labelled as `Pr(>|t|)` are different than the one we have computed above. If we multiply by 2 the  $p$ -values above (to account for the *difference than* implicit in the alternative hypothesis, as opposed to simply the tail-area probability under the sampling distribution in correspondence with  $H_0$ ), we obtain

```

      p-value
beta0 2.742177e-07
beta1 4.520223e-16
beta2 1.132338e-08

```

which are in fact consistent with the table reported by `lm`. To highlight the fact that these  $p$ -value are highly significant, `lm` uses a “star-based” rating system — where `***` indicates that a  $p$ -value is between 0 and 0.001, `**` that it is between 0.001 and 0.01, `*` that it is between 0.01 and 0.05 and `.` that it is between 0.05 and 0.1.

#### Warning

The analysis based on the MLE and  $p$ -values shows some slight discrepancies with the Bayesian analysis shown above. By and large the results are extremely consistent — notice that because of the different scaling of the covariate, the intercepts  $\beta_0$  are not directly comparable. As for the slopes (which, on the contrary *are* directly comparable), the Bayesian analysis estimates the effect of fathers’ height to be 0.3783, while the frequentist model indicates a value of 0.3799. For the mothers’ height effect, the Bayesian model estimates 0.2832, while the MLE-based analysis indicates that it is 0.2832. More importantly, all the coefficients are estimated with large precision, so that the entire intervals are above 0, in both approaches.

**In this case**, because we have a relatively large dataset, with evidence that consistently points towards the same direction; and we have used relatively vague priors, the numerical outputs of the two approaches are highly comparable. But this need not be the case in general, particularly when the data size is small and they provide only limited evidence. From the Bayesian point of view, this is perfectly reasonable: if we do not have overwhelming evidence, it is sensible that including prior knowledge does exert some influence on the final decisions.

This feature is also particular important when we need to complement limited evidence from observed data (e.g. in terms of short follow up, or large non-compliance).

### 5.3 Generalised linear regression

One of the main assumptions underlying the linear regression analysis seen above when viewed as a statistical model is that the underlying outcome is suitably modelled using a Normal distribution — this implies that we can assume that  $Y$  is reasonably symmetric, continuous and unbounded, i.e. it can, at least theoretically, take on values in the range  $(-\infty; \infty)$ .

However, as we have seen in Chapter 2, there are many cases in which variables are not suitably described by a Normal distribution — notably Bernoulli/Binomial or Poisson counts, but (as you will see if you take STAT0019) also other variables describing skewed phenomena (e.g. costs or times to event). In these instances, we can slightly extend the set up of Equation 5.4 and Equation 5.5 to account for this extra complexity. This can be accomplished using the following structure.

$$\left\{ \begin{array}{ll} y_i \overset{iid}{\sim} p(y_i | \boldsymbol{\theta}_i, \mathbf{X}_i) & \text{is the model to describe sampling variability for the outcome} \\ \boldsymbol{\theta}_i = (\mu_i, \boldsymbol{\alpha}) & \text{is the vector of model parameters} \\ \mu_i = E[Y_i | \mathbf{X}_i] & \text{is the mean of the outcome given the covariates } \mathbf{X}_i \\ \boldsymbol{\alpha} & \text{is a vector of other potential model parameters, e.g. variances, etc.} \\ & \text{(NB this may or may not exist for a specific model)} \\ g(\mu_i) = \mathbf{X}_i \boldsymbol{\beta} & \text{is the linear predictor on the scale defined by } g(\cdot). \end{array} \right. \quad (5.13)$$

#### **i** Link functions

A function such that  $g(x) = x$  is referred to as the **identity** function. We can see that linear regression, presented in Section 5.2, is in fact a special case of the wider class of structures described in Equation 5.13, which is often referred to as *Generalised Linear Models* (GLMs). A GLM in which  $p(y_i | \boldsymbol{\theta}, \mathbf{X}) = \text{Normal}(\mu_i, \sigma^2)$  and  $g(\mu_i) = \mu_i = E[Y | \mathbf{X}_i] = \mathbf{X}_i \boldsymbol{\beta}$  is the linear regression model of Equation 5.4 and Equation 5.5.

Upon varying the choice of the distribution  $p(\cdot)$  and the transformation function  $g(\cdot)$ , we can model several outcomes.

#### 5.3.1 Logistic regression

When the outcome is a binary or Binomial variable, we know that the mean  $\theta$  is the probability that a random individual in the relevant population experiences the event of interest. Thus, by necessity,  $\theta_i = E[Y_i | \mathbf{X}_i]$  is bounded by 0 from below and 1 from above (i.e. it **cannot** be below 0 or above 1). For this reason, we should not use a linear regression to model this type of outcome (although sometimes this is done, particularly in Econometrics, in the context of “two-stage least square analysis”, which you may encounter during your studies).

One convenient way to model binary outcomes in a regression context is to use the general structure of Equation 5.13, where the outcome is modelled using  $y_i \overset{iid}{\sim} \text{Bernoulli}(\theta_i)$  and

$$g(\theta_i) = g(E[Y_i | \mathbf{X}_i]) = \text{logit}(\theta_i) = \log\left(\frac{\theta_i}{1 - \theta_i}\right) = \mathbf{X}_i \boldsymbol{\beta} \quad (5.14)$$

(notice that in the Bernoulli model there is only one parameter and so we indicate here  $\boldsymbol{\theta} = \theta$ ).

Equation 5.14 is referred to as **logistic regression**. Figure 5.7 shows graphically the mapping from the original range of the parameter  $\theta$ , on the  $x$ -axis, to the rescaled parameter  $g(\theta) = \text{logit}(\theta)$ , on the  $y$ -axis. As is possible to see, the original range is mapped onto the whole set of numbers from  $-\infty$  (in correspondence of the value  $\theta = 0$ ) to  $\infty$  (for  $\theta = 1$ ).

Registered S3 methods overwritten by 'bmhe':

```
method      from
print.bugs  R2WinBUGS
print.rjags R2jags
```

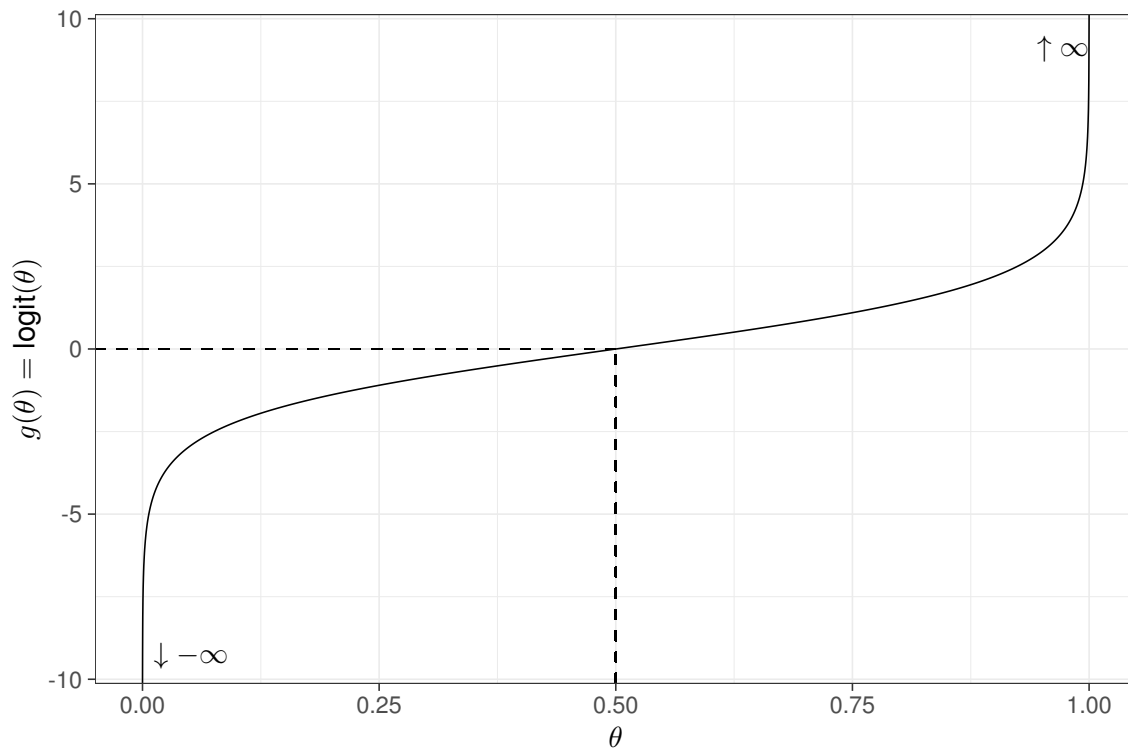


Figure 5.7: Graphical description of the shape and features of the logit function. For  $\theta \rightarrow 0$ , then  $\text{logit}(\theta) \rightarrow -\infty$ , while for  $\theta \rightarrow 1$ , then  $\text{logit}(\theta) \rightarrow +\infty$ . For  $\theta = 0.5$ , then  $\text{logit}(\theta) = 0$

This is obvious by noting that

$$\theta = 0 \Rightarrow \log\left(\frac{\theta}{1-\theta}\right) = \log\left(\frac{0}{1}\right) = \log(0) \rightarrow -\infty$$

and

$$\theta = 1 \Rightarrow \log\left(\frac{\theta}{1-\theta}\right) = \log\left(\frac{1}{0}\right) = \log(\infty) \rightarrow \infty.$$

In addition, the graph in Figure 5.7 shows that the resulting function of  $\theta$  is reasonably symmetric around the central point ( $\theta = 0.5$ ).

#### **i** Odds vs probabilities

If  $\theta$  represents a probability (e.g. that a given event  $E$  happens), the quantity

$$O = \left(\frac{\theta}{1-\theta}\right) = \frac{\Pr(E)}{1-\Pr(E)}$$

is called the **odds** for the event  $E$  and it represents a measure of how more likely  $E$  is to happen than not. If  $\Pr(E) = \theta = 0.5$ , then  $O = 0.5/0.5 = 1$ . If  $\Pr(E)$  is small, then  $O$  is also very small, while if  $E$  is very likely, then  $O$  is increasingly bigger. The range in which  $O$  is defined is characterised by the extreme values 0 (in correspondence of which  $E$  is impossible, i.e.  $\Pr(E) = 0$  and thus  $O = 0/1 = 0$ ) and  $\infty$ , when  $E$  is certain, i.e.  $\Pr(E) = 1$  and thus  $O = 1/0 = \infty$ .

The quantity  $\log\left(\frac{\theta}{1-\theta}\right) = \log O$  is the **log-odds** for the event  $E$ . By applying the log transformation to  $O$ , we map its range in the interval  $[\log(0) = -\infty; \log(\infty) = \infty]$ .

Equation 5.14 clarifies that logistic regression implies that we are modelling the log-odds using a linear predictor or, in other words, that we are assuming a linear relationship between the mean outcome and the covariates, *on the log odds scale*.

For example, consider again Galton's data, but this time, our outcome is given by a new variable, which takes value 1 if the original child's height is above a threshold of 71 inches (180 cm) and 0 otherwise. We can create the suitable data in R using the following command.

```
# Creates a variable "y2" taking value 1 if the original child's height > median
y2=ifelse(galton$Height>71,1,0)
# Summarises the new variable
table(y2)
```

```
y2
 0  1
801 97
```

As is possible to see, nearly 10.80% (i.e. 97/898) of the sample “experiences the event”, i.e. is taller than the set threshold.

### ! Logistic regression coefficients

The interpretation of the regression coefficients in a logistic regression needs a bit of care, given the change in the scale we define for the linear predictor.

The intercept is still related to the expected mean of the outcome for the profile of covariates all set to 0. So, in Galton's example, centering the covariates for simplicity of interpretation, this would be the expected value for the height of a child of “average” father and mother (in terms of height). However, this time the expected mean of the outcome is the *probability* of the underlying event of interest, e.g. that a child is taller than the threshold of 71 inches. So, for an individual with *centred* covariates set to 0 (e.g. for a child whose parents' *centred* height is 0), then the regression line (on the log-odds scale) is simply

$$\text{logit}(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0$$

and thus, recalling that  $\exp(\log(x)) = x$ , we can invert the logit function to give

$$\begin{aligned}\exp\left[\log\left(\frac{\theta_i}{1-\theta_i}\right)\right] &= \exp(\beta_0) \Rightarrow \\ \theta_i &= \exp(\beta_0)(1-\theta_i) \Rightarrow \\ \theta_i[1 + \exp(\beta_0)] &= \exp(\beta_0) \Rightarrow\end{aligned}$$

$$\theta_i = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \quad (5.15)$$

(the right hand side of Equation 5.15 is often referred to as the *expit* or *inverse logit* transformation). This particular rescaling of intercept represents the expected mean outcome (= the probability that the event under study occurs, for an individual with covariates set to 0 — or to the overall average, if we are centring them).

The “slope” has also a slightly different interpretation in a logistic regression. We have seen before that a generic slope  $\beta_k$  represents the difference in the expected outcome for two individuals whose  $k$ -th covariate varies by one unit, “all else being equal”, i.e. where all the other covariates are set to the same value.

For instance, if in the current example we compared two fathers, one for whom the centred height was 0 (= 69.23 inches) and one for whom the centred height was 1 (= 70.23 inches), then we would have

$$\text{logit}(\theta \mid X_1^* = 0) = \beta_0 + \beta_1 \times 0 = \beta_0,$$

for the first one; and

$$\text{logit}(\theta \mid X_1^* = 1) = \beta_0 + \beta_1 \times 1,$$

for the second one. The difference between the expected outcomes would then be

$$\begin{aligned} \Delta_X &= \text{logit}(\theta \mid X_1^* = 1) - \text{logit}(\theta \mid X_1^* = 0) \\ &= \log\left(\frac{\theta}{1-\theta} \mid X_1^* = 1\right) - \log\left(\frac{\theta}{1-\theta} \mid X_1^* = 0\right) \\ &= \log\left(\frac{\theta_1}{1-\theta_1}\right) - \log\left(\frac{\theta_0}{1-\theta_0}\right) \\ &= (\beta_0 + \beta_1) - (\beta_0) \\ &= \beta_1, \end{aligned}$$

indicating the probability of the event in correspondence of the covariate profile  $X_1^* = j$  (for  $j = 0, 1$ ) with the notation  $\theta_j$ , for simplicity.

Recalling that for any two positive numbers  $(a, b)$  we can show that  $\log(a) - \log(b) = \log(a/b)$ , we can then write

$$\log\left(\frac{\theta_1}{1-\theta_1} \Big/ \frac{\theta_0}{1-\theta_0}\right) = \beta_1$$

The quantity  $\left(\frac{\theta_1}{1-\theta_1} \Big/ \frac{\theta_0}{1-\theta_0}\right)$  is the ratio of two odds — this is called the **odds ratio** (OR) and describes how much more likely an event  $E$  is to happen among those individuals who present the characteristic associated with  $X_1^* = 1$  (which happens with probability  $\theta_1$ ) than among those who have  $X_1^* = 0$  (which is associated with probability  $\theta_0$ ). Taking the log of this quantity gives the **log-OR**, which we can now see is the same as the value of the slope  $\beta_1$ .

Theoretically, the log-OR ranges between  $-\infty$  and  $+\infty$ , with larger values indicating that the event is more likely to happen when individuals are associated with larger values of the covariate. In this case, the taller the father, the more likely their child to be taller than 71 inches, by an amount  $\exp(\beta_1)$ .

In general, negative values for a log-OR indicate that the covariate is *negatively associated* with the outcome, while positive log-OR suggest that the covariate is *positively associated* with the outcome. If we transform this on to the natural scale by exponentiating the log-OR, which means that  $\text{OR} \in [0, \infty)$ , we can interpret an  $\text{OR} > 1$  to indicate that the covariate has a positive effect on the (probability of the) outcome, while if  $\text{OR} < 1$  then the opposite is true. When  $\text{log-OR} = 0$  (or, equivalently  $\text{OR} = 1$ ), then there is no association between the covariate and the outcome.



### 5.3.1.1 Bayesian approach

In a Bayesian context, theoretically the model of Equation 5.14 does not pose particular problems: we need to specify prior distributions on the coefficients  $\beta$  and, much as in the case of linear regression, we can use Normal distributions — note that Equation 5.14 is defined on the log-odds scale, which as suggested above, does have an unbounded range, which makes it reasonable to assume a Normal distribution for the coefficients.

Equation 5.15 is helpful when we want to set up a prior distribution for  $\beta_0$  in a meaningful way for the main parameter  $\theta$ . For example, imagine we wanted to encode the assumption that, before getting to see any data, we expected only a 5 to 20% chance for a random person in the population to be taller than 71 inches. If we set a Normal( $\text{logit}(0.105), \text{sd}=0.4$ ) for  $\beta_0$ , then we can check that what we are actually implying is a prior for the underlying “baseline” probability (i.e. for the child of “average” parents) as the one represented in Figure 5.8.

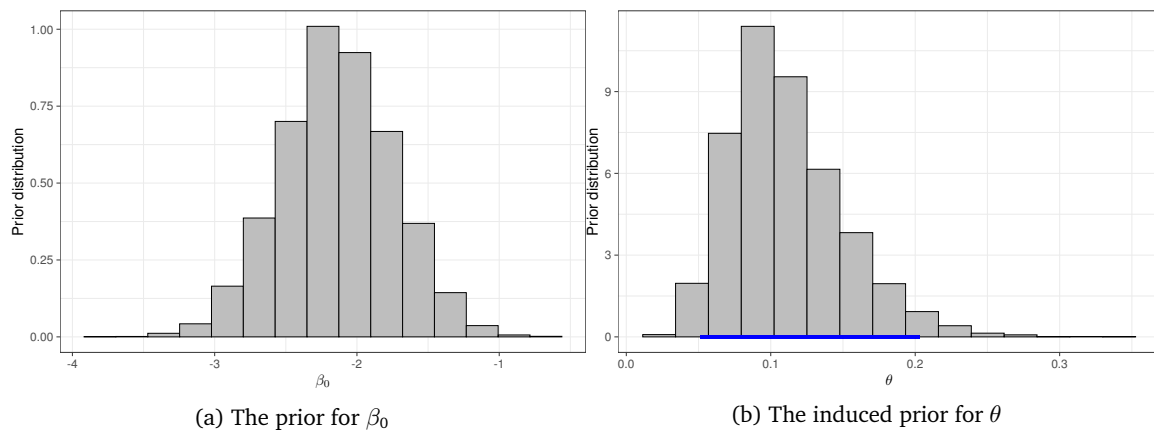


Figure 5.8: The prior distribution for  $\beta_0$  in panel (a) is used to encode the information that  $\theta$  is most likely between around 0.05 and 0.2, when  $X_{1i}^* = X_{2i}^* = 0$ , indicated by the dark blue horizontal line at the bottom on panel (b)

The dark blue horizontal line just above the  $x$ -axis in panel (b) indicates the 95% prior interval estimate for the parameter  $\theta$ , which is derived by the prior depicted in panel (a). As is possible to see, the current choice for the mean and standard deviation of the Normal prior do induce a 95% interval covering approximately the required range. The values for the parameters of the Normal prior distribution for  $\beta_0$  can be found by trial-and-error, e.g. using the following R code:

```
# Defines the mean and sd of the prior
b0=log(0.105/(1-0.105))
s0=0.4
# Simulates 10000 values from the prior
beta0=rnorm(10000,b0,s0)
# Rescales the prior to the scale of the main parameter theta
theta=exp(beta0)/(1+exp(beta0))
# Checks the implied 95% interval estimate for theta
cbind(quantile(theta,.025),quantile(theta,.975))
```

	[, 1]	[, 2]
2.5%	0.05049197	0.2023662

and changing the imposed value for  $b_0$  and  $s_0$  until the resulting approximate 95% interval returns values close enough to the required limits (5-20%). This procedure is often referred to as *forward sampling*. Note that setting some mildly “informative” prior for the intercept may be helpful in stabilising the inference, particularly when the data are made by only few records (i.e. small sample size), or, even more importantly, when the number of observed “successes” is small.

As for the slopes, we may elicit some informative prior and encode genuine prior knowledge in the model — for example, we may have some strong belief in a given treatment effect and thus we may set up a prior for the corresponding coefficient that is concentrated above 0. This would indicate a large prior probability that the resulting OR is greater than 1 and thus a positive association between the treatment and the outcome. However, in general terms, we may be unwilling to specify too strong a prior on a given treatment effect because we would like to be a bit more conservative and let the data drive the inference on this particular relationship.

For example we could set up  $\beta_1, \beta_2 \stackrel{iid}{\sim} \text{Normal}(0, sd = 2)$ . This assumes that, before seeing any data, we are not expecting a particular effect of either father’s or mother’s height on the probability of a child being taller than 71 inches (because these distributions are centred around 0 — recall that  $\beta_1$  and  $\beta_2$  represent the log-ORs!). However, we are implying some variance around it to guarantee the possibility that the effect is either positive or negative.

Notice that we are imposing a standard deviation of 2 — you may think that this is rather strict and we are in fact including some strong prior on these distributions. However, recall that the slopes are defined on the log-odds scale and so a value of 2 for a standard deviation is actually pretty large, when rescaled back to the original probability scale. Adapting Equation 3.8, we know that for a Normal distribution with mean 0 and standard deviation of 5, 95% of the probability is approximately included in the interval  $[-1.96 \times 2; 1.96 \times 2] = [-3.92; 3.92]$ , which when we map back on to the probability scale applying the inverse logit transformation implies a prior 95% interval for the OR of  $[0.0198; 50.40]$ . That is extremely vague!

### ! The “Parachute effect”

While we do not want to impose too strong priors on the “treatment effect”, there is much information that we can use to restrict the reasonable range of a log-OR.

For instance, consider an experiment in which you test the effectiveness of parachutes, when jumping off a flying plane. You define your outcome  $Y$  as 0 if the individual jumps off the plane and dies and 1 if they survive. You also have a covariate  $X_i$  taking value 1 if individual  $i$  is given a parachute and 0 otherwise.

In a situation such as this, you may expect the treatment to have a very large effect — almost everyone with a parachute can be reasonably expected to survive, while almost (or probably just!) everyone without one is most likely to die. So if you define  $\theta_1 = \Pr(Y = 1 \mid X = 1)$  and  $\theta_0 = \Pr(Y = 1 \mid X = 0)$ , we could reasonably estimate something like  $\theta_1 = 0.9$  and  $\theta_0 = 0.1$  (the actual numbers are irrelevant — but the magnitude is important and arguably realistic). Thus, in this case the OR would be

$$\begin{aligned} \text{OR} &= \frac{\theta_1}{1 - \theta_1} \bigg/ \frac{\theta_0}{1 - \theta_0} \\ &= \frac{0.9}{1 - 0.9} \bigg/ \frac{0.1}{1 - 0.1} \\ &= \frac{0.9}{0.1} \bigg/ \frac{0.1}{0.9} = 9 \bigg/ \frac{1}{9} = 81. \end{aligned}$$

This means that people with a parachute are 81 times more likely to survive the jump off the plane than those without a parachute. This massive OR comes about, of course, because of the

assumptions we are making about such a large treatment effect — almost everyone without the parachute dies, while almost everyone with survives. In a case such as this, we may be prepared to believe a very large OR. But in most cases, interventions do **not** have such dramatic effects. And thus, it is reasonable to imagine that ORs greater than 3 or perhaps 4 are already relatively unlikely to be observed in practice!

Interestingly, there is no closed form for the posterior distributions of the model parameters, in a logistic regression. Thus we need to resort to simulations, e.g. MCMC. The details of the MCMC model used to run the analysis are not important here, but the results are summarised in Table 5.3.

Table 5.3: A summary of the posterior distributions for the model parameters

	Mean	SD	2.5%	97.5%
$\beta_0$ (intercept)	-0.0719	0.0687	-0.2010	0.0605
$\beta_1$ (logOR for father's height)	0.1676	0.0303	0.1087	0.2283
$\beta_2$ (logOR for mothers's height)	0.0987	0.0304	0.0376	0.1595

In a logistic regression, typically we do not worry too much about the intercept (given the caveat above and the rescaling necessary to make sense of its value in terms of the underlying probability of “success”). As for the log-ORs, we can see that both the covariates are associated with a positive point estimate and both the entire 95% interval estimates are also positive, thus indicating a probability  $\geq 0.95$  that the posterior distributions of  $\beta_1$  and  $\beta_2$  are positive.

Of course, once we have the simulations for the log-OR, we can simply exponentiate each simulated values to obtain a full posterior distribution for the ORs. So for example if we had stored the output of the MCMC model in two suitable objects called `beta1` and `beta2`, then we could simply rescale them to compute numerically probabilities associated with them, for example as in the following R code.

```
# Constructs the ORs from the original simulations obtained by the model
OR1=exp(beta1)
OR2=exp(beta2)
# Tail-area probability to estimate Pr(OR1<1). This is the proportion of
# simulations for OR1 that are below 1
sum(OR1<1)/length(OR1)
```

```
[1] 0
```

```
# Tail-area probability to estimate Pr(OR2<1). This is the proportion of
# simulations for OR2 that are below 1
sum(OR2<1)/length(OR2)
```

```
[1] 0.0005
```

Once the objects `OR1` and `OR2` are available, we could plot histograms of the posterior distributions, as shown in Figure 5.9. As is possible to see, in both cases, none or very little of the posterior distribution is below 1, indicating a “highly significant” result in terms of the positive association between the covariates and the outcome.

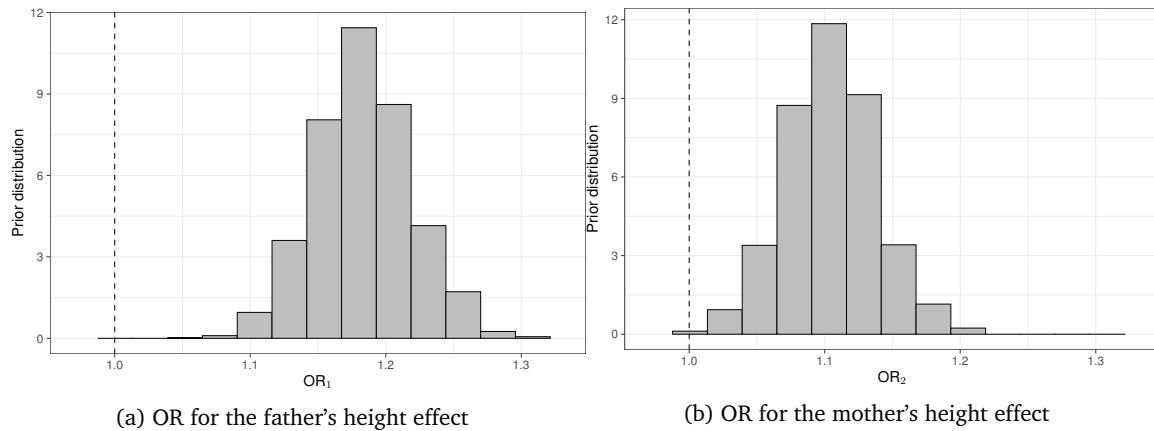


Figure 5.9: Histograms for the rescaled posterior distributions for ORs

### 5.3.1.2 Likelihood approach

If we consider a Likelihood approach, the idea is of course to maximise the likelihood function for the model parameters, to determine the MLE, which, again, would be considered a very good candidate for optimality even under a pure Frequentist approach. Unfortunately, for a logistic regression model, it is not possible to obtain maximum values analytically. Thus, we resort to numerical maximisation — usually based on a clever approximation method referred to as *Newton-Rapson algorithm*.

In practice, we do not need to programme this algorithm ourselves, but rather we rely on existing routines or programmes. For example R has a built-in command `glm`, that can be used to obtain the MLEs for GLMs.

The following code shows how to perform a GLM analysis for the model shown in Section 5.3.1.1 — all the assumptions are identical, except of course that we do not specify any prior distribution in this case.

Call:

```
glm(formula = formula, family = "binomial", data = data.frame(y2,
X))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7428	-1.1254	-0.7077	1.1306	1.6407

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
XIntercept	-0.008175	0.068537	-0.119	0.90506
XFather	0.166414	0.029209	5.697	0.000000122 ***
XMother	0.097989	0.030380	3.225	0.00126 **

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1244.9 on 898 degrees of freedom  
Residual deviance: 1197.1 on 895 degrees of freedom  
AIC: 1203.1

Number of Fisher Scoring iterations: 4

The computer output is fairly similar to that presented for the outcome of `lm` in the case of a linear regression. The main parameters are presented in the core of the table in terms of the point estimate (labelled as Estimate), the standard deviation (Std. Error), the value of the test statistic used to test the null hypothesis that each is equal to 0 (z value) and the “two-sided”  $p$ -value (indicated as  $\text{Pr}(>|z|)$  — see the discussion at the end of Section 5.2.2).

In this particular instance, the results are fairly similar to the numerical output of the Bayesian analysis. The intercept shows some slight differences — this is mainly due to the informative prior used above. However, for the two slopes, given that the prior was centred around 0 and in fact fairly vague, the effects of the covariates is estimated to very similar numerical values. Even the analysis of the  $p$ -values is consistent with the Bayesian analysis, in this case — the effect of the father’s height is more “significant”, much as the posterior distribution for  $\beta_1$  had shown a lower probability of being less than 0 in the Bayesian model.

Notice that unlike in the case of the linear regression, the  $p$ -values are computed this time based on a Wald test (see Section 4.4.4), computed as the ratio of the MLE minus the null value (0, in this case) to its standard deviation. So for example, in this case, the test statistic for  $X_2$  (the mother’s effect) can be computed as

$$W = \frac{\hat{\theta} - 0}{\sqrt{\text{Var}[\hat{\theta}]}} = \frac{0.098}{0.0304} = 3.225,$$

as shown in the computer output. We can compare this to the standard Normal to determine the  $p$ -value. In particular, as mentioned above, R constructs the two-sided version of the  $p$ -value, essentially using the following code.

```
# Stores the summary of the model output in the object "s"
s=summary(model)
# Constructs the W statistic using the elements of the object "s"
w=s$coefficients[3,1]/s$coefficients[3,2]
# Computes the two-sided p-value based on Normal(0,1) approximate distribution
P=2*(pnorm(q=w,mean=0,sd=1,lower.tail=FALSE))
# Prints the output (which is the same as in the computer output)
P

[1] 0.001257877
```

### 5.3.2 Poisson and other GLMs

Logistic regression is only one of the models embedded in the wider family of GLMs. The general principles are essentially the same in all circumstances — specify a distribution  $p(y_i | \theta_i, \mathbf{X}_i)$  and a suitable map from the *natural* scale of the mean to an unbounded range, where we can claim at least approximately linearity in the covariates.

Another common example is Poisson regression, where we have observed data on the outcome  $y_1, \dots, y_n \stackrel{iid}{\sim} \text{Poisson}(\theta_i)$ . Here the parameter  $\theta_i$  represents at once the mean and the variance of the underlying Poisson distribution. Because  $\theta_i \geq 0$  (as it represents a rate, i.e. the intensity with which the counts are observed), a suitable transformation is simply to take the log and model

$$g(\theta_i) = g(\text{E}[Y_i | \mathbf{X}_i]) = \log(\theta_i) = \mathbf{X}_i \beta. \quad (5.16)$$

In comparison to logistic regression, the interpretation of the coefficients is slightly simpler: using a reasoning similar to that shown for the intercept and slope of a logistic regression, we can show that, in the case of a Poisson GLM:

- The intercept  $\beta_0$  is the log rate for an individual whose covariates are set to 0;
- The slope  $\beta_k$  is the log *relative risk* corresponding to increasing the covariate by one unit. So if you compare two individuals, the first of whom has the value of the covariate  $X_k = 1$  and the other for whom  $X_k = 0$ , then  $\beta_k = \log\left(\frac{\theta_1}{\theta_0}\right)$ , where  $\theta_j$  is the rate associated with a covariates profile of  $X_k = j$ , for  $j = 0, 1$ .

The log-linear model embedded in the Poisson structure can be applied to several other distributions to describe sampling variability in the observed outcome. Other relevant examples include the case of time-to-event outcomes, for which suitable models are Weibull or Gamma, among others. Because all of these are defined as positive, continuous variables, their mean is also positive and so it is useful to rescale the linear predictor on the log scale.

---

## References

- Bayes, Thomas. 1763. "An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S." *Philosophical Transactions of the Royal Society of London*, no. 53: 370–418.
- Casella, George, and Roger L Berger. 2002. *Statistical inference*. Vol. 2. Pacific Grove, CA, US: Duxbury.
- Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian data analysis*. Boca Raton, FL, US: Chapman; Hall/CRC.
- Gelman, Andrew, and Jennifer Hill. 2007. "Data Analysis Using Regression and Hierarchical/Multilevel Models."
- Goodman, Steven N. 1999. "Toward evidence-based medical statistics. 1: The P value fallacy." *Annals of Internal Medicine* 130: 995–1004.
- Keynes, John Maynard. 1923. *A Tract on Monetary Reform*. London, Macmillan.
- Kruschke, John. 2014. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. San Diego, CA, US: Academic Press.
- Leemis, L, and J McQueston. 2008. "Univariate Distribution Relationships." *The American Statistician* 62 (1): 45–53. <http://www.math.wm.edu/~leemis/2008amstat.pdf>.
- Lunn, David, Chris Jackson, Nicky Best, David Spiegelhalter, and Andrew Thomas. 2012. *The BUGS book: A practical introduction to Bayesian analysis*. Boca Raton, FL, US: Chapman; Hall/CRC.
- Spiegelhalter, David J, Keith R Abrams, and Jonathan P Myles. 2004. *Bayesian approaches to clinical trials and health-care evaluation*. Chichester, UK: John Wiley & Sons.
- Wasserstein, Ronald L, Nicole A Lazar, et al. 2016. "The ASA's statement on p-values: context, process, and purpose." *The American Statistician* 70 (2): 129–33.