Analysing data using linear models

Stéphanie M. van den Berg

Fifth edition (SPSS and R)
(October 7, 2020)

University of Twente
Department of Research Methodology, Measurement and Data Analysis

For source code and updates: github.com/pingapang/book
Email: stephanie.vandenberg@utwente.nl

First edition:	October 2018
Second edition:	November 2018
Third edition:	January 2019
Fourth edition:	November 2019
Fifth edition:	October 2020

Preface

This book is for bachelor students in social, behavioural and management sciences that want to learn how to analyse their data, with the specific aim to answer research questions. The book has a practical take on data analysis: how to do it, how to interpret the results, and how to report the results. All techniques are presented within the framework of linear models: this includes simple and multiple regression models, linear mixed models and generalised linear models. This approach is illustrated using SPSS, and in some cases also R.

Contents

Chapter 1

Variables, variation and co-variation

1.1 Units, variables, and the data matrix

Data is the plural of datum, and datum is the Latin translation of 'given'. That the world is round, is a given. That you are reading these lines, is a given, and that my dog's name is Philip, is a given. Sometimes we have a bunch of given facts (data), for example the names of all students in a school, and their marks for a particular course. We could put these data in a table, like the one in Table 1.1. There we see information ('facts') about seven students. And of these seven students we know two things: their name and their grade. You see that the data are put in a matrix with seven (horizontal) rows and two (vertical) columns. Each row stands for one student, and each column stands for one property.

In data analysis, we always put data in such a matrix format. In general, we put the objects of our study in rows, and their properties in columns. The objects of our study we call *units*, and the properties we call *variables*.

Table 1.1: Data matrix with 7 units and 2 variables.

name	grade
Mark Zimmerman	5
Daisy Doe	8
Mohammed Solmaz	5
Monique Gambin	9
Inga Svensson	10
Piet van der Keuken	2
Floor de Vries	6

Let's look at the first column in Table 1.1. We see that it regards the variable **name**. We call the property **name** a variable, because it varies across our units (the students): in this case, every unit has a different value for the variable

name. In sum, a variable is a property of units that shows different values for different units.

The second column represents the variable **grade**. Grade is here a variable, because it takes different values for different students. Note that both Mark Zimmerman and Mohammed Solmaz have the same value for this variable.

What we see in Table 1.1 is called a *data matrix*: it is a matrix (a collection of rows and columns) that contains information on units (in the rows) in the form of variables (in the columns).

A unit is something we'd like to say something about. For example, I might want to say something about students and how they score on a course. In that case, students are my *units of analysis*.

If my interest is in schools, the data matrix in Table 1.2 might be useful, which shows a different row for each school with a couple of variables. Here again, we see a variable for grade on a course, but now averaged per school. In this case, school is my unit of analysis.

Table 1.2: Data matrix on schools.

school	number_students	grade_average	teacher
1	5	6.1	Alice Monroe
2	8	5.9	Daphne Stuart
3	5	6.9	Stephanie Morrison
4	9	5.9	Clark Davies
5	10	6.4	David Sanchez Gomez
6	2	6.1	Metin Demirci
7	6	5.2	Frederika Karlsson
8	9	6.8	Advika Agrawal

1.2 Data matrices in R

In R, data matrices are called data frames. A data frame consists of different vectors, one vector for each variable, and each vector contains values. Each vector/variable is stored as a column in a data frame. In the tidyverse version of R that we use in this book, we work with a particular form of a data frame: a tibble. Below we see some R code that creates a tibble: we first load the tidyverse package, then we create the vectors studentID, course, grade, and shirtsize, and then combine these 4 vectors into a tibble.

```
library(tidyverse)
studentID <- seq(4132211, 4132215)
course <- c("Chemistry", "Physics", "Math", "Math", "Chemistry")
grade <- c(4, 6, 3, 6, 8)
shirtsize <- c("medium", "small", "large", "medium", "small")
tibble(studentID, course, shirtsize, grade)
```

```
## # A tibble: 5 x 4
##   studentID course    shirtsize grade
##       <int> <chr>     <chr>     <dbl>
## 1   4132211 Chemistry medium        4
## 2   4132212 Physics   small         6
## 3   4132213 Math      large         3
## 4   4132214 Math      medium        6
## 5   4132215 Chemistry small         8
```

From the output, you see that the tibble has dimensions 5×4: that means it has 5 rows (units) and 4 columns (variables). Under the variable names, it can be seen how the data are stored. The variable studentID is stored as a numeric variable, more specifically as an integer ($< int >$). The course variable is stored as a character variable ($< chr >$), because the values consist of text. The same is true for shirtsize. The last variable, grade, is stored as $< dbl >$ which stands for 'double'. Whether a numeric variable is stored as integer or double depends on the amount of computer memory that is allocated to a variable. Double variables have a decimal part (e.g., 2.0), integers don't (e.g., 2).

1.3 Multiple observations: wide format and long format data matrices

In many instances, units of analysis are observed more than once. This means that we have more than one observation for the *same* variable for the *same* unit of analysis. Storing this information in the rows and columns of a data matrix can be done in two ways: using *wide format* or using *long format*. We first look at wide format, and then see that generally, long format is to be preferred.

Suppose we measure depression levels in four men four times during cognitive behavioural therapy. Sometimes you see data presented in the way of Table 1.3, where there are four separate variables for depression level, one for each measurement: **depression_1**, **depression_2**, **depression_3**, and **depression_4**.

Table 1.3: Data matrix with depression levels in wide format.

client	depression_1	depression_2	depression_3	depression_4
1	5	6	9	3
2	9	5	8	7
3	9	0	9	3
4	9	2	8	6

This way of representing data on a variable that was measured more than once is called *wide format*. We call it *wide* because we simply add columns when we have more measurements, which increases the width of the data matrix. Each new observation of the same variable on the same unit of analysis leads to a new column in the data matrix.

Table 1.4: Data matrix with depression levels in long format.

client	time	depression
1	1	5
1	2	6
1	3	9
1	4	3
2	1	9
2	2	5
2	3	8
2	4	7
3	1	9
3	2	0
3	3	9
3	4	3
4	1	9
4	2	2
4	3	8
4	4	6

Note that this is only one way of looking at this problem of measuring depression four times. Here, you can say that there are really four depression variables: there is depression measured at time point 1, there is depression measured at time point 2, and so on, and these four variables vary only across units of analysis. This way of thinking leads to a wide format representation.

An alternative way of looking at this problem of measuring depression four times, is that depression is really only one variable and that it varies across units of analysis (some people are more depressed than others) and that it *also* varies across time (at times you feel more depressed than at other times).

Therefore, instead of adding columns, we could simply stick to one variable and only add rows. That way, the data matrix becomes longer, which is the reason that we call that format *long format*. Table 1.4 shows the same information from Table 1.3, but now in long format. Instead of four different variables, we have only one variable for depression level, and one extra variable **time** that indicates to which time point a particular depression measure refers to. Thus, both Tables 1.3 and 1.4 tell us that the second depression measure for client number 3 was 0.

Now let's look at a slightly more complex example, where the advantage of long format becomes clear. Suppose the depression measures were taken on different days for different clients. Client 1 was measured on Monday, Tuesday, Wednesday and Thursday, while client 2 was measured on Thursday, Friday, Saturday and Sunday. If we would put that information into a wide format table, it would look like Figure 1.5, with missing values for measures on Monday thru Wednesday for client 2, and missing values for measures on Friday thru Sunday for patient 1.

Table 1.5: Data matrix with depression levels in wide format.

client	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1	5	6	9	3			
2				9	5	8	7

Table 1.6 shows the same data in long format. The data frame is considerably smaller. Imagine that we would also have weather data for the days these patients were measured: whether it was cloudy or sunny, whether it rained or not, and what the maximum temperature was. In long format, storing that information is easy, see Table 1.7. Try and see if you can think of a way to store that information in a wide table!

Table 1.6: Data matrix with depression levels in long format.

client	depression	day
1	5	Monday
1	6	Tuesday
1	9	Wednesday
1	3	Thursday
2	9	Thursday
2	5	Friday
2	8	Saturday
2	7	Sunday

Table 1.7: Data matrix with depression levels in wide format, including data on the time of measurement.

client	depression	day	maxtemp	rain
1	5	Monday	23	rain
1	6	Tuesday	24	no rain
1	9	Wednesday	23	rain
1	3	Thursday	25	no rain
2	9	Thursday	25	no rain
2	5	Friday	22	no rain
2	8	Saturday	21	rain
2	7	Sunday	22	no rain

Thus, storing data in long format is often more efficient in terms of storage of information. Another reason for preferring long format over wide format is the most practical one for data analysis: when analysing data using linear models, software packages require your data to be in long format. In this book, all the analyses with linear models require your data to be in long format. However, we will also come across some analyses apart from linear models that require your data to be in wide format. If your data happen to be in the wrong format,

rearrange your data first. Of course you should never do this by hand as this will lead to typing errors and would take too much time. Statistical software packages have helpful tools for rearranging your data from wide format to long format, and vice versa.

1.4 Wide and long format in R

Making a data matrix longer or wider can be done with the functions pivot_longer() and pivot_wider(), respectively. These functions are part of the tidyr package, and available when you load the tidyverse collection of packages.

```
library(tidyverse)
```

1.4.1 From wide to long

The relig_income dataset stores counts based on a survey which (among other things) asked people about their religion and annual income:

```
relig_income
```

```
## # A tibble: 18 x 11
##    religion `<$10k` `$10-20k` `$20-30k` `$30-40k` `$40-50k` `$50-75k` `$75-100k`
##    <chr>      <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>      <dbl>
##  1 Agnostic      27        34        60        81        76       137        122
##  2 Atheist       12        27        37        52        35        70         73
##  3 Buddhist      27        21        30        34        33        58         62
##  4 Catholic     418       617       732       670       638      1116        949
##  5 Dont k~      15        14        15        11        10        35         21
##  6 Evangel~     575       869      1064       982       881      1486        949
##  7 Hindu          1         9         7         9        11        34         47
##  8 Histori~     228       244       236       238       197       223        131
##  9 Jehovah~      20        27        24        24        21        30         15
## 10 Jewish        19        19        25        25        30        95         69
## 11 Mainlin~     289       495       619       655       651      1107        939
## 12 Mormon        29        40        48        51        56       112         85
## 13 Muslim         6         7         9        10         9        23         16
## 14 Orthodox      13        17        23        32        32        47         38
## 15 Other C~       9         7        11        13        13        14         18
## 16 Other F~      20        33        40        46        49        63         46
## 17 Other W~       5         2         3         4         2         7          3
## 18 Unaffil~     217       299       374       365       341       528        407
## # ... with 3 more variables: `$100-150k` <dbl>, `>150k` <dbl>, `Don't
## #   know/refused` <dbl>
```

This dataset contains three variables:

1. religion, stored in the rows,

2. income spread across the column names, and

3. count stored in the cell values.

To put the values that we see in the columns into one single column, we use pivot_longer():

```
relig_income %>%
  pivot_longer(cols = -religion, # columns that need to be restructured
               names_to = "income", # name of new variable with old column names
               values_to = "count") # name of new variable with values
```

```
## # A tibble: 180 x 3
##    religion income             count
##    <chr>    <chr>              <dbl>
##  1 Agnostic <$10k                 27
##  2 Agnostic $10-20k               34
##  3 Agnostic $20-30k               60
##  4 Agnostic $30-40k               81
##  5 Agnostic $40-50k               76
##  6 Agnostic $50-75k              137
##  7 Agnostic $75-100k             122
##  8 Agnostic $100-150k            109
##  9 Agnostic >150k                 84
## 10 Agnostic Don't know/refused    96
## # ... with 170 more rows
```

- The *cols* argument describes which columns need to be reshaped. In this case, it is every column except religion.

- The *names_to* argument gives the name of the variable that will be created using the column names, i.e. income.

- The *values_to* argument gives the name of the variable that will be created from the data stored in the cells, i.e. count.

1.4.2 From long to wide

The us_rent_income dataset contains information about median income and rent for each state in the US for 2017 (from the American Community Survey, retrieved with the tidycensus package).

```
us_rent_income
```

```
## # A tibble: 104 x 5
##    GEOID NAME       variable estimate   moe
##    <chr> <chr>      <chr>       <dbl> <dbl>
##  1 01    Alabama    income      24476   136
##  2 01    Alabama    rent          747     3
##  3 02    Alaska     income      32940   508
##  4 02    Alaska     rent         1200    13
##  5 04    Arizona    income      27517   148
##  6 04    Arizona    rent          972     4
##  7 05    Arkansas   income      23789   165
##  8 05    Arkansas   rent          709     5
##  9 06    California income      29454   109
## 10 06    California rent         1358     3
## # ... with 94 more rows
```

Here both estimate and moe are values columns, so we can supply them to the function argument values_from:

```
us_rent_income %>%
  pivot_wider(names_from = variable,
              values_from = c(estimate, moe))

## # A tibble: 52 x 6
##    GEOID NAME                 estimate_income estimate_rent moe_income moe_rent
##    <chr> <chr>                          <dbl>         <dbl>      <dbl>    <dbl>
##  1 01    Alabama                        24476           747        136        3
##  2 02    Alaska                         32940          1200        508       13
##  3 04    Arizona                        27517           972        148        4
##  4 05    Arkansas                       23789           709        165        5
##  5 06    California                     29454          1358        109        3
##  6 08    Colorado                       32401          1125        109        5
##  7 09    Connecticut                    35326          1123        195        5
##  8 10    Delaware                       31560          1076        247       10
##  9 11    District of Columbia           43198          1424        681       17
## 10 12    Florida                        25952          1077         70        3
## # ... with 42 more rows
```

- The *names_from* argument gives the name of the variable that will be used for the new column names, i.e. variable
- The *values_from* argument gives the name(s) of the variable(s) that store the value that you wish to see spread out across several columns. Here we have two such variables, i.e. moe and estimate

For more examples, see the vignette on pivoting.

```
vignette("pivot")
```

1.5 Measurement level

Data analysis is about variables and the relationships among them. In essence, data analysis is about describing how different values in one variable go together with different values in one or more other variables (co-variation). For example, if we have the variable age with values 'young' and 'old', and the variable happiness with values 'happy' and 'unhappy', we'd like to know whether 'happy' mostly comes together with either 'young' or 'old'. Therefore, data analysis is about variation and co-variation in variables.

Linear models are important tools when describing co-varying variables. When we want to use linear models, we need to distinguish between different kinds of variables. One important distinction is about the measurement level of the variable: numeric, ordinal or categorical.

1.5.1 Numeric variables

Numeric variables have values that describe a measurable quantity as a number, like 'how many' or 'how much'. A numeric variable can be a *count variable*, for instance the number of children in a classroom. A count variable can only consist of discrete, natural numbers: 0, 1, 2, 3, etcetera. But a numeric variable can also be a *continuous variable*. Continuous variables can take any value from the set of real numbers, for instance values like -200.765, -9.78, -2, 0.001, 4, and 7.8. The number of decimals can be as large as the instrument of measurement allows. Examples of continuous variables include height, time, age, blood pressure and temperature. Note that in all these examples, *quantities* (age, height, temperature) are expressed as the number of a particular *measurement unit* (years, inches, degrees).

Whether a numeric variable is a count variable or a continuous variable, it is always expressing a *quantity*, and therefore numeric variables can be called *quantitative* variables.

For numeric variables, there is a further distinction between *interval variables* and *ratio variables*. The distinction is rather technical. The difference between interval and ratio variables is that for ratio variables, the ratio between two measurement values is meaningful, and for interval variables it is not. An example of a ratio variable is height. You could measure height in two persons where one measures 1 meter and the other measures 2 meters. It is then meaningful to say that the second person is twice as tall as the first person. This is meaningful, because had we chosen a different measurement unit, the ratio would be the same. For instance, suppose we express the heights of the two persons in inches, we would get 39.37 and 78.74 inches respectively. The ratio remains 2: namely 78.74/39.37. The same ratio would hold for measurements

in feet, miles, millimetres or even light years. Thus, whatever the unit of measurement you use, the ratio of height for these individuals would always be 2. Therefore, if we have a variable that measures height in meters, we are dealing with a ratio variable.

Now let's look at an example of an interval variable. Suppose we measure the temperature in two classrooms: one is 10 degrees Celsius and the other is 20 degrees Celsius. The ratio of these two temperatures is $20/10 = 2$, but does that ratio convey meaningful information? Could we state for example that the second classroom is twice as warm as the first classroom? The answer is no, and the reason is simple: had we expressed temperature in Fahrenheit, we would have gotten a very different ratio. Temperatures of 10 and 20 degrees Celsius correspond to 50 and 68 degrees Fahrenheit, respectively. These Fahrenheit temperatures have a ratio of 68/50=1.36. Based on the Fahrenheit metric, the second classroom would now be 1.36 times warmer than the first classroom. We therefore say that the ratio does not have a meaningful interpretation, since the ratio depends on the metric system that you use (Fahrenheit or Celsius). It would be strange to say that there is twice more warmth in classroom B than in classroom A, but only if you measure temperature in Celsius, not when you measure it in Fahrenheit!

The reason why the ratios depend on the metric system, is because both the Celsius and Fahrenheit metrics have arbitrary zero-points. In the Celsius metric, 0 degrees does not mean that there is no warmth, nor is that implied in the Fahrenheit metric. In both metrics, a value of 0 is still warmer than a value of -1.

Contrasting this to the example of height: a height of 0 is indeed the absence of height, as you would not even be able to see a person with a height of 0, whatever metric you would use. Thus, the difference between ratio and interval variables is that ratio variables have a meaningful zero point where zero indicates the absence of the quantity that is being measured. This meaningful zero-point makes it possible to make meaningful statements about ratios (e.g., 4 is twice as much as 2) which gives ratio variables their name.

What ratio and interval variables have in common is that they are both numeric variables, expressing quantities in terms of units of measurements. This implies that the distance between 1 and 2 is the same as the distances between 3 and 4, 4 and 5, etcetera. This distinguishes them from ordinal variables.

1.5.2 Ordinal variables

Ordinal variables are also about quantities. However, the important difference with numeric variables is that ordinal variables are not measured in units. An example would be a variable that would quantify size, by stating whether a T-shirt is small, medium or large. Yes, there is a quantity here, size, but there is no unit to state *exactly* how much of that quantity is present in that T-shirt.

Even though ordinal variables are not measured in specific units, you can still have a meaningful order in the values of the variable. For instance, we know

that a large T-shirt is larger than a medium T-shirt, and a medium T-shirt is larger than a small T-shirt.

Similar for age, we could code a number of people as young, middle-aged or old, but on the basis of such a variable we could not state by *how much* two individuals differ in age. As opposed to numeric variables that are often continuous, ordinal variables are usually *discrete*: there isn't an infinite number of levels of the variable. If we have sizes small, medium and large, there are no meaningful other values in between these values.

Ordinal variables often involve subjective measurements. One example would be having people rank five films by preference from one to five. A different example would be having people assess pain: "On a scale from 1 to 10, how bad is the pain?"

Ordinal variables often look numeric. For example, you may have large, medium and small T-shirts, but these values may end up in your data matrix as '3', '2' and '1', respectively. However, note that with a truly numeric variable there should be a unit of measurement involved (3 of what? 2 of what?), and that numeric implies that the distance between 3 and 2 is equal to the distance between 2 and 1. Here you would not have that information: you only know that a large T-shirt (coded as '3') is larger than a medium T-shirt (coded as '2'), but how large that difference is, and whether that difference is that same as the difference between a medium T-shirt ('2') is larger than a small T-shirt ('1'), you do not know. Therefore, even though we see numbers in our data matrix, the variable is called an ordinal variable.

1.5.3 Categorical variables

Categorical variables are not about quantity at all. Categorical variables are about *quality*. They have values that describe 'what type' or 'which category' a unit of belongs to. For example, a school could either be publicly funded or not, or a person could either have the Swedish nationality or not. A variable that indicates such a dichotomy between publicly funded 'yes' or 'no', or Swedish nationality 'yes' or 'no', is called a *dichotomous* variable, and is a subtype of a categorical variable. The other subtype of a categorical variable is a *nominal* variable. Nominal comes from the Latin *nomen*, which means name. When you name the nationality of a person, you have a nominal variable. Table 1.8 shows an example of both a dichotomous variable (Swedish) that always has only two different values, and a nominal variable (Nationality), that can have as many different values as you want (usually more than two).

Another example of a nominal variable could be the answer to the question: "name the colours of a number of pencils". Nothing quantitative could be stated about a bunch of pencils that are only assessed regarding their colour. In addition, there is usually no logical order in the values of such variables, something that we do see with ordinal variables.

Table 1.8: Nationalities.

ID	Swedish	Nationality
1	Yes	Swedish
2	Yes	Swedish
3	No	Angolan
4	No	Norwegian
5	Yes	Swedish
6	Yes	Swedish
7	No	Danish
8	No	Unknown

1.5.4 Treatment of variables in data analysis

For data analysis with linear models, you have to decide for each variable whether you want to treat it as numeric or as categorical.[1] The easiest choice is for numeric variables: numeric variables should always be treated as numeric.

Categorical data should always be treated as categorical. However, the problem with categorical variables is that they often *look* like numeric variables. For example, take the categorical variable country. In your data file, this variable could be coded with strings like "Netherlands", "Belgium", "Luxembourg", etc. But the variable could also be coded with numbers: 1, 2 and 3. In a codebook that belongs to a data file, it could be stated that 1 stands for "Netherlands", 2 for "Belgium", and 3 for "Luxembourg" (these are the value labels), but still in your data matrix your variable would look numeric. You then have to make sure that, even though the variable *looks* numeric, it should be *interpreted* as a categorical variable and therefore be *treated* like a categorical variable.

The most difficult problem involves ordinal variables: in linear models you can either treat them as numeric variables or as categorical variables. The choice is usually based on common sense and whether the results are meaningful. For instance, if you have an ordinal variable with 7 levels, like a Likert scale, the variable is often coded with numbers 1 through 7, with value labels 1="completely disagree", 2="mostly disagree", 3="somewhat disagree", 4="ambivalent", 5="somewhat agree", 6="mostly agree", and 7="completely agree". In this example, you could choose to treat this variable as a categorical variable, recognising that this is not a numeric variable as there is no measurement unit. However, if you feel this is awkward, you could choose to treat the variable as numeric, but be aware that this implies that you feel that the difference between 1 and 2 is the same as the difference between 2 and 3. In general, with ordinal data like Likert scales or sizes like, Small, Medium and Large, one generally chooses to use categorical treatment for low numbers of categories, say 3 or 4 categories, and numerical treatment for variables with many categories, say 5 or more. However, this should not be used as a rule of thumb: first think about the meaning of your variable and the objective of your data analysis project,

[1] In data analysis, it is possible to treat variables as ordinal, but only in more advanced models and methods than treated in this book.

and only then take the most reasonable choice. Often, you can start with numerical treatment, and if the analysis shows peculiar results[2], you can choose categorical treatment in secondary analyses.

In the coming chapters, we will come back to the important distinction between categorical and numerical treatment (mostly in Chapter **??**). For now, remember that numeric variables are always treated as numeric variables, categorical variables are always treated as categorical variables (even when they appear numeric), and that for ordinal variables you have to think before you act.

1.6 Measurement level in R

In a previous section we saw the creation of a data frame. Let's store the resulting data frame as an object called course_results.

```
studentID <- seq(4132211, 4132215)
course <- c("Chemistry", "Physics", "Math", "Math", "Chemistry")
grade <- c(4, 6, 3, 6, 8)
shirtsize <- c("medium", "small", "large", "medium", "small")
course_results <- tibble(studentID, course, shirtsize, grade)
course_results
```

```
## # A tibble: 5 x 4
##   studentID course    shirtsize grade
##       <int> <chr>     <chr>     <dbl>
## 1   4132211 Chemistry medium        4
## 2   4132212 Physics   small         6
## 3   4132213 Math      large         3
## 4   4132214 Math      medium        6
## 5   4132215 Chemistry small         8
```

We see that the variable studentID is stored as integer. That means that the values are stored as numeric values. However, the values are quite meaningless, they are only used to identify persons. If we want to treat this variable as a categorical variable in data analysis, it is necessary to change this variable into a factor variable. We can do this by typing:

```
course_results$studentID <-
  course_results$studentID %>%
  factor()
```

When we look at this variable after the transformation, we see that this new categorical variable has 5 different categories (levels).

[2]For instance, you may find that the assumptions of your linear model are not met, see Chapter **??**.

```
course_results$studentID
```

```
## [1] 4132211 4132212 4132213 4132214 4132215
## Levels: 4132211 4132212 4132213 4132214 4132215
```

When we look at the variable course, we see that it is stored as a character variable. If we want R to treat it as a categorical variable in data analysis, we can also transform this variable into a factor variable. We could use the same code as above, or we could use the function mutate().

```
course_results <- course_results %>%
  mutate(course = factor(course))
```

The shirtsize variable is stored as character, but we tell R that this is an ordinal variable. For this we need to turn it into a factor variable, indicating that there is an order in the values, where small is the lowest quantity, and large the highest quantity.

```
course_results <- course_results %>%
  mutate(shirtsize = factor(shirtsize,
                            levels = c("small", "medium", "large"),
                            ordered = TRUE)
  )
course_results$shirtsize
```

```
## [1] medium small  large  medium small
## Levels: small < medium < large
```

The last variable grade is stored as double. Variables of this type will be treated as numeric in data analyses. If we're fine with that for this variable, we leave it as it is. If we want the variable to be treated as ordinal, then we need the same type of factor transformation as for shirtsize. For now, we leave it as it is. The resulting data frame then looks like this:

```
course_results
```

```
## # A tibble: 5 x 4
##   studentID course    shirtsize grade
##   <fct>     <fct>     <ord>     <dbl>
## 1 4132211   Chemistry medium        4
## 2 4132212   Physics   small         6
## 3 4132213   Math      large         3
## 4 4132214   Math      medium        6
## 5 4132215   Chemistry small         8
```

Now both studentID and course are stored as factors and will be treated as categorical. Shirtsize is stored as an ordinal factor and will be treated accordingly. Grade is still stored as double and will therefore be treated as numeric.

1.7 Frequency tables, frequency plots and histograms

Variables have different values. For example, age is a (numeric, ratio) variable: lots of people have different ages. Suppose we have an imaginary town with 1000 children. For each age measured in years, we can count the number of children who have that particular age. The results of the counting are in Table 1.9. The number of observed children with a certain age, say 8 years, is called the *frequency* of age 8. The table is therefore called a frequency table. Generally in a frequency table, values that are not observed are omitted (i.e., the frequency of children with age 16 is 0).

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Table 1.9: Frequency table for age, with proportions and cumulative proportions.

age	frequency	proportion	cum_frequency	cum_proportion
0	2	0.002	2	0.002
1	7	0.007	9	0.009
2	20	0.020	29	0.029
3	50	0.050	79	0.079
4	105	0.105	184	0.184
5	113	0.113	297	0.297
6	159	0.159	456	0.456
7	150	0.150	606	0.606
8	124	0.124	730	0.730
9	108	0.108	838	0.838
10	70	0.070	908	0.908
11	34	0.034	942	0.942
12	32	0.032	974	0.974
13	14	0.014	988	0.988
14	9	0.009	997	0.997
15	2	0.002	999	0.999
17	1	0.001	1000	1.000

The data in the frequency table can also be represented using a frequency plot. Figure 1.1 gives the same information, not in a table but in a graphical way. On the horizontal axis we see several possible values for age in years, and on the vertical axis we see the number of children (the count) that were observed for each particular age. Both the frequency table and the frequency plot tell us something about the *distribution* of age in this imaginary town with 1000 children. For example, both tell us that the oldest child is 17 years old. Furthermore, we see that there are quite a lot of children with ages between 5 and 8, but not so many children with ages below 3 or above 14. The advantage

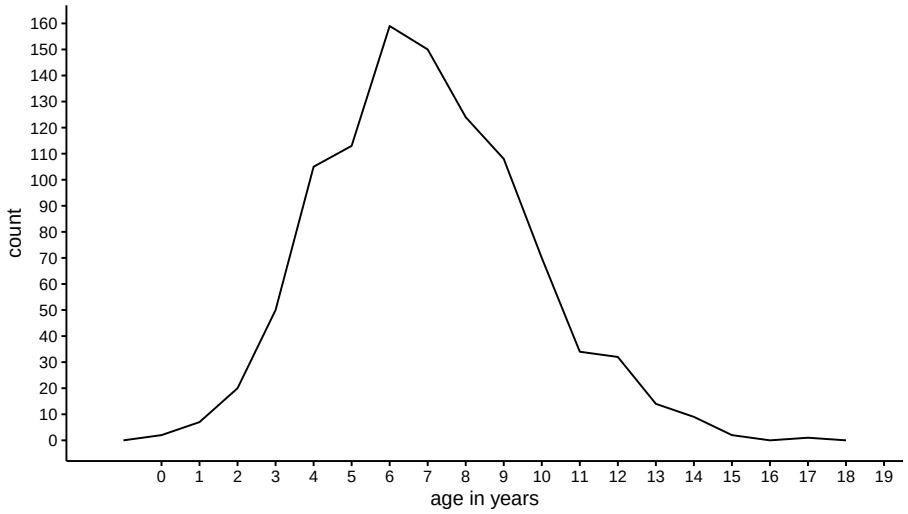


Figure 1.1: A frequency plot

of the table over the graph is that we can get the exact number of children of a particular age very easily. But on the other hand, the graph makes it easier to get a quick idea about the shape of the distribution, which is hard to make out from the table.

Instead of frequency plots, one often sees *histograms*. Histograms contain the same information as frequency plots, except that *groups of values* are taken together. Such a group of values is called a *bin*. Figure 1.2 shows the same age data, but uses only 9 bins: for the first bin, we take values of age 0 and 1 together, for the second bin we take ages 2 and 3 together, etcetera, until we take ages 16 and 17 together for the last bin. For each bin, we compute how often we observe the ages in that bin.

Histograms are very convenient for continuous data, for instance if we have values like 3.473, 2.154, etcetera. Or, more generally, for variables with values that have very low frequencies. Suppose that we had measured age not in years but in days. Then we could have had a data set of 1000 children where each and every child had a unique value for age. In that case, the length of the frequency table would be 1000 rows (each value observed only once) and the frequency plot would be very flat. By using age measured in years, what we have actually done is putting all children with an age less than 365 days into the first bin (age 0 years) and the children with an age of at least 365 but less than 730 days into the second bin (age 1 year). And so on. Thus, if you happen to have data with many many values with very low frequencies, consider binning the data, and using a histogram to visualise the distribution of your numeric variable.

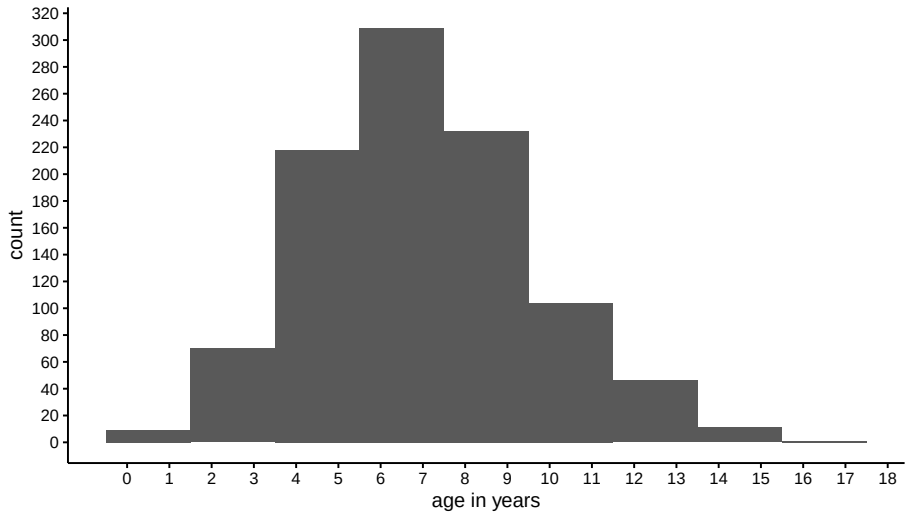


Figure 1.2: A histogram

1.8 Frequencies, proportions and cumulative frequencies and proportions

When we have the frequency for each observed age, we can calculate the *relative frequency* or *proportion* of children that have that particular age. For example, when we look again at the frequencies in Table 1.9 we see that there are two children who have age 0. Given that there are in total 1000 children, we know that the *proportion* of people with age 0 equals $\frac{2}{1000} = 0.002$. Thus, the proportion is calculated by taking the frequency and dividing it by the total number.

We can also compute *cumulative frequencies*. You get cumulative frequencies by accumulating (summing) frequencies. For instance, the cumulative frequency for the age of 3, is the frequency for age 3 plus all frequencies for younger ages. Thus, the cumulative frequency of age 3 equals 50 + 20 (for age 2) + 7 (for age 1) + 2 (for age 0) = 79. The cumulative frequencies for all ages are presented in Table 1.9.

We can also compute *cumulative proportions*: if we take for each age the proportion of people who have that age *or less*, we get the fifth column in Table 1.9. For example, for age 2, we see that there are 20 children with an age of 2. This corresponds to a proportion of 0.020 of all children. Furthermore, there are 9 children who have an even younger age. The proportion of children with an age of 1 equals 0.007, and the proportion of children with an age of 0 equals 0.002. Therefore, the proportion of all children with an age of 2 or less equals $0.020 + 0.007 + 0.002 = 0.029$, which is called the cumulative proportion for the age of 2.

1.9 Frequencies and proportions in R

The mtcars data set contains information about a number of cars: miles per gallon (mpg), number of cylinders (cyl), etcetera.

```
mtcars
```

```
##                     mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4          21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag      21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710         22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive     21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout  18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant            18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360         14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D          24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230           22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280           19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C          17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE         16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL         17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC        15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood 10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4  8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial  14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128           32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic        30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla     33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona      21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger   15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin        15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28         13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird   19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9          27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2      26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa       30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L     15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino       19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora      15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E         21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

The object is a data frame. We can turn it into a tibble as follows:

```
mtcars <- mtcars %>% as_tibble()
```

The function as_tibble() is available when you load the tidyverse package. From now on, we assume that you load the tidyverse package at the start of

every R session.

If we want to know how many cars belong to which category of number of cylinders, we can use the function count():

```
mtcars %>%
  count(cyl)

## # A tibble: 3 x 2
##     cyl     n
##   <dbl> <int>
## 1     4    11
## 2     6     7
## 3     8    14
```

The new variable **n** is the frequency. We see that the value 4 occurs 11 times, the value 6 occurs 7 times and the value 8 occurs 14 times. Thus, in this data set there are 11 cars with 4 cylinders, 7 cars with 6 cylinders, and 14 cars with 8 cylinders.

We obtain proportions when we divide the frequencies by the total number of cars (the sum of all the values in the **n** variable):

```
mtcars %>%
  count(cyl) %>%
  mutate(proportion = n/sum(n))

## # A tibble: 3 x 3
##     cyl     n proportion
##   <dbl> <int>      <dbl>
## 1     4    11      0.344
## 2     6     7      0.219
## 3     8    14      0.438
```

Cumulative frequencies and cumulative proportions can be obtained using the cumsum() function:

```
mtcars %>%
  count(cyl) %>%
  mutate(proportion = n/sum(n)) %>%
  mutate(cumfreq = cumsum(n),
         cumprop = cumsum(proportion))

## # A tibble: 3 x 5
##     cyl     n proportion cumfreq cumprop
##   <dbl> <int>      <dbl>   <int>   <dbl>
## 1     4    11      0.344      11   0.344
## 2     6     7      0.219      18   0.562
## 3     8    14      0.438      32   1
```

A frequency plot can be made using ggplot combined with geom_line():

```
mtcars %>%
  count(cyl) %>%
  mutate(proportion = n/sum(n)) %>%
  ggplot(aes(x = cyl, y = n)) +
  geom_line()
```

A histogram of the **mpg** variable can be made using geom_histogram():

```
mtcars %>%
  ggplot(aes(x = mpg)) +
  geom_histogram(breaks = seq(5, 40, 5))
```

It is wise to play around with the number of bins that you'd like to make, or with the boundaries of the bins. Here we choose boundaries $5, 10, 15, \ldots, 40$.

1.10 Quartiles, quantiles and percentiles

Suppose we want to split the group of 1000 children into 4 equally-sized subgroups, with the 25% youngest children in the first group, the 25% oldest children in the last group, and the remaining 50% of the children in two equally sized middle groups. What ages should we then use to divide the groups? First, we can order the 1000 children on the basis of their age: the youngest first, and the oldest last. We could then use the concept of *quartiles* (from quarter, a fourth) to divide the group in four. In order to break up all ages into 4 subgroups, we need 3 points to make the division, and these three points are called quartiles. The first quartile is the value below which 25% of the observations fall, the second quartile is the value below which 50% of the observations fall, and the third quartile is the value below which 75% of the observations fall.[3]

Let's first look at a smaller but similar problem. For example, suppose your observed values are 10, 5, 6, 21, 11, 1, 7, 9. You first order them from low to high so that you obtain 1, 5, 6, 7, 9, 10, 11, 21. You have 8 values, so the first 25% of your values are the first two. The highest value of these two equals 5, and this we define as our first quartile.[4] We find the second quartile by looking at the values of the first 50% of the observations, so 4 values. The first 4 values are 1, 5, 6, and 7. The last of these is 7, so that is our second quartile. The first 75% of the observations are 1, 5 ,6 ,7 , 9, and 10. The value last in line is 10, so our fourth quartile is 10.

[3]The fourth quartile would be the value below which *all* values are, so that would be the largest value in the row (the age of the last child in the row).

[4]Note that we could also choose to use 6, because 1 and 5 are lower than 6. Don't worry, the method that we show here to compute quartiles is only one way of doing it. In your life, you might stumble upon alternative ways to determine quartiles. These are just arbitrary agreements made by human beings. They can result in different outcomes when you have small data sets, but usually not when you have large data sets.

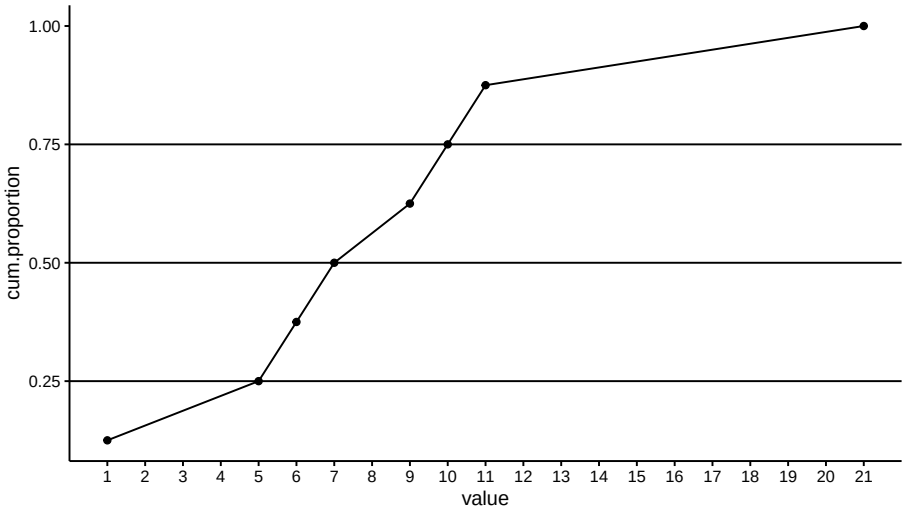


Figure 1.3: Cumulative proportions.

The quartiles as defined here can also be found graphically, using cumulative proportions. Figure 1.3 shows for each observed value the cumulative proportion. It also shows where the cumulative proportions are equal to 0.25, 0.50 and 0.75. We see that the 0.25 line intersects the other line at the value of 5. This is the first quartile. The 0.50 line intersects the other line at a value of 7, and the 0.75 line intersects at a value of 10. The three percentiles are therefore 5, 7 and 10.

If you have a large data set, the graphical way is far easier than doing it by hand. If we plot the cumulative proportions for the ages of the 1000 children, we obtain Figure 1.4. We see a nice S-shaped curve. We also see that the three horizontal quartile lines no longer intersect the curve at specific values, so what do we do? By eye-balling we can find that the first quartile is somewhere between 4 and 5. But which value should we give to the quartile? If we look at the cumulative proportion for an age of 4, we see that its value is slightly below the 0.25 point. Thus, the proportion of children with age 4 or younger is lower than 0.25. This means that the child that happens to be the 250th cannot be 4 years old. If we look at the cumulative proportion of age 5, we see that its value is slightly above 0.25. This means that the proportion of children that is 5 years old or younger is slightly more than 0.25. Therefore, of the the total of 1000 children, the 250th child must have age 5. Thus, by definition, the first quantile is 5. The second quartile is somewhere between 6 an 7, so by using the same reasoning as for the first quartile we know that 50% of the youngest children is 7 years old or younger. The third quartile is somewhere between 8 and 9 and this tells us that the youngest 75% of the children is age 9 or younger. Thus, we can call 5, 7 and 9 our three quartiles.

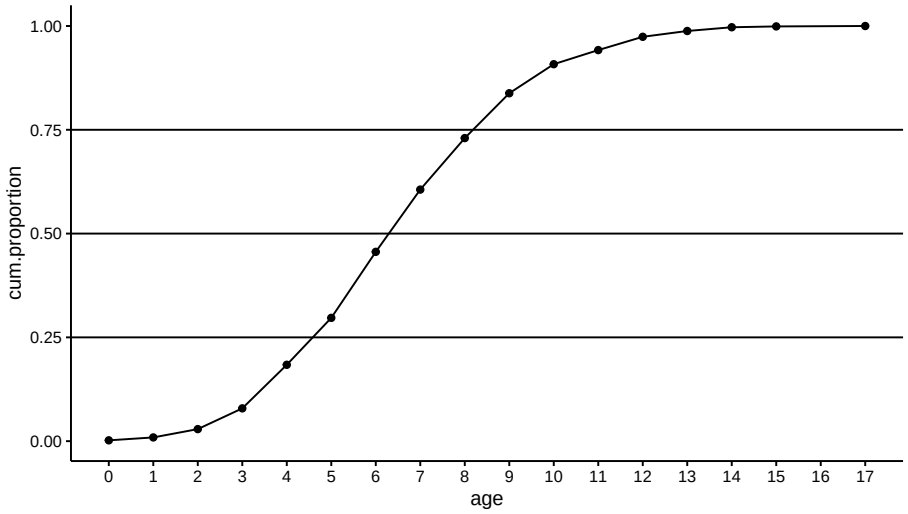


Figure 1.4: Cumulative proportions.

Alternatively, we could also use the frequency table (Table 1.9). First, if we want to have 25% of the children that are the youngest, and we know that we have 1000 children in total, we should have $0.25 \times 1000 = 250$ children in the first group. So if were to put all the children in a row, ordered from youngest to oldest, we want to know the age of the 250th child.

In order to find the age of this 250th child, and we look at Table 1.9, we see that 29.7 % of the children have an age of 5 or less (297 children), and 18.4 % of the children have an age of 4 or less (184 children). This tells us that, since 250 comes after 184, the 250th child must be older than 4, and because 250 comes before 297, it must be younger than or equal to 5, hence the child is 5 years old.

Furthermore, if we want to find a cut-off age for the oldest 25%, we see from the table, that 83.8% of the children (838 children) have an age of 9 or less, and 73.0% of the children (730) have an age of 8 or less. Therefore, the age of the 750th child (when ordered from youngest to oldest) must be 9.

What we just did for quartiles, (i.e. 0.25, 0.50, 0.75) we can do for any proportion between 0 and 1. We then no longer call them quartiles, but *quantiles*. A quantile is the value below which a given proportion of observations in a group of observations fall. From this table it is easy to see that a proportion of 0.606 of the children have an age of 7 or less. Thus, the 0.606 quantile is 7. One often also sees *percentiles*. Percentiles are very much like quantiles, except that they refer to percentages rather than proportions. Thus, the 20th percentile is the same as the 0.20 quantile. And the 0.81 quantile is the same as the 81st percentile.

The reason that quartiles, quantiles and percentiles are important is that they are very short ways of saying something about a distribution. Remember

that the best way to represent a distribution is either a frequency table or a frequency plot. However, since they can take up quite a lot of space sometimes, one needs other ways to briefly summarise a distribution. Saying that "the third quartile is 454" is a condensed way of saying that "75% of the values is either 454 or lower". In the next sections, we look at other ways of summarising information about distributions.

Another way in which quantiles and percentiles are used is to say something about *individuals*, relative to a group. Suppose a student has done a test and she comes home saying she scored in the 76th percentile of her class. What does that mean? Well, you don't know her score exactly, but you do know that of her classmates, 76 percent had the same score or lower. That means she did pretty well, compared to the others, since only 24 percent had a higher score.

1.11 Quantiles in R

Obtaining quartiles, quantiles and percentiles can be done with the quantile() function:

```
quantile(mtcars$mpg,
         probs = c(0.25, 0.50, 0.75, 0.90))

##     25%     50%     75%     90%
## 15.425 19.200 22.800 30.090
```

1.12 Measures of central tendency

The mean, the median and the mode are three different measures that say something about the *central tendency* of a distribution. If you have a series of values: around which value do they tend to cluster?

1.12.1 The mean

Suppose we have the values 1, 2 and 3, then we compute the mean by first adding these numbers and then divide them by the number of values we have. In this case we have three values, so the mean is equal to $(1 + 2 + 3)/3 = 2$. In statistical formulas, the mean of a variable is indicated by a bar above that variable. So if our values of variable Y are 1, 2 and 3, then we denote the mean by $\overline{Y}$ (pronounced as 'y-bar'). When taking the sum of a set of values, statistical formulas show the summation sign Σ (the Greek letter sigma). So we often see the following formula for the mean of a set of n values for variable Y[5]:

$$\overline{Y} = \frac{\Sigma_{i=1}^{n} Y_i}{n} \tag{1.1}$$

[5]Variables are symbolised by capitals, e.g., Y. Specific values of a variable are indicated in lowercase, e.g., y.

In words, in order to compute $\overline{Y}$, we take every value for variable Y from $i = 1$ to $i = n$ and sum them, and the result is divided by n. Suppose we have variable Y with the values 6, -3, and 21, then the mean of Y equals:

$$\overline{Y} = \frac{\Sigma_i Y_i}{n} = \frac{Y_1 + Y_2 + Y_3}{n} = \frac{6 + (-3) + 21}{3} = \frac{24}{3} = 8 \tag{1.2}$$

1.12.2 The median

The mean is only one of the measures of central tendency. An alternative measure of central tendency is the *median*. The median is nothing but the middle value of an ordered series. Suppose we have the values 45, 567, and 23. Then what value lies in the middle when ordered? Let's first order them from small to large to get a better look. We then get 23, 45 and 567. Then it's easy to see that the value in the middle is 45.

Suppose we have the values 45, 45, 45, 65, and 23. What is the middle value when ordered? We first order them again and see what value is in the middle: 23, 45, 45, 45 and 65. Obviously now 45 is the median. You can also see that half of the values is equal or smaller than this value, and half of the values is equal or larger than this value. The median therefore is the same as the second quartile.

What if we have two values in the middle? Suppose we have the values 46, 56, 45 and 34. If we order them we get 34, 45, 46 and 56. Now there are two values in the middle: 45 and 46. In that case, we take the mean of these two middle values, so the median is 45.5.

When do you use a median and when do you use a mean? For numeric variables that have a more or less symmetric distribution (i.e., a frequency plot that is more or less symmetric), the mean is most often used. Actually, for distributions that are more or less symmetric the mean and median are very similar. For numeric variables that do not have a symmetric distribution, it is usually more informative to use the median. An example of such a situation is income. Figure 1.5 shows a typical distribution of yearly income. The distribution is highly asymmetric, it is severely skewed to the right. The bulk of the values are between 20,000 and 40,000, with only a very few extreme values on the high end. Even though there are only a few people with a very high income, the few high values have a huge effect on the mean.

The mean of the distribution turns out to be 23604. The largest value in the distribution is an income of 75051. Imagine what would happen to the mean and the median if we would change only this one value, that is, the highest observed income. Which would be most affected, do you think: the mean or the median?

Well, if we would change this value into 85051, you see an immediate impact on the mean: the mean is then 23614. This means that the mean is very sensitive to extreme values. One single change in a data set can have a huge effect on the mean. The median on the other hand is much more stable. The median remains unaffected by changes in the extremes. This because it only looks at

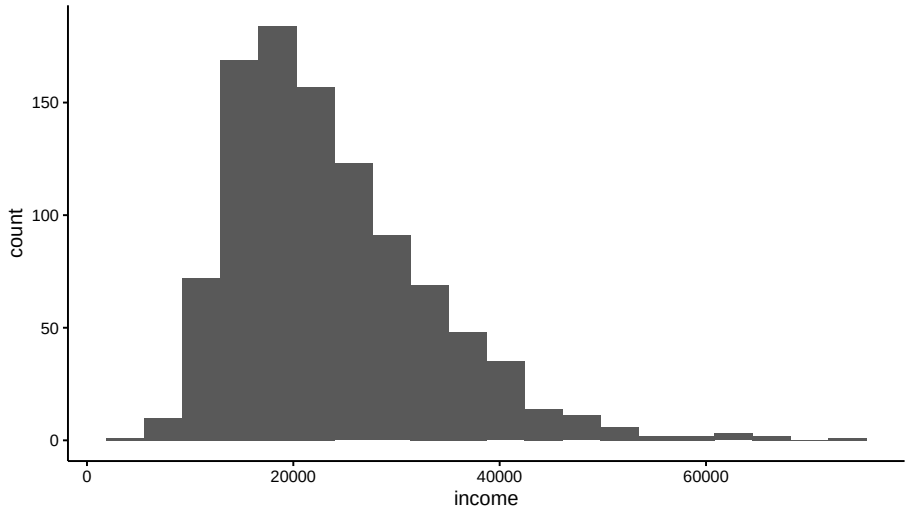


Figure 1.5: Distribution of yearly income.

the middle value. The middle value is unaffected by a change in the extreme values, as long as the order of the values remains the same and the middle value remains the same.

This can be seen even more clearly by looking at the example in Table 1.10. There we have three values, X1, X2 and X3, for which we compute both the mean and the median. First, suppose we have the values 4, 5, and 8 (like in the first row of Table 1.10). Obviously, the median is 5. Next, instead of 4, 5 and 8, we could have values 4, 5 and 80, or 4, 5 and 800, or 4, 5 and 8000. Regardless, the middle value of this series remains 5. In contrast, the mean would be very much affected by having either an 8, an 80, an 800 or an 8000 in the series. In sum: the median is a more stable measure of central tendency than the mean.

Table 1.10: Four series of values and their respective medians and means.

X1	X2	X3	median	mean
4	5	8	5	5.7
4	5	80	5	29.7
4	5	800	5	269.7
4	5	8000	5	2669.7

1.12.3 The mode

A third measure of central tendency is the *mode*. The mode is defined as the value that we see most frequently in a series of values. For example, if we have the series 4, 7, 5, 5, 6, 6, 6, 4, then the value observed most often is 6 (three

times). Modes are easily inferred from frequency tables: the value with the largest frequency is the mode. They are also easily inferred from frequency plots: the value on the horizontal axis for which we see the highest count (on the vertical axis).

The mode can also be determined for categorical variables. If we have the observed values 'Dutch', 'Danish', 'Dutch', and 'Chinese', the mode is 'Dutch' because that is the value that is observed most often.

If we look back at the distribution in Figure 1.5, we see that the peak of the distribution is around the value of 19,000. However, whether this is the mode, we cannot say. Because income is a more or less continuous variable, every value observed in the Figure occurs only once: there is no value of income with a frequency more than 1. So technically, there is no mode. However, if we split the values into 20 bins, like we did for the histogram in Figure 1.5, we see that the fifth bin has the highest frequency. In this bin there are values between 17000 and 21000, so our mode could be around there. If we really want a specific value, we could decide to take the average value in the fifth bin. There are many other statistical tricks to find a value for the mode, where technically there is none. The point is that for the mode, we're looking for the value or the range of values that are most frequent. Graphically, it is the value under the peak of the distribution. Similar to the median, the mode is also quite stable: it is not affected by extreme values and is therefore to be preferred over the mean in the case of asymmetric distributions.

1.13 Relationship between measures of tendency and measurement level

There is a close relationship between measures of tendency and measurement level. For numeric variables, all three measures of tendency are meaningful. Suppose you have the numeric variable age measured in years, with the values 56, 68, 68, 99 and 100. Then it is meaningful to say that the average age is 78.2 years, that the median age is 68 years, and that the mode is 68 years.

For ordinal variables, it is quite different. Suppose you have 5 T-shirts, with the following sizes: M, S, M, L, XL. Then what is the average size? There are no numeric values here to put in the algebraic formula. But we can determine the median: if we order the values from small to large we get the set S, M, M, L, XL and we see that the middle value is M. So M is our median in this case. [6] The other meaningful measure of tendency for ordinal variables is the mode.

For categorical variables, both the mean and the median are pointless to report. Suppose we have the nominal variable Study Programme with observed values "Medicine", "Engineering", "Engineering", "Mathematics", and "Biology". It would be impossible to derive a numerical mean, nor would it be

[6] However, suppose that our collection of T-shirts had the following sizes: S, M, L, L. Then there would be no single middle value in we would have to average the M and L values, which would be impossible!

possible to determine the middle value to determine the median, as there is no logical or natural order.[7] It is meaningful though to report a mode. It would be meaningful to state that the study programme mentioned most often in the news is "Psychology", or that the most popular study programme in India is "Engineering". Thus, for categorical variables, both dichotomous and nominal variables, only the mode is a meaningful measure of central tendency.

As stated earlier, the appearance of a variable in a data matrix can be quite misleading. Categorical variables and ordinal variables can often look like numeric variables, which makes it very tempting to compute means and medians where they are completely meaningless. Take a look at Table 1.11. It is entirely possible to compute the average University, Size, or Programme, but it would be utterly senseless to report these values.

It is entirely possible to compute the median University, Size, or Programme, but it is only meaningful to report the median for the variable Size, as Size is an ordinal variable. Reporting that the median size is equal to 2 is saying that about half of the study programmes is of medium size or small, and about half of the study programmes is of medium size or large.

It is entirely possible to compute the mode for the variables University, Size, or Programme, and it is always meaningful to report them. It is meaningful to say that in your data there is no University that is observed more than others. It is meaningful to report that most study programmes are of medium size, and that most study programmes are study programme number 2 (don't forget to look up and write down which study programme that actually is!).

Table 1.11: Study programmes and their relative sizes (1=small, 2=medium, 3=large) for six different universities.

University	Size	Programme
1	1	2
2	3	2
3	2	3
4	2	3
5	3	4
6	2	1

1.14 Measures of central tendency in R

The mean and median for numeric variables can be obtained as follows:

```
mtcars %>%
  summarise(mean_cyl = mean(cyl),
            median_cyl = median(cyl))
```

[7] Unless you see one? But then it would not be a categorical value but an ordinal variable.

```
## # A tibble: 1 x 2
##   mean_cyl median_cyl
##      <dbl>      <dbl>
## 1     6.19          6
```

R does not have an in-built function to calculate modes. So we create our own function getmode(). This function takes a vector as input and gives the mode value as output.

```
getmode <- function(variable){
  unique_values <- unique(variable)
  unique_values[
    match(variable, unique_values) %>%
      tabulate() %>%
      which.max()
    ]
}

mtcars %>%
  summarise(mode_cyl = getmode(cyl))

## # A tibble: 1 x 1
##   mode_cyl
##      <dbl>
## 1        8
```

1.15 Measures of variation

Above we saw that we can summarise distributions by measures of central tendency. Here we discuss how we can summarise distributions of numeric variables by a measure that describes their *variation*. Variables show variation, by definition, but how much variation do they actually show?

Suppose we measure the height of 3 children, and their heights (in cms) are 120, 120 and 120. There is no variation in height: all heights are the same. There are no differences. Then the average height is 120, the median height is 120, and the mode is 120. The variation is 0: non-existing, absent.

Now suppose their heights are 120, 120, 135. Now there are differences: one child is taller than the other two, who have the same height. There is some variation now. We know how to quantify the mean, which is 125, we know how to quantify the median, which is 120, and we know how to quantify the mode, which is also 120. But how do we quantify the variation? Is there a lot of variation, or just a little, and how do we measure it?

1.15.1 Range and interquartile distance

One thing you could think of is measuring the distance or difference between the lowest value and the highest value. We call this the *range*. The lowest value is 120, and the highest value is 135, so the range of the data is equal to $135 - 120 = 15$. As another example, suppose we have the values 20, 20, 21, 20, 19, 20 and 454. Then the range is equal to $454 - 19 = 435$. That's a large range, for a series of values that for the most part hardly differ from another.

Instead of measuring the distance from the lowest to the highest value, we could also measure the distance between the first and the third quartile: how much does the third quartile *deviate* from the first quartile? This distance or deviation is called the *interquartile distance*. Suppose that we have a large number of systolic blood pressure measurements, where 25% are 120 or lower, and 75% are 147 or lower, then the interquartile distance is equal to $147 - 120 = 27$.

Thus, we can measure variation using the range or the interquartile distance. A third measure for variation is *variance*, and variance is based on the *sum of squares*.

1.15.2 Sum of squares

What we call a sum of squares is actually a sum of squared deviations. But deviations from what? We could for instance be interested in how much the values 120, 120, 135 vary around the mean of these values. The mean of these three values equals 125. The first value differs $120 - 125 = -5$, the second value also differs $120 - 125 = -5$, and the third value differs $135 - 125 = 10$.

Whenever we look at deviations from the mean, some deviations are positive and some deviations will be negative (except when there is no variation). If we want to measure variation, it should not matter whether deviations are positive or negative: any deviation should add to the total variation in a positive way. Moreover, if we would add up all deviations from the mean, we would always end up with 0, as you can see in our example. Adding up -5, -5 and +10 would lead to a sum of 0. This would mean no variation. However, as you can see, there is variation. So that is why we would better make all deviations positive, and this can be done by taking the square of the deviations, since a negative number squared is always positive. So for our three values 120, 120 and 135, we get the deviations -5, -5 and +10, and if we square these deviations, we get 25, 25 and 100. If we add these three squares, we obtain the sum 150.

In most cases, the sum of squares (SS) refers to the sum of squared deviations from the mean. In brief, suppose you have n values of a variable Y, you first take the mean of those values (this is $\overline{Y}$), you subtract this mean from each of these n values ($Y_i - \overline{Y}$), then you take the squares of these deviations, $(Y_i - \overline{Y})^2$, and then add them up (take the sum of these squared deviations, $\Sigma(Y_i - \overline{Y})^2$. In formula form, this process looks like:

$$SS = \Sigma_i^n (Y_i - \overline{Y})^2 \tag{1.3}$$

As an example, suppose you have the values 10, 11 and 12, then the mean is 11. Then the deviations from the mean are -1, 0 and +1. If you square them you get $(-1)^2 = 1$, $0^2 = 0$ and $(+1)^2 = 1$, and if you sum these three values, you get $SS = 1 + 0 + 1 = 2$. In formula form:

$$\begin{aligned} SS &= (Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + (Y_3 - \overline{Y})^2 \\ &= (10 - 11)^2 + (11 - 11)^2 + (12 - 11)^2 = (-1)^2 + 0^2 + 1^2 = 2 \end{aligned} \tag{1.4}$$

Now let's use some values that are more different from each other, but with the same mean. Suppose you have the values 9, 11 and 13. The average value is still 11, but the deviations from the mean are larger. The deviations from 11 are -2, 0 and +2. Taking the squares, you get $(-2)^2 = 4$, $0^2 = 0$ and $(+2)^2 = 4$ and if you add them you get $SS = 4 + 0 + 4 = 8$.

$$\begin{aligned} SS &= (Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + (Y_3 - \overline{Y})^2 \\ &= (9 - 11)^2 + (11 - 11)^2 + (13 - 11)^2 = (-2)^2 + 0^2 + 2^2 = 8 \end{aligned} \tag{1.5}$$

Thus, the more the values differ from each other, the larger the deviations from the mean. And the larger the deviations from the mean, the larger the sum of squares. The sum of squares is therefore a nice measure of how much values differ from each other.

1.15.3 Variance and standard deviation

The sum of squares can be seen as a measure of total variation: all (squared) deviations from a certain value are added up. This means that the more data values you have, the larger the sum of squares. Often-times, you are not interested in the total variation, but you're interested in the average variation. Suppose we have the values 10, 11 and 24. The mean is then $45/3 = 15$. We have two values that are smaller than the mean and one value that is larger than the mean, so two negative deviations and one positive deviation. Squaring them makes them all positive. The squared deviations are 25, 16, and 81. The third value has a huge squared deviation (81) compared to the other two values. If we take the *average* squared deviation, we get $(25 + 16 + 81)/3 \approx 40.67$. So the average squared deviation is equal to 40.67. This value is called the *variance*. So the variance of a bunch of values is nothing but the SS divided by the number of values, n. The variance is *the average squared deviation from the mean*. The

symbol used for the variance is usually σ^2 (pronounced as 'sigma squared').[8]

$$\text{Var}(Y) = \frac{SS}{n} = \frac{\Sigma_i (Y_i - \overline{Y})^2}{n} \tag{1.6}$$

As an example, suppose you have the values 10, 11 and 12, then the average value is 11. Then the deviations are -1, 0 and 1. If you square them you get $(-1)^2 = 1$, $0^2 = 0$ and $1^2 = 1$, and if you add these three values, you get $SS = 1 + 0 + 1 = 2$. If you divide this by 3, you get the variance: $\frac{2}{3}$. Put differently, if the squared deviations are 1, 0 and 1, then the average squared deviation (i.e., the variance) is $\frac{1+0+1}{3} = \frac{2}{3}$.

As another example, suppose you have the values 8, 10, 10 and 12, then the average value is 10. Then the deviations from 10 are -2, 0, 0 and +2. Taking the squares, you get 4, 0, 0 and 4 and if you add them you get $SS = 8$. To get the variance, you divide this by 4: $8/4 = 2$. Put differently, if the squared deviations are 4, 0, 0 and 4, then the average squared deviation (i.e., the variance) is $\frac{4+0+0+4}{4} = 2$.

Often we also see another measure of variation: the *standard deviation*. The standard deviation is the square root of the variance and is therefore denoted as σ:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\text{Var}(Y)} = \sqrt{\frac{\Sigma_i (Y_i - \overline{Y})^2}{n}} \tag{1.7}$$

The standard deviation is often used to indicate how deviant a particular value is from the rest of the values. Take for instance an IQ score of 105. Is that a high IQ score or a low IQ score? Well, if someone tells you that the average person has an IQ score of 100, you know that a score of 105 is above average. However, still you do not know whether it is much higher than average, or just slightly higher than average. Suppose I tell you that the standard deviation of IQ scores is 15, then you know that a score of 105 is a third of a standard deviation above the mean. Therefore, in order to know how deviant a particular value is relative to a the rest of the values, one needs both a measure of central tendency and a measure of variation. In psychological testing, IQ testing for instance, one usually uses the mean and the standard deviation to express someone's score as the number of standard deviations above or below the average score. This process of counting the number of standard deviations is called *standardisation*. If we go back to the IQ score of 105, and if we want to standardise the score in terms of standard deviations from the mean, we saw that a score of 105 was a third of a standard deviation above the mean, so $+\frac{1}{3}$. As another example,

[8] Online you will often find the formula $\frac{\Sigma_i (Y_i - \overline{Y})^2}{n-1}$. The difference is that here we are talking about the definition of the variance of an observed variable Y, and that elsewhere one talks about trying to figure out what the variance might be of all values of Y when we only see a small portion of the values of Y. When we use all values of Y, we talk about the *population* variance, denoted by σ^2. When we only see a small part of the values of Y, we talk about a *sample* of Y-values. We will come back to the distinction between population variance and sample variance and why they differ in Chapter 2.

suppose the mean is 100 and we observe an IQ score of 80, we see that we are 20 points below the average of 100. This is equal to $20/15 = 4/3$ standard deviations below the average, so our standardised measure equals $-4/3$ (note the negative sign: it indicates we are below the mean). In general, a standardised score can be computed by subtracting the mean and dividing the result by the standard deviation. A standardised score for a particular value of Y, $Y = y$, is usually denoted by the z-score:

$$z = \frac{y - \overline{Y}}{\sigma} \tag{1.8}$$

1.16 Variance, standard deviation, and standardisation in R

The functions var() and sd() calculate the variance and standard deviation for a variable.

```
mtcars %>%
  summarise(var_mpg = var(mpg),
            std_mpg = sd(mpg))

## # A tibble: 1 x 2
##   var_mpg std_mpg
##     <dbl>   <dbl>
## 1    36.3    6.03
```

However, these functions use the formulas $\frac{\Sigma_i(Y_i-\overline{Y})^2}{n-1}$ and $\sqrt{\frac{\Sigma_i(Y_i-\overline{Y})^2}{n-1}}$, respectively. We will discuss this further in Chapter 2. If you want to use the formula $\frac{\Sigma_i(Y_i-\overline{Y})^2}{n}$, you need to write your own function that computes the sum of squares (SS) and divides by n:

```
var_n <- function(variable){
  SS <- (variable - mean(variable))**2 %>%
    sum()
  return(SS/length(variable)) # dividing by N
}

mtcars %>%
  summarise(var_mpg = var_n(mpg),
            std_mpg = sqrt(var_n(mpg))) # taking the square root

## # A tibble: 1 x 2
##   var_mpg std_mpg
##     <dbl>   <dbl>
## 1    35.2    5.93
```

Note that you get different results. For large data sets (large n), the differences will be negligible.

Standardised measures can be obtained using the `scale()` function:

```
mtcars %>%
  mutate(z_mpg = scale(mpg)) %>%
  select(mpg, z_mpg)

## # A tibble: 32 x 2
##      mpg z_mpg[,1]
##    <dbl>     <dbl>
##  1  21       0.151
##  2  21       0.151
##  3  22.8     0.450
##  4  21.4     0.217
##  5  18.7    -0.231
##  6  18.1    -0.330
##  7  14.3    -0.961
##  8  24.4     0.715
##  9  22.8     0.450
## 10  19.2    -0.148
## # ... with 22 more rows
```

1.17 Density plots

Earlier in this chapter we saw that when we have a number of values for a numeric variable, frequency tables and frequency plots fully describe all values of the variable that are observed. A histogram is a helpful tool to visualise the distribution of a variable when there are so many different values that a frequency table would be too long and a frequency plot would become too cluttered.

A histogram can then be used to give a quick graphical overview of the distribution. The bin width is usually chosen rather arbitrarily. Figure 1.6 shows a histogram of one million values of a numeric variable, say yearly **wage** for an administrative clerk. Figure 1.7 shows a histogram for the exact same data, but now using a much smaller bin size. You see that when you have a lot of values, a million in this case, you can choose a very small bin size, and in some cases this can result in a very clear shape of the distribution.

The shape of the distribution that we discern in Figure 1.7 can be represented by a *density plot*. Density plots are an elegant representation of how the frequency of certain values are distributed across a continuum. They are particularly suited for large amounts of non-discrete (continuous) values, typically more than 1000. Figure 1.8 shows a density plot of the one million wages. They more or less 'smooth' the histogram: drawing a smooth line connecting the dots of the histogram in Figure 1.7 while looking through your eyelashes. On the

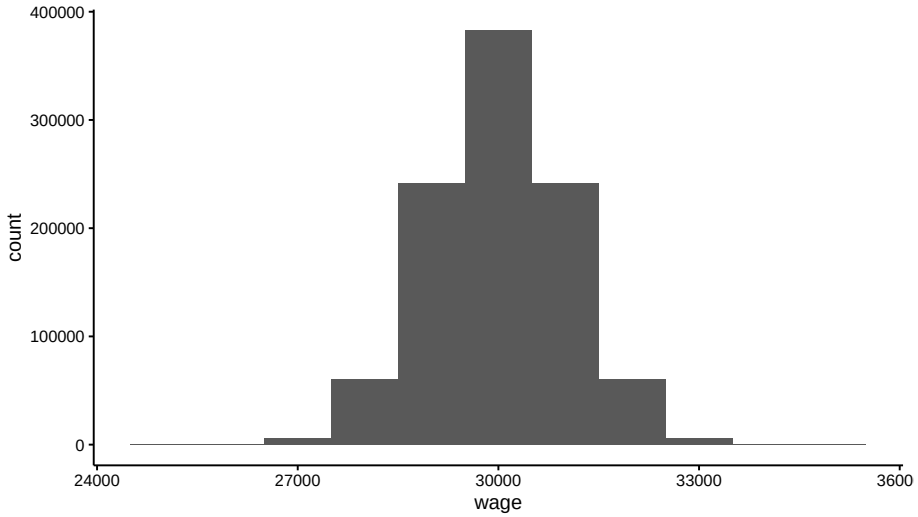


Figure 1.6: A histogram of wages with bin size 1000.

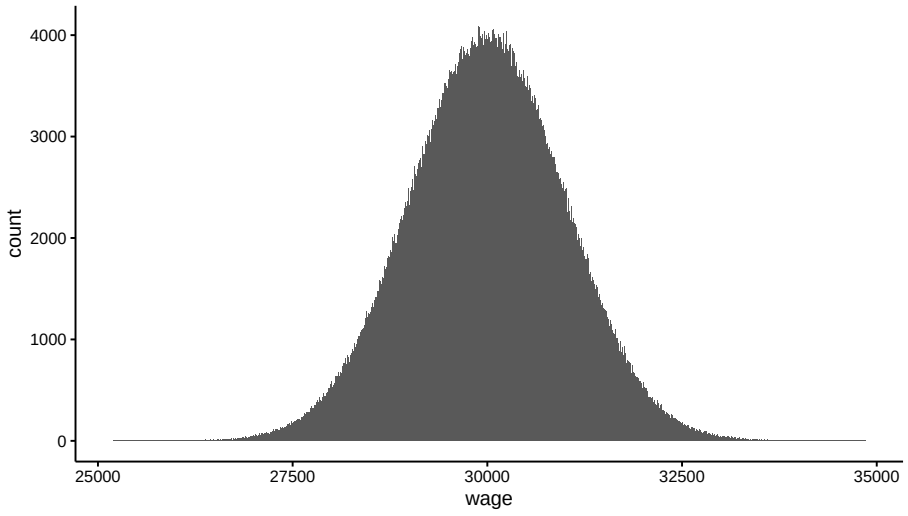


Figure 1.7: A histogram of wages with bin size 10.

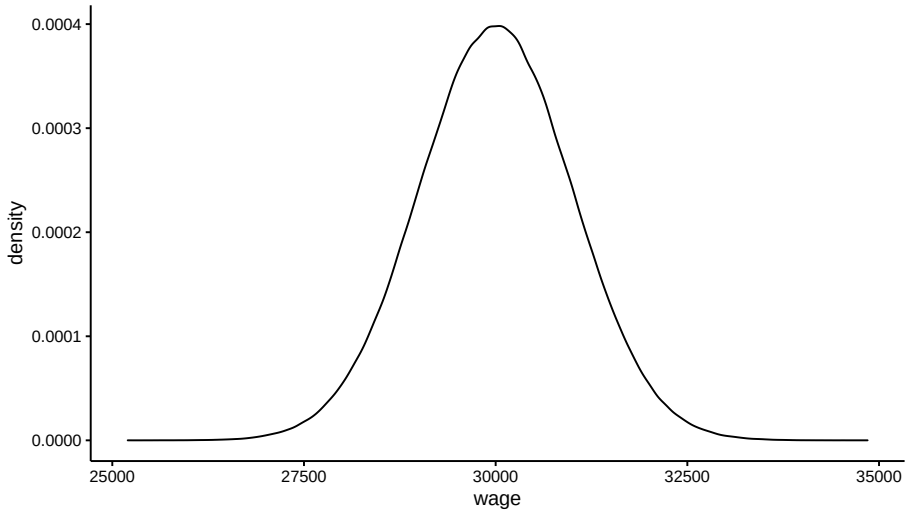


Figure 1.8: A density plot of the wage variable.

vertical axis, we no longer see 'count' or 'frequency', but 'density'. The quantity *density* is defined such that the area under the curve equals 1. Density plots are particularly suited for large data sets, where one is no longer interested in the particular counts, but more interested in relative frequencies: how often are certain values observed, relative to other values. From this density plot, it is very clear that, relatively speaking, there are more values around 30,000 than around 27,500 or 32,500.

1.18 Density plots in R

Density plots can be obtained using geom_density():

```
mtcars %>%
  ggplot(aes(x = mpg)) +
  geom_density()
```

1.19 The normal distribution

Sometimes distributions of observed variables bear close resemblance to *theoretical* distributions. For instance, Figure 1.8 bears close resemblance to the theoretical *normal* distribution with mean 30,000 and standard deviation 1000.

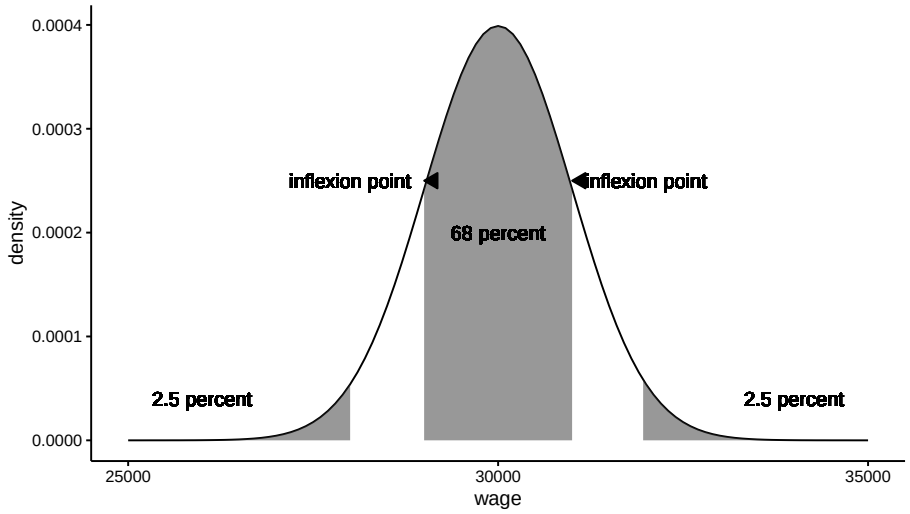


Figure 1.9: The theoretical normal distribution with mean 30,000 and standard deviation 1000.

This theoretical shape can be described with the mathematical function

$$f(x) = \frac{1}{\sqrt{2\pi 1000^2}} e^{-\frac{(x-30000)^2}{2\times 1000^2}} \tag{1.9}$$

which you are allowed to forget immediately. It is only to illustrate that distributions observed in the wild (empirical distributions) sometimes resemble mathematical functions (theoretical distributions).

The density function of that distribution is plotted in Figure 1.9. Because of its bell-shaped form, the normal distribution is sometimes informally called 'the bell curve'. The histogram in Figure 1.8 and the normal density function in Figure 1.9 look so similar, they are practically indistinguishable.

Mathematicians have discovered many interesting things about the normal distribution. If the distribution of a variable closely resembles the normal distribution, you can infer many things. One thing we know about the normal distribution is that the mean, mode and median are always the same. Another thing we know from theory is that the inflexion points[9] are one standard deviation away from the mean. Figure 1.9 shows the two inflexion points. From theory we also know that if a variable has a normal distribution, 68% of the observed values lies between these two inflexion points. We also know that 5% of the observed values lie more than 1.96 standard deviations away from the mean (2.5% on both sides, see Figure 1.9). Theorists have constructed tables that make it easy to see what proportion of values lies more than $1, 1.1, 1.2 \ldots, 3.8, 3.9, \ldots$

[9]The inflexion point is where concave turns into convex, and vice versa. Mathematically, the inflexion point can be found by equating the second derivative of a function to 0.

standard deviations away from the mean. These tables are easy to find online or in books, and these are fully integrated into statistical software like SPSS and R. Because all these percentages are known for the number of standard deviations, it is easier to talk about the *standard normal distribution*.

In such tables online or in books, you find information only about this standard normal distribution. The standard normal distribution is a normal distribution where all values have been *standardised* (see Section 1.15.3). When values have been standardised, they automatically have a mean of 0 and a standard deviation of 1. As we saw in Section 1.15.3, such standardised values are obtained if you subtract the mean score from each value, and divide the result by the standard deviation. A standardised value is usually denoted as a z-score. Thus in formula form, a value $Y = y$ is standardised by using the following equation:

$$z = \frac{y - \overline{Y}}{\sigma} \tag{1.10}$$

Table 1.12: Standardising scores.

Y	mean	Y_minus_mean	Z
7.2	10.4	-3.2	-0.7
8.8	10.4	-1.5	-0.3
17.8	10.4	7.4	1.6
10.4	10.4	-0.0	-0.0
10.6	10.4	0.3	0.1
18.6	10.4	8.2	1.7
12.3	10.4	1.9	0.4
3.7	10.4	-6.7	-1.4
6.6	10.4	-3.8	-0.8
7.8	10.4	-2.6	-0.5

Table 1.12 shows an example set of values for Y that are standardised. The mean of the Y-values turns out to be 10.38, and the standard deviation 4.77. By subtracting the mean, we ensure that the average z-score becomes 0, and by subsequently dividing by the standard deviation, we make sure that the standard deviation of the z-scores becomes 1.

This standardisation makes it much easier to look up certain facts about the normal distribution. For instance, if we go back to the normally distributed wage values, we see that the average is 30,000, and the standard deviation is 1,000. Thus, if we take all wages, subtract 30,000 and divide by 1,000, we get standardised wages with mean 0 and standard deviation 1. The result is shown in Figure 1.10. We know that the inflexion points lie at one standard deviation below and above the mean. The mean is 30,000, and the standard deviation equals 1,000, so the inflexion points are at $30000 - 1000 = 29000$ and $30000 + 1000 = 31000$. Thus we know that 68% of the wages are between 29,000 and 31,000.

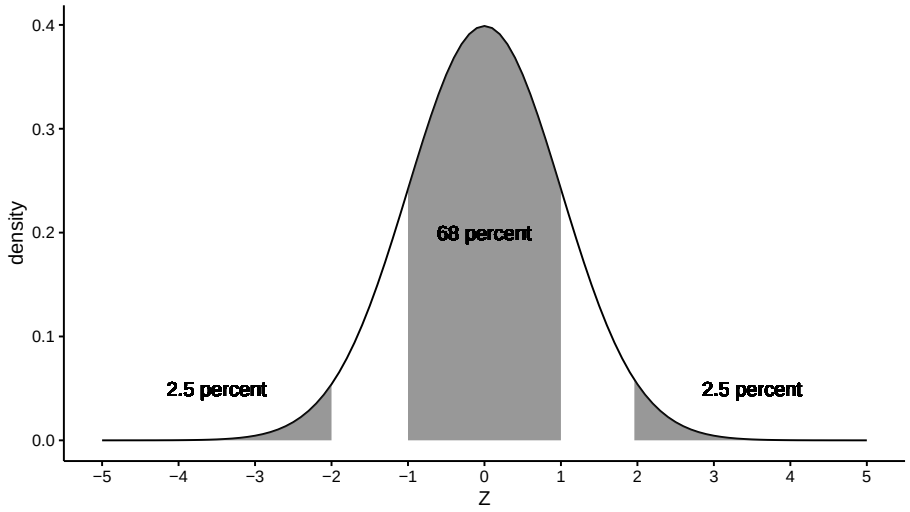


Figure 1.10: The standard normal distribution.

How do we know that 68% of the observations lie between the two inflexion points? Similar to proportions and cumulative proportions, we can plot the cumulative normal distribution. Figure 1.11 shows the cumulative proportions curve for the normal distribution. Note that we no longer see dots because the variable Z is continuous.

We know that the two inflexion points lie one standard deviation below and above the mean. Thus, if we look at a z-value of 1, we see that the cumulative probability equals about 0.84. This means that 84 % of the z-values are lower than 1. If we look at a z-value of -1, we see that the cumulative probability equals about 0.16. This means that 16 % of the z-values are lower than -1. Therefore, if we want to know what percentage of the z-values lie between -1 and 1, we can calculate this by subtracting 0.16 from 0.84, which equals 0.68, which corresponds to 68%.

All quantiles for the standard normal distribution can be looked up online[10] or in Appendix A, but also using R. Table 1.13 gives a short list of quantiles. From this table, you see that 1% of the z-values is lower than -2.33, and that 25% of the z-values is lower than -0.67. We also see that half of all the z-values is lower than 0.00 and that 10% of the z-values is larger than 1.28, and that the 1% largest values are higher than 2.33.

Although tables are readily found online, it's helpful to memorise the so-called *68 – 95 – 99.7 rule*. It says that 68% of normally distributed values are at most 1 standard deviation away from the mean, 95% of the values are at most 2 standard deviations away (more precisely, 1.96), and 99.7% of the values

[10]See for example www.normaltable.com or www.mathsisfun.com/data/standard-normal-distribution-table.html

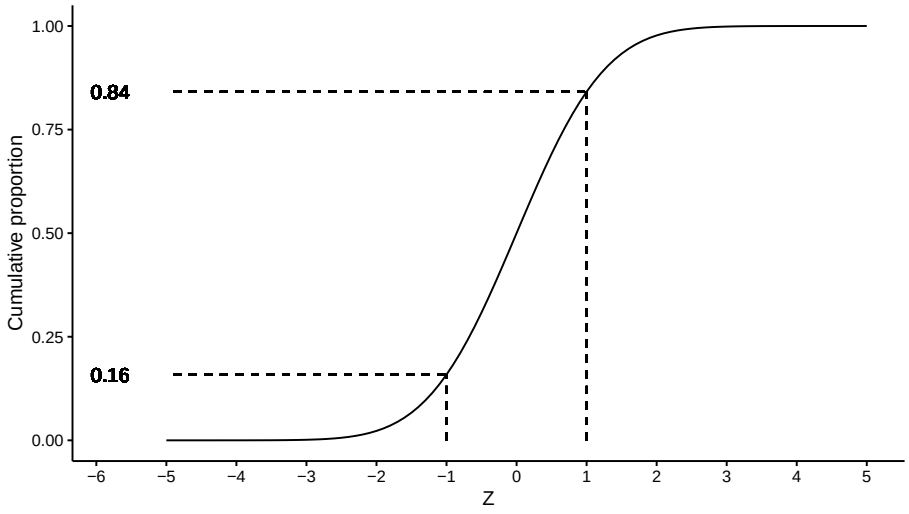


Figure 1.11: The cumulative standard normal distribution.

are at most 3 standard deviations away. In other words, 68% of standardised values are between -1 and +1, 95% of standardised values are between -2 and +2 (-1.96 and +1.96), 99.7% of standardised values are between -3 and +3.

Table 1.13: Some quantiles for the standard normal distribution.

Z	cum_proportion
-2.33	0.01
-1.28	0.10
-0.67	0.25
0.00	0.50
0.67	0.75
1.28	0.90
2.33	0.99

Thus, if we return to our wages with mean 30,000 and standard deviation 1,000, we know from Table 1.13 that 99% of the wages are below 30000 + 2.33 times the standard deviation = $30000 + 2.33 \times 1000$=32330.

Returning back to the IQ example of Section 1.15.3. Suppose we have IQ scores that are normally distributed with a mean of 100 and a standard deviation of 15. What IQ score would be the 90th percentile? From Table 1.13 we see that the 90th percentile is a z-value of 1.28. Thus, the 90th percentile for our IQ scores lies 1.28 standard deviations above the mean (above because the z-value is positive). The mean is 100 so we have to look at 1.28 standard deviations above that. The standard deviation equals 15, so we have to look at an IQ score of $100 + 1.28 \times 15$, which equals 119.2. This tells us that 90% of the IQ scores

are equal to or lower than 119.2.

As a last example, suppose we have a personality test that measures extraversion. If we know that test scores are normally distributed with a mean of 18 and a standard deviation of 2, what would be the 0.10 quantile? From Table 1.13 we see that the 0.10 quantile is a z-value of -1.28. This tells us that the 0.10 quantile for the personality scores lies at 1.28 standard deviations below the mean. The mean is 18, so the 0.10 quantile for the personality scores lies at 1.28 standard deviations below 18. The standard deviation is 2, so this amounts to $18 - 1.28 \times 2 = 15.44$. This tells us that 10% of the scores on this test are 15.44 or lower.

Such handy tables are also available for other theoretical distributions. Theoretical distributions are at the core of many data analysis techniques, including linear models. In this book, apart from the normal distribution, we will also encounter other theoretical distributions: the t-distribution (Chapter 2), the F-distribution (Chapter **??**), the chi-square distribution (Chapters 2, **??**, **??**, **??** and **??**) and the Poisson distribution (**??**).

1.20 Obtaining quantiles of the normal distribution using R

Quantiles of a normal distribution with a certain mean and standard deviation (sd) can be obtained using the qnorm() function:

```
qnorm(c(0.05, 0.50, 0.95), mean = 100, sd = 15)
```

```
## [1]  75.3272 100.0000 124.6728
```

This means that if you have a normal distribution with mean 100 and standard deviation 15, 5% of the values are 75.3272 or less, 50% of the values are 100 or less, and 95% of the values are 124.6728 or less.

If you want to know the cumulative proportion for a certain value of a variable that is normally distributed, you can use pnorm():

```
pnorm(-1, mean = 0, sd = 1)
```

```
## [1] 0.1586553
```

So 15.86% of the values from a standard normal distribution (mean 0, standard deviation 1), are -1 or less.

1.21 Visualising numeric variables: the box plot

We started this chapter with variables that can be stored in a data matrix. With a variable with a large number of values on a large number of units of analysis, it is hard to get a an intuitive feel for the data. Making a frequency table

is one way of summarising a variable, computing measures of central tendency and variation is another way. Visualisation is probably the best way of getting a quick and dirty feel for the information contained in a large data matrix. Earlier in this chapter we came across frequency plots, histograms, and density plots to visualise the distribution of a single variable. A fourth plot for a single variable that we discuss in this book is the *box plot*.

A box plot gives a quick overview of the distribution of a numeric variable in terms of its quartiles. Figure 1.12 gives an example of a box plot of (part of) the wage data. The white box represents the interquartile range. The top of the white box equals the third quartile, and the bottom of the white box equals the first quartile. Therefore, we know that half of the workers have a wage between 29,400 and 30,800 The horizontal black line within the white box represents the second quartile (the median), so half of the workers earn less than 30,100.

A box plot also shows whiskers: two vertical lines sprouting from the white box. There are several ways to draw these two whiskers. One way is to draw the top whisker to the largest value (the maximum) and the bottom whisker to the smallest value (the minimum). Another way, used in Figure 1.12, is to have the upper whisker extend from the third quartile to the observed value equal to at most 1.5 times the interquartile range away from the median, and the lower whisker extend from the first quartile to the value at most 1.5 times the interquartile range below the median (the interquartile range is of course the height of the white box). The dots are outlying values, or simply called *outliers*: values that are even further away from the median. This is displayed in Figure 1.12. There you see first and third quartiles of 29,400 and 30,800, respectively, so an interquartile range (IQR) of $30800 - 29400 = 1400$. Multiplying this IQR by 1.5 we get $1.5 \times 1400 = 2100$. The whiskers therefore extend to $29400 - 2100 = 27300$ and $30800 + 2100 = 32900$.

Thus, the box plot is a quick way of visualising in what range the middle half of the values are (the range in the white box), where most of the values are (the range of the white box plus the whiskers), and where the extreme values are (the outliers, individually plotted as dots). Note that the white box always contains 50% of the values. The whiskers are only extensions of the box by a factor of 1.5. In many cases you see that they contain most of the values, but sometimes they miss a lot of values. You will see that when you notice a lot of outliers.

1.22 Box plots in R

A box plot can be made using geom_boxplot():

```
mtcars %>%
  ggplot(aes(x = "", y = mpg)) +
  geom_boxplot() +
  xlab("")
```

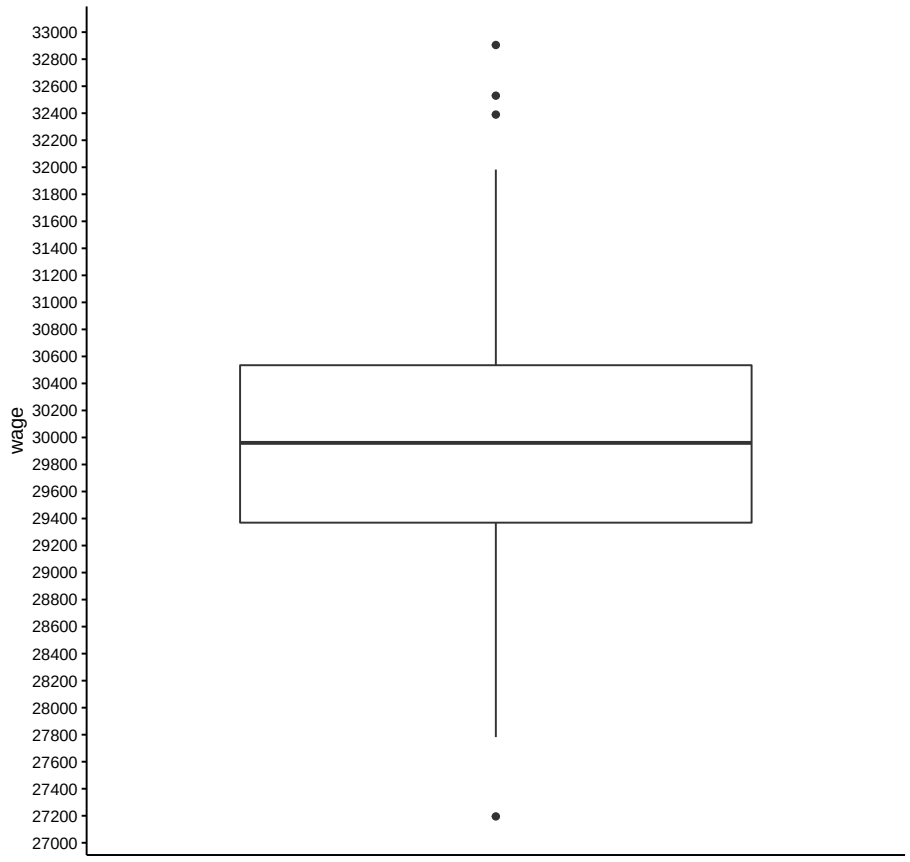


Figure 1.12: A boxplot of the wages earned by a sample of 150 administrative clerks

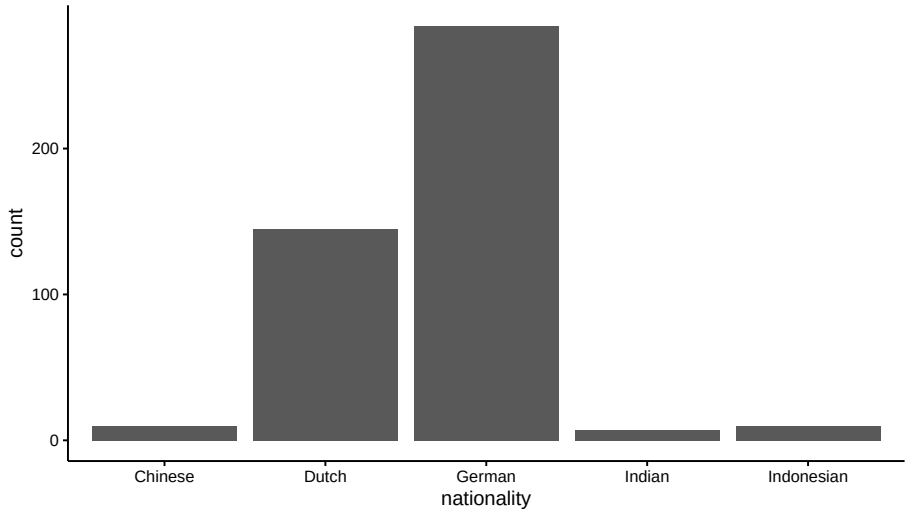


Figure 1.13: A bar chart of the observed nationalities in a lecture hall.

1.23 Visualising categorical variables

The histogram, the density plot and the box plot can be used for numeric variables, but also for ordinal variables that you'd like to treat numerically. For categorical variables and ordinal variables that can't be treated numerically, we need other types of plots.

For example, suppose we are in a lecture hall with 456 students and we count the number of Dutch, German, Belgian, Indian, Chinese and Indonesian students. We could summarise the results in a frequency table (see Table 1.14), but a *bar chart* shows the distribution in a more dramatic way, see Figure 1.13.

Table 1.14: A frequency table of nationalities.

nationality	n
Chinese	10
Dutch	145
German	284
Indian	7
Indonesian	10

Sometimes, counts of values of a categorical variable are displayed as a *pie chart*, see Figure 1.14. Pie charts are however best avoided. First, because compared to bar charts, they show no information about the actual counts; you only observe relative sizes of the counts. Second, it is very hard to see from a pie chart what the exact proportions are. For example, from the bar chart in Figure 1.13 it is easily seen that the ratio German students to Dutch students

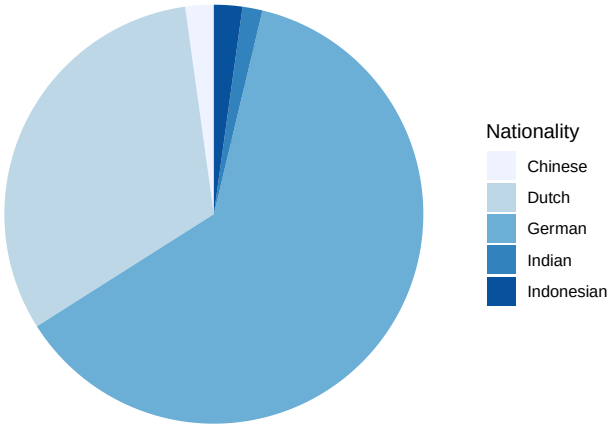


Figure 1.14: A pie chart of nationalities.

is about 2 to 1. Research shows that this ratio cannot be read with the same precision from the pie chart in Figure 1.14. In sum, pie charts are best replaced by bar charts.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Ordinal variables are often visualised using bar charts. Figure 1.15 shows the variation of the answers to a Likert questionnaire item, where Nairobi inhabitants are asked "To what degree do you agree with the statement that the climate in Iceland is agreeable?". With ordinal variables, make sure that the labels are in the natural order.

1.24 Visualising categorical and ordinal variables in R

If a categorical variable is stored as numeric, turn it into a factor first. Then R will treat it as categorical. A bar plot with the frequencies on the y-axis can be made with geom_bar():

```
mtcars %>%
  mutate(cyl = factor(cyl, ordered = TRUE)) %>%
  ggplot(aes(x = cyl)) +
  geom_bar()
```

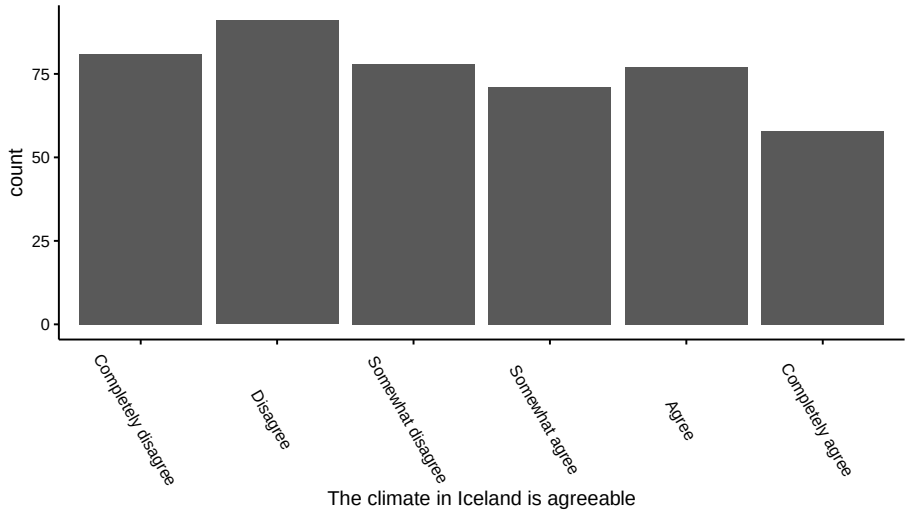
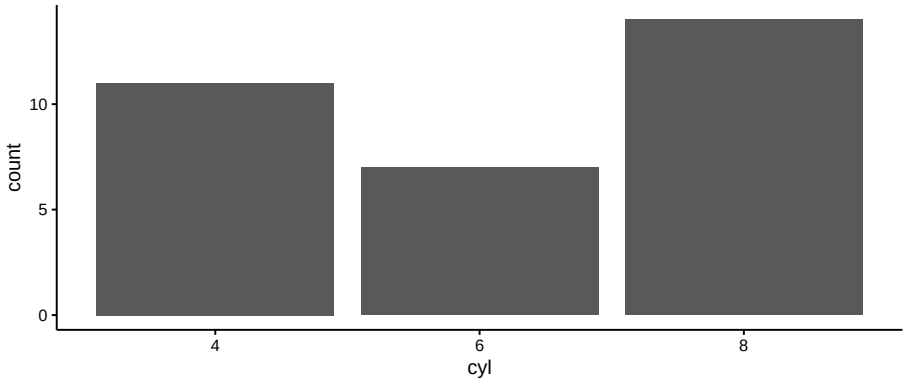


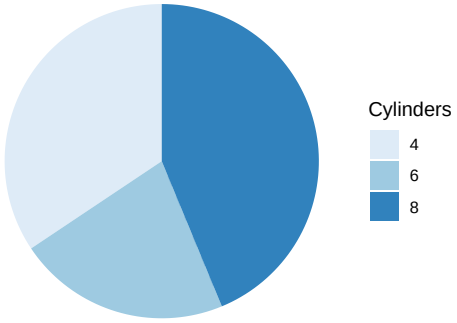
Figure 1.15: Opinions on the climate in Iceland.



If you really want a pie chart, then do:

```
mtcars %>%
  count(cyl) %>%
  mutate(proportion = n/sum(n)) %>%
  ggplot(aes(x = "",
             y = proportion,
             fill = factor(cyl))) +
  geom_col(width = 1) +
  coord_polar(theta = "y") +
  xlab("") +
  ylab("") +
  theme_void() +
```

```
  scale_fill_brewer(palette = "Blues") +
  labs(fill = "Cylinders")
```



1.25 Visualising co-varying variables

1.25.1 Categorical by categorical: cross-table

Variables are properties that vary: from person to person, or from location to location, or from time to time, or from object to object. Sometimes when you have two variables, you see that they co-vary: when one variable changes, the other variable changes too. For example, suppose I have 20 pencils. These pencils may vary in colour: twelve of them are red, and eight of them are blue. Therefore, colour is a variable with values "red" and "blue". The twenty pencils also vary in length: four are unused and therefore still long, and sixteen of them have been used many times so that they are short. Therefore, length is also a variable, with values "long" and "short". Note that these variables have been measured using the same pencils. In theory I could have long blue pencils, long red pencils, short blue pencils and short red pencils. Let's look at the pencils that I have: for each combination of length and colour, I count the number of pencils. The result I put in Table 1.15.

Table 1.15: Cross-tabulation of colour and length for twenty pencils.

	blue	red
long	4	0
short	8	8

Such a table is called a *cross-table*. For every combination of two variables, I see the number of objects (units of analysis) that have that combination. From the table we see that there is not a single pencil that is both red and long (count is 0). At the same time you see that all long pencils are blue. A cross-table is therefore a nice way to show how two variables co-vary. From this particular

table for instance, you can easily see that once you know that a pencil is long, you automatically know it is blue.

Cross-tables are a nice visualisation of how two categorical variables co-vary. But what if one of the two variables is not a categorical variable?

1.25.2 Categorical by numerical: box plot

Suppose instead of determining length by values "short" and "long", we could measure the exact length of the pencils in centimetres. The results are displayed in Table 1.16. We see that the table is much larger than Table 1.15. We also see quite a few cells with zeros. In most cases, for every particular combination of length and colour we only see a count of 1 pencil. In general, you see that when one of the variables is numeric, the cross-table becomes very large and in addition it becomes sparse, that is, with many zeros. With such a large and sparse table, it is hard to get a quick impression of how two variables co-vary.

Table 1.16: Cross-tabulation of colour and length for twenty pencils.

	blue	red
2	0	1
2.7	1	0
3.3	1	0
3.4	0	1
3.5	0	1
3.6	1	0
4.1	1	1
4.4	1	1
4.5	1	1
4.7	0	1
5.2	1	0
5.7	1	0
5.8	0	1
9	4	0

The alternative for two variables where one is categorical and the other one is numeric, is to create a *box plot*. Figure 1.16 shows a box plot of the pencil data. A box plot gives a quick overview of the distribution of the pencils: one distribution of the blue pencils, and one distribution of the red pencils. Let's have a look at the distribution of the blue pencils on the left side of the plot. The white box represents the interquartile range (IQR), so that we know that half of the blue pencils have a height between 4 and 9. The horizontal black line within the white box represents the median (the middle value), so half of the blue pencils are smaller than 4.85. The vertical lines are called whiskers. These typically indicate where the data points are that lie at most 1.5 times the IQR away from the median. For the blue pencils, we see no whisker on top of the white box. That means that there are no data points that lie more than

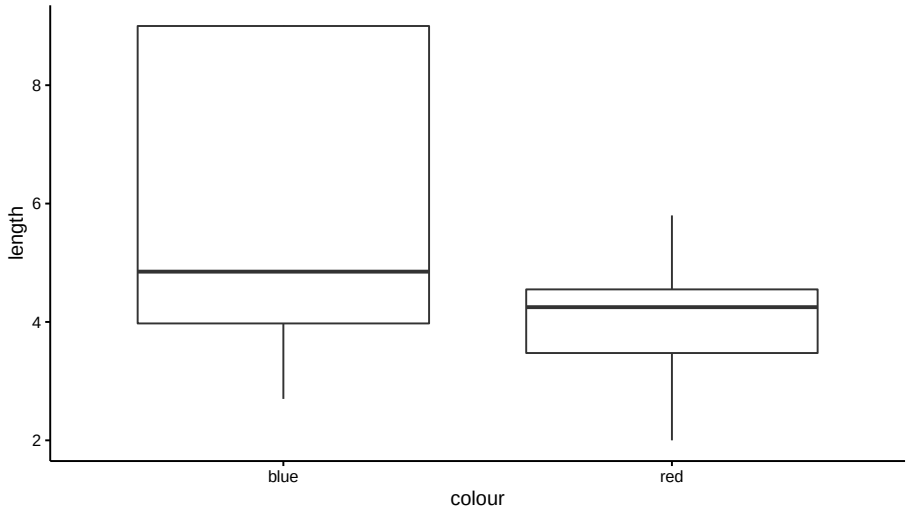


Figure 1.16: A boxplot of the pencil data.

1.5 times the IQR above the median of 4.85 (here the IQR equals 5.03). We see a whisker on the bottom of the white box, to the lowest observed value of 2.7. This value is less than 1.5 times $5.03 = 7.545$ away from the median of 4.85 so it is included in the whisker. It is the lowest observed value for the blue pencils so the whisker ends there.

From a box plot like this it is easy to spot differences in the distribution of a quantitative measure for different levels of a qualitative measure. From Figure 1.16 we easily spot that the red pencils (varying between 2 and 6 cm) tend to be shorter than the blue pencils (varying between 3 and 9 cm). Thus, in these pencils, length and colour tend to co-vary: red pencils are often short and blue pencils are often long.

1.25.3 Numeric by numeric: scatter plot

Suppose we also measure the weight of my pencils in grams. Table 1.17 shows the cross-tabulation of length and weight. This is a very sparse table (i.e., with lots of zeros), which makes it very hard to see any systematic co-variation in weight and length. Figure 1.17 shows a box plot of weight and length. Also this plot seems a bit strange, because for every observed weight value under 4 grams, there is only one observation, so that only the median can be plotted.

Therefore, in cases where we have two numeric variables, we generally use a *scatter plot*. Figure 1.18 shows a scatter plot of weight by length. Now, the relationship between height and length is easily understood: it appears there is a *linear* relationship between weight. For every increase in weight, there is also an increase in length. The relationship is called linear because we could summarise the relationship by drawing a straight line through the dots. This

Table 1.17: Cross-tabulation of length (rows) and weight (columns) for twenty pencils.

	3.3	3.4	3.5	3.6	3.7	4
2	1	0	0	0	0	0
2.7	0	1	0	0	0	0
3.3	0	1	0	0	0	0
3.4	0	1	0	0	0	0
3.5	0	0	1	0	0	0
3.6	0	0	1	0	0	0
4.1	0	0	2	0	0	0
4.4	0	0	2	0	0	0
4.5	0	0	2	0	0	0
4.7	0	0	0	1	0	0
5.2	0	0	0	1	0	0
5.7	0	0	0	0	1	0
5.8	0	0	0	0	1	0
9	0	0	0	0	0	4

line is shown in Figure 1.19.

You see that by visualising two variables, important patterns may emerge that you can easily overlook when only looking at the values. Cross-tables, box plots and scatter plots are powerful tools to find regularities but also oddities in your data that you'd otherwise miss. Some such patterns can be summarised by straight lines, as we see in Figure 1.19. The remainder of this book focuses on how we can use straight lines to summarise data, but also how to make predictions for data that we have not seen yet.

1.26 Visualising two variables using R

A scatter plot for two numeric variables can be made using geom_point():

```
mtcars %>%
  ggplot(aes(x = wt, y = mpg)) +
  geom_point()
```

A box plot for one categorical and one numeric variable can be made using geom_boxplot():

```
mtcars %>%
  mutate(cyl = factor(cyl)) %>%
  ggplot(aes(x = cyl, y = mpg)) +
  geom_boxplot()
```

A cross table for two categorical variables can be made using table():

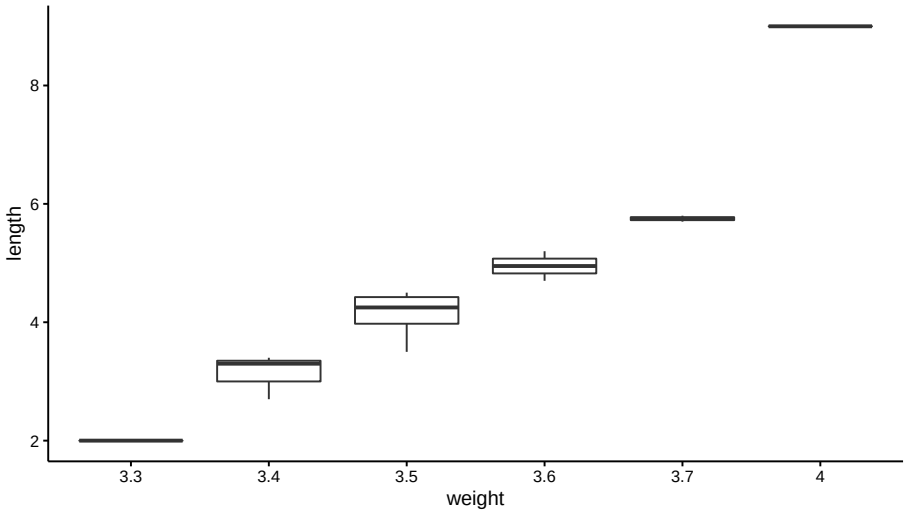


Figure 1.17: A boxplot of the pencil data.

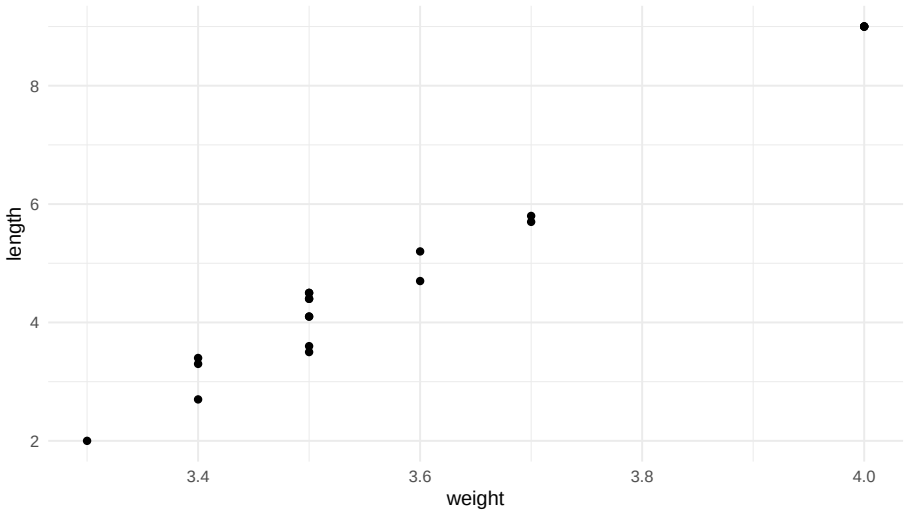


Figure 1.18: A scatterplot of length and weight.

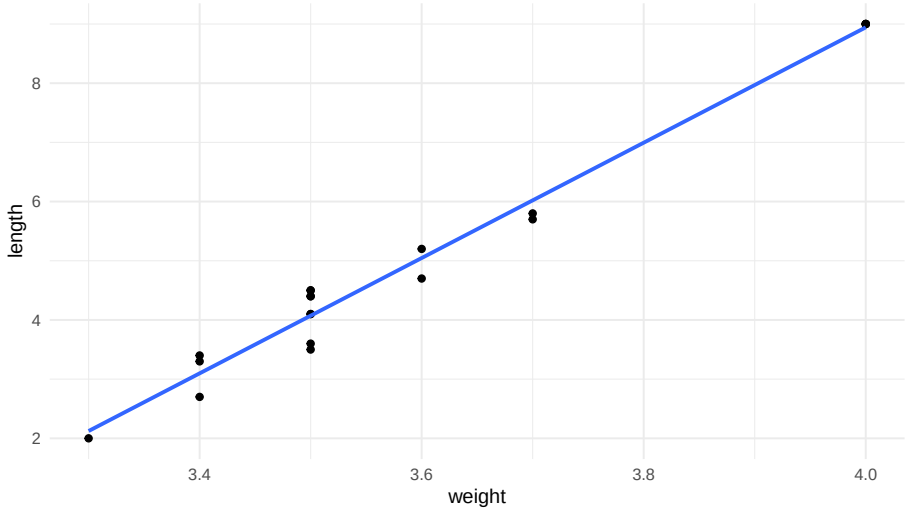


Figure 1.19: A scatterplot of length and weight, with a straight line that summarises the relationship.

```
table(mtcars$cyl, mtcars$gear)
```

```
##
##      3  4  5
##   4  1  8  2
##   6  2  4  1
##   8 12  0  2
```

Note that the number of cylinders (first-named variable) is in the rows (here 4, 6 and 8 cylinders), and the number of gears (second-named variable) is in the columns (3, 4, and 5 gears).

1.27 Overview of the book

Chapter 2 will introduce the problem of *inference*: if you only have a small selection of data points, what can they tell us about the rest of the data? We will use the example of a mean computed using a small number of numerical data points and try to figure out what the mean is likely to be if we would have all the data points. Chapter 3 discusses the same problem but then for a proportion.

Chapter 4 will show how we can use a straight line to summarise the relationship between two numeric variables (simple regression). Such a straight line is a simple form of a *linear model*. We also describe how we can use straight lines (linear models) to summarise relationships between more than two numeric

variables (multiple regression). In Chapters **??** and **??** we will discuss how you can draw conclusions about linear models for data that you have not seen. For example, in the previous section we described the relationship between weight and height of twenty pencils. The question that you may have is whether this linear relationship also holds for *all* pencils of the same make, that is, whether the same linear model holds for both the observed twenty pencils and the total collection of pencils.

In Chapter **??** we will show how we can use straight lines to summarise relationships with independent variables that we want to treat as categorical.

Chapter **??** focuses on moderation: how one variable can affect the effect that a second variable has on the outcome variable.

Chapter **??** shows how you can make elaborate statements about differences between groups of observations, in case one of the variables is a categorical variable.

Chapter **??** discusses when it is appropriate to use linear models to summarise your data, and when it is not. It introduces methods that enable you to decide whether to trust a linear model or not. Chapter **??** then discusses alternative methods that you can use when linear models are not appropriate.

Chapters **??** and **??** show how to deal with variables that are measured more than once in the same unit of analysis (the same participant, the same pencil, the same school, etc.). For example, you may measure the weight of a pencil before and after you have made a drawing with it. Models that we use for such data are called *linear mixed models*. Similar to linear models, linear mixed models are not always appropriate. Therefore, Chapter **??** discusses alternative methods to study variables that are repeatedly measured in the same research unit.

Chapters **??** and **??** discuss *generalised linear models*. These are models where the dependent variable is not numeric and continuous. Chapter **??** discusses a method that is appropriate when the dependent variable has only two values, say "yes" and "no", or "pass" and "fail". Chapter **??** discusses a method that can be used when the dependent variable is a count variable and therefore discrete, for example the number of children in a classroom, or the number of harvested zucchini from one plant.

Chapter **??** discusses relatively new statistical methodology that is needed when you have a lot of variables. In such cases, traditional inferential data analysis as discussed in the previous chapters often fails.

Chapter 2

Inference about a mean

2.1 The problem of inference

The human body is heavily controlled by hormones. One of the hormones involved in a healthy reproductive system is luteinising hormone (LH). This hormone is present in both females and males, but with different roles. In females, a sudden rise in LH levels triggers ovulation (the release of an egg from an ovary). We have a data set on luteinising hormone (LH) levels in one anonymous female. The data are given in Figure 2.1. In this data set, we have 48 measures, taken at 10-minute intervals. We see that LH levels show quite some variation over time. Suppose we want to know the mean level of luteinising hormone level in this woman, how could we do that?

The easiest way is to compute the mean of all the values that we see in this graph. If we do that here, we get the value 2.4. That value is displayed as the red line in Figure 2.1. However, is that really the mean of the hormone levels during that time period? The problem is that we only have 48 measures; we do not have information about the hormone levels *in between* measurements. We see some very large differences between two consecutive measures, which makes the level of hormone look quite unstable. We lack information about hormone levels in between measurements because we do not have data on that. We only have information about hormone levels at the times where we have observed data. For the other times, we have unobserved or missing data.

Suppose that instead of the mean of the *observed* hormone levels, we want to know the mean of *all* hormone levels during this time period: not only those that are measured at 10-minute intervals, but also those that are not measured (unobserved/missing).

You could imagine that if we would measure LH not every 10 minutes, but every 5 minutes, we would have more data, and the mean of those measurements would probably be somewhat different than 2.4. Similarly, if we would take measurements every minute, we again would obtain a different mean. Suppose we want to know what the true mean is: the mean that we would get if we

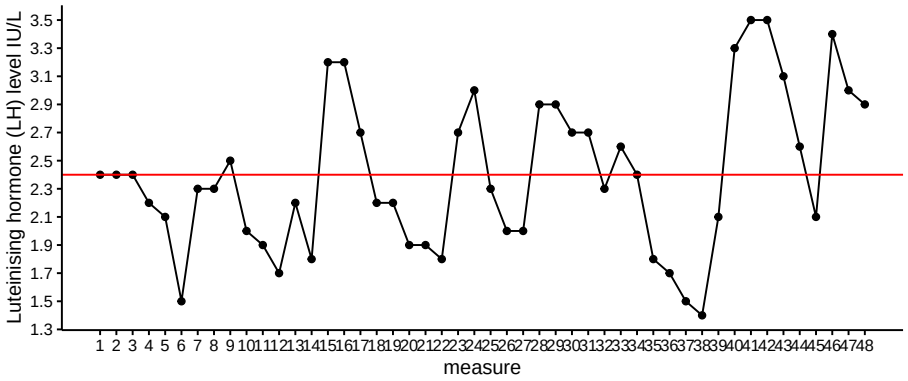


Figure 2.1: Luteinising hormone levels measured in one female, 48 measures taken at 10-minute intervals.

would measure LH continuously, that is, an infinite number of measurements. Unfortunately we only have these 48 measures to go on. We would like to infer from these 48 measures, what the mean is of LH level *had we measured continuously*.

This is the problem of *inference*: how to infer something about complete data, when you only see a small subset of the data. The problem of *statistical inference* is when you want to say something about an imagined complete data set, the *population*, when you only observe a relatively small portion of the data, the *sample*.

In order to show you how to do that, we do a thought experiment. Imagine a huge data set on African elephants where we measured the height of each elephant currently living (today around 415,000 individuals). Let's imagine that for this huge data set, the mean and the variance are computed: a mean of 3.25 m and a variance of 0.14 (recall, from Chapter 1, that the variance is a measure of spread, based on the sums of squared differences between values and the mean). We call this data set of all African elephants currently living the *population* of African elephants.

Now that we know that the actual mean equals 3.25 and the actual variance equals 0.14, what happens if we only observe 10 of these 415,000 elephants? In our thought experiment we randomly pick 10 elephants. Random means that every living elephant has an equal chance of being picked. This random *sample* of 10 elephants is then used to compute a mean and a variance. Imagine that we do this exercise a lot of times: every time we pick a new random sample of 10 elephants, and you can imagine that each time we get slightly different values for our mean, but also for our variance. This is illustrated in Table 2.1, where we show the data from 5 different samples (in different columns), together with 5 different means and 5 different variances.

What we see from this table is that the 5 *sample means* vary around the population mean of 3.25, and that the 5 variances vary around the population

variance of 0.14. We see that therefore the mean based on only 10 elephants gives a rough approximation of the mean of *all* elephants: the sample mean gives a rough approximation of the population mean. Sometimes it is too low, sometimes it is too high. The same is true for the variance: the variance based on only 10 elephants is a rough approximation, or *estimate*, of the variance of *all* elephants: sometimes it is too low, sometimes it is too high.

Table 2.1: Imaginary data on elephant height when 5 random samples (columns) of 10 elephants (rows) are drawn from the population data.

	1	2	3	4	5
1	3.77	2.52	3.26	3.61	3.16
2	3.61	3.41	3.09	3.33	2.74
3	3.12	2.91	3.14	3.22	3.91
4	2.95	3.20	2.85	3.40	3.60
5	2.53	3.45	2.69	3.20	3.19
6	3.12	3.11	3.45	2.31	2.94
7	3.31	3.22	2.98	3.65	4.39
8	2.59	3.76	2.81	2.20	3.24
9	2.91	3.44	3.63	3.12	3.21
10	3.36	2.84	4.15	2.73	2.75
mean	3.13	3.19	3.20	3.08	3.31
variance	0.14	0.12	0.18	0.23	0.24

2.2 Sampling distribution of mean and variance

How high and how low the sample mean can be, is seen in Figure 2.2. There you see a histogram of all sample means when you draw 10,000 different samples of each consisting of 10 elephants and for each sample compute the mean. This distribution is a *sampling distribution*. More specifically, it is the sampling distribution of the sample mean.

The red vertical line indicates the mean of the population data, that is, the mean of 3.25 (the population mean). The blue line indicates the mean of all these sample means together (the mean of the sample means). You see that these lines practically overlap.

What this sampling distribution tells you, is that if you randomly pick 10 elephants from a population, measure their heights, and compute the mean, this mean is *on average* a good estimate (approximation) of the mean height in the population. The mean height in the population is 3.25, and when you look at the sample means in Figure 2.2, they are generally very close to this value of 3.25. Another thing you may notice from Figure 2.2 is that the sampling distribution of the sample mean looks symmetrical and resembles a normal distribution.

Now let's look at the sampling distribution of the sample variance. Thus, every time we randomly pick 10 elephants, we not only compute the mean but

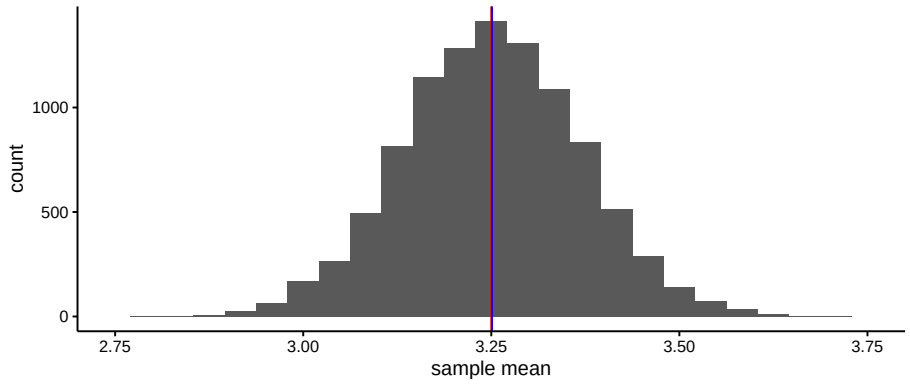


Figure 2.2: A histogram of 10,000 sample means when the sample size equals 10.

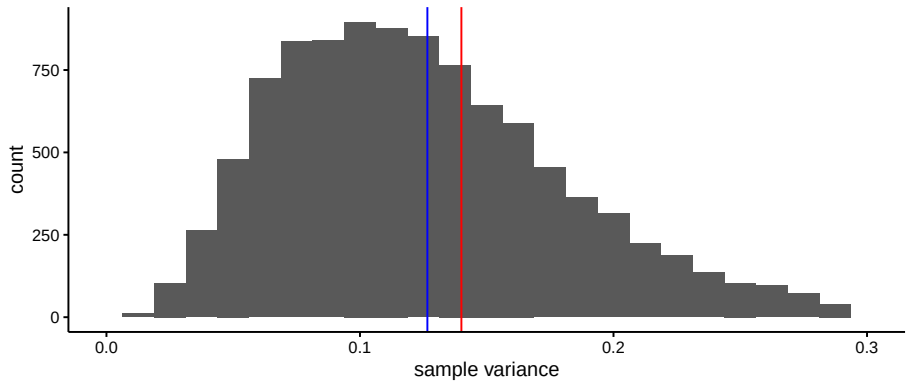


Figure 2.3: A histogram of 10,000 sample variances when the sample size equals 10. The red line indicates the population variance. The blue line indicates the mean of all variances observed in the 10,000 samples.

also the variance. Figure 2.3 shows the sampling distribution. The red line shows the variance of the height in the population, and the blue line shows the mean variance observed in the 10,000 samples. Clearly, the red and blue line do not overlap: the mean variance in the samples is slightly lower than the actual variance in the population. We say that the sample variance underestimates the population variance a bit. Sometimes we get a sample variance that is lower than the population value, sometimes we get a value that is higher than the population value, but on average we are on the low side.

Overview

- **population**: all values, both observed and unobserved
- **population mean**: the mean of all values (observed and unobserved values)
- **sample**: a limited number of observed values
- **sample size**: the number of observed values
- **sample mean**: the mean of the values in the sample
- **random sample**: values that you observe when you randomly pick a subset of the population
- **random**: each value in the population has an equal probability of being observed
- **sampling distribution of the sample mean**: the distribution of means that you get when you randomly pick new samples from a population and for each sample compute the mean
- **sampling distribution of the sample variance**: the distribution of variances that you get when you randomly pick new samples from a population and for each sample compute the variance

2.3 The effect of sample size

What we have seen so far is that when the population mean is 3.25 m and we observe only 10 elephants, we may get a value for the sample mean of somewhere around 3.25, but on average, we're safe to say that the sample mean is a good approximation for the population mean. In statistics, we call the sample mean an *unbiased estimator* of the population mean, as the expected value (the average value we get when we take a lot of samples) is equal to the population value.

Unfortunately the same could not be said for the variance: the sample variance is not an unbiased estimator for the population variance. We saw that on

average, the values for the variances are too low.

Another thing we saw was that the distribution of the sample means looked symmetrical and close to normal. If we look at the sampling distribution of the sample variance, this was less symmetrical, see Figure 2.3. It actually has the shape of a so-called χ^2-(pronounced 'chi-square') distribution, which will be discussed in Chapters **??**, **??**, **??** and **??**. Let's see what happens when we do not take samples with 10 elephants each time, but 100 elephants.

Stop and think: What will happen to the sampling distributions of the mean and the variance? For instance, in what way will Figure 2.2 change when we use 100 elephants instead of 10?

Figure 2.4 shows the sampling distribution of the sample mean. Again the distribution looks normal, again the blue and red lines overlap. The only difference with Figure 2.2 is the spread of the distribution: the values of the sample means are now much closer to the population value of 3.25 than with a sample size of 10. That means that if you use 100 elephants instead of 10 elephants to estimate the population mean, on average you get much closer to the true value!

Now stop for a moment and think: is it logical that the sample means are much closer to the population mean when you have 100 instead of 10 elephants?

Yes, of course it is, with 100 elephants you have much more information about elephant heights than with 10 elephants. And if you have more information, you can make a better approximation (estimation) of the population mean.

Figure 2.5 shows the sampling distribution of the sample variance. Compared to a sample size of 10, the shape of the distribution now looks more symmetrical and closer to normal. Second, similar to the distribution of the means, there is much less variation in values: all values are now closer to the true value of 0.14. And not only that: it also seems that the bias is less, in that the blue and the red lines are closer to each other.

Here we see three phenomena. The first is that if you have a statistic like a mean or a variance and you compute that statistic on the basis of randomly picked sample data, the distribution of that statistic (i.e., the sampling distribution) will generally look like a normal distribution if sample size is large enough.

It can actually be proven that the distribution of the mean will become a normal distribution if sample size becomes large enough. This phenomenon is known as the Central Limit Theorem. It is true for any population, no matter what distribution it has.[1] Thus, this means that height in elephants itself does not have to be normally distributed, but the sampling distribution of the sample mean will be normal for large sample sizes (e.g., 100 elephants).

The second phenomenon is that the sample mean is an unbiased estimator

[1]This is true except for the case that you have fewer than 3 data points and for a few special cases, that you don't need to know about in this book.

of the population mean, but that the variance of the sample data is not an unbiased estimator of the population variance. Let's denote the variance of the sample data as S^2. Remember from Chapter 1 that the formula for the variance is

$$S^2 = \text{Var}(Y) = \frac{\Sigma(y_i - \bar{y})^2}{n} \tag{2.1}$$

We saw that the bias was large for small sample size and small for larger sample size. So somehow we need to correct for sample size. It turns out that the correction is a multiplication with $\frac{n}{n-1}$:

$$s^2 = \frac{n}{n-1} S^2 \tag{2.2}$$

where s^2 is the corrected estimator of population variance, S^2 is the variance observed in the sample, and n is sample size. When we rewrite this formula and cancel out n, we get a more direct way to compute s^2:

$$s^2 = \frac{\Sigma(y_i - \bar{y})^2}{n-1} \tag{2.3}$$

Thus, if we are interested to know the variance or the standard deviation in the population, and we only have sample data, it is better to take the sums of squares and divide by $n-1$, and not by n.

$$\widehat{\sigma^2} = s^2 = \frac{\Sigma(y_i - \bar{y})^2}{n-1} \tag{2.4}$$

where $\widehat{\sigma^2}$ (pronounced 'sigma-squared hat') signifies the estimator of the population variance (the little hat stands for estimator or estimated value).

The third phenomenon is that if sample size increases, the variability of the sample statistic gets smaller and smaller: the values of the sample means and the sample variances get closer to their respective population values. We will delve deeper into this phenomenon in the next section.

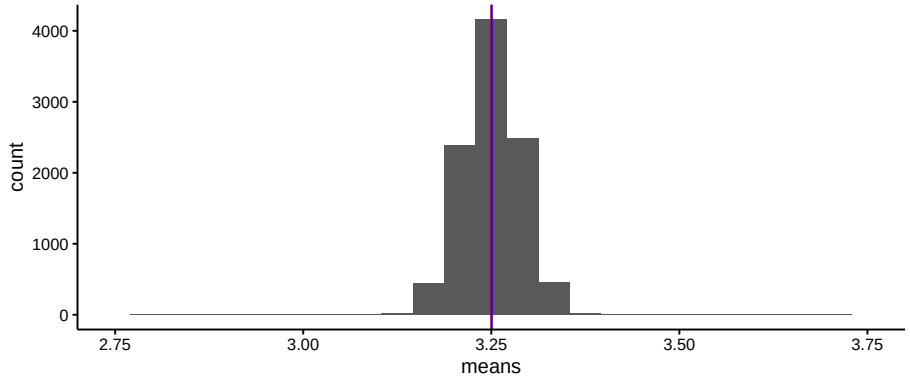


Figure 2.4: A histogram of 10,000 sample means when the sample size equals 100.

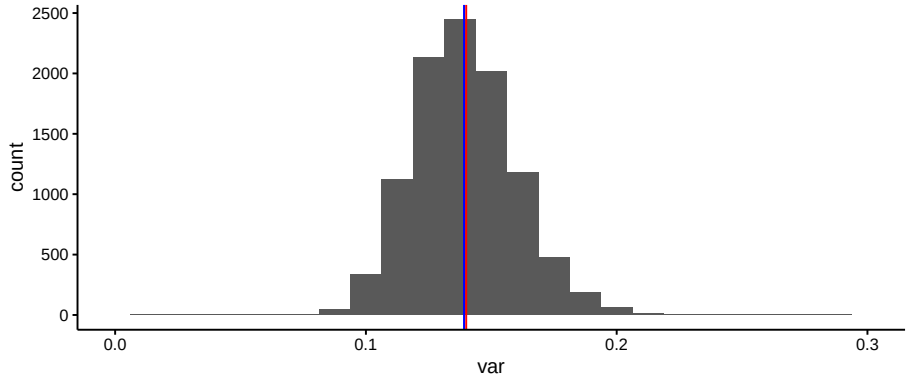


Figure 2.5: A histogram of 10,000 sample variances when the sample size equals 100.

Overview

- **Central Limit Theorem**: says that the sampling distribution of the sample mean will be normally distributed for infinitely large sample sizes.
- **estimator**: a quantity that you compute based on sample data, that you hope says something about a quantity in the population data. For instance, you can use the sample mean and hope that it is close to the population mean. You use the sample mean as an approximation of the population mean.
- **estimate**: the actual value that you get when computing an estimator. For instance, we can use the sample mean as the estimator of the population mean. The formula for the sample mean is $\frac{\Sigma y_i}{n}$ so this formula is our estimator. Based on a sample of 10 values, you might get a sample mean of 3.5. Then 3.5 is the estimate for the population mean.
- **unbiased estimator**: an estimator that has the population value as expected value (the mean that you get when averaging over many samples). For example, the sample mean is an unbiased estimator for the population mean because if you draw an infinite number of samples, the mean of the sample means will be equal to the population mean.
- **biased estimator**: an estimator that does not have the population value as expected value. For example, the variance calculated using a sample is a biased estimator for the population variance because if you draw an infinite number of samples, the mean of the variances will not be equal to the population variance.
- S^2: the variance of the values in the sample, computed by taking the sums of squares and divide by sample size n.
- s^2: an unbiased estimator for the population variance, often confusingly called the 'sample variance', computed by taking the sums of squares and divide by $n - 1$.

2.4 The standard error

In Chapter 1 we saw that a measure for spread and variability was the variance. In the previous section we saw that with sample size 100, the variability of the sample mean was much lower than with sample size 10. Let's look at this more closely.

When we look at the sampling distribution in Figure 2.2 with sample size 10, we see that the means lie between 2.8 and 3.71. If we compute the standard deviation of the sample means, we obtain a value of 0.118. This standard deviation of the sample means is technically called the *standard error*, in this

case the *standard error of the mean*. It is a measure of how uncertain we are about a population mean when we only have sample data to go on. Think about this: why would we associate a large standard error with very little certainty? In this case we have only 10 data points for each sample, and it turns out that the standard error of the mean is a function of both the sample size n and the population variance σ^2.

$$\sigma_{\bar{y}} = \sqrt{\frac{\sigma^2}{n}} \tag{2.5}$$

Here, the population variance equals 0.14 and sample size equals 10, so the $\sigma_{\bar{y}}$ equals $\sqrt{\frac{0.14}{10}} = 0.118$, close to our observed value. If we fill in the formula for a sample size of 100, we obtain a value of 0.037. This is a much smaller value for the spread and this is indeed observed in Figure 2.4. Figure 2.6 shows the standard error of the mean for all sample sizes between 1 and 200.

In sum, the standard error of the mean is the standard deviation of the sample means, and serves as a measure of the uncertainty about the population mean. The larger the sample size, the smaller the standard error, the closer a sample mean is expected to be around the population mean, the more certain we can be about the population mean.

Similar to the standard error of the mean, we can compute the standard error of the variance. This is more complicated – especially if the population distribution is not normal – and we do not treat it here. Software can do the computations for you, and later in this book you will see examples of the standard error of the variance.

Summarising the above: when we have a population mean, we usually see that the sample mean is close to it, especially for large sample sizes. If you do not understand this yet, go back before you continue reading.

The larger the sample size, the closer the sample means are to the population means. If you turn this around, if you don't know the population mean, you can use a large sample size, calculate the sample mean, and then you have a fairly good estimate for the population. This is useful for our problem of the LH levels, where we have 48 measures. The mean of the 48 measurements could be a good approximation of the mean LH level in general.

As an indication of how close you are to the population mean, the standard error can be used. The standard error of the mean is the standard deviation of the sampling distribution of the sample mean. The smaller the standard error, the more confident you can be that your sample mean is close to the population mean. In the next section, we look at this more closely. If we use our sample mean as our best guess for the population mean, what would be a sensible range of other possible values for the population mean, given the standard error?

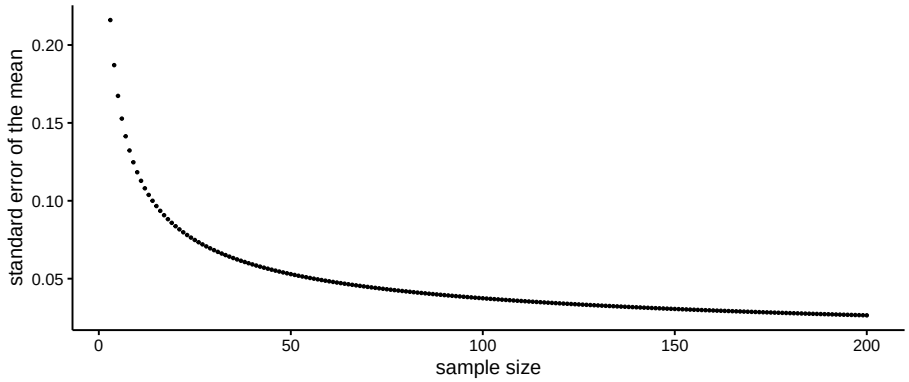


Figure 2.6: Relationship between sample size and the standard error of the mean, when the population variance equals 0.14.

Overview

- **standard error of the mean**: the standard deviation of the distribution of sample means (the sampling distribution of the sample mean). Says something about how spread out the values of the sample means are. It can be used to quantify the uncertainty about the population mean when we only have the sample mean to go on.
- **standard error of the variance**: the standard deviation of the sampling distribution of the sample variance. Says something about how spread out the values of the sample variances are. It can be used to quantify the uncertainty about the population variance when we only have the variance of the sample values to go on.

2.5 Confidence intervals

If we take a sample mean as our best guess of the population mean, we know that we are probably a little bit off. If we have a large standard error we know that the population mean could be very different from our best guess, and if we have a small standard error we know that the true population mean is pretty close to our best guess, but could we quantify this in a better way? Could we give a range of plausible values for the population mean?

In order to do that, let's go back to the elephants: the true population mean is 3.25 m with variance 0.14. What would possible values of sample means look like if sample size is 4? Of course it would look like the sampling distribution of the sample mean with a sample size of 4. Its mean would be the population mean of 3.25 and its standard deviation would be equivalent to the standard

error, computed as a function of the population variance and sample size, in our case $\sqrt{\frac{0.14}{4}} = 0.19$. Now imagine that for a bunch of samples we compute the sample means. We know that the means for large sample sizes will look more or less like a normal distribution, but how about for a small sample size like $n = 4$? If it would look like a normal distribution too, then we could use the knowledge about the standard normal distribution to say something about the distribution of the sample means.

For the moment, let's assume the sample size is not 4, but 4000. From the Central Limit Theorem we know that the distribution of sample means is almost identical to a normal distribution, so let's assume it is normal. From the normal distribution, we know that 68% of the observations lies between 1 *standard deviation* below and 1 *standard deviation* above the mean (see Section 1.19 and Figure 1.9). If we would therefore standardise our sample means, we could say something about their distribution given the standard error, since the standard error is the standard deviation of the sampling distribution. Thus, if the sampling distribution looks normal, then we know that 68% of the sample means lies between one *standard error* below the population mean and one *standard error* above the population mean.

So suppose we take a large number of samples from the population, compute means and variances for each sample, so that we can compute standardised scores. Remember from Chapter 1 that a standardised score is obtained by subtracting an observed score from the mean and divide by the standard deviation:

$$z_y = \frac{y - \bar{y}}{sd_y} \tag{2.6}$$

If we apply standardisation of the sample means, we get the following: for a given sample mean $\bar{y}$ we subtract the population mean μ and divide by the standard deviation of the sample means (the standard error):

$$z_{\bar{y}} = \frac{\bar{y} - \mu}{\sigma_{\bar{y}}} \tag{2.7}$$

If we then have a bunch of standardised sample means, their distribution should have a standard normal distribution with mean 0 and variance 1. We know that for this standard normal distribution, 68% of the values lie between -1 and +1, meaning that 68% of the values in a non-standardised situation lie between -1 and +1 standard deviations from the mean (see Section 1.19). That implies that 68% of the sample means lie between -1 and +1 standard deviations (standard errors!) from the population mean. Thus, 68% of the sample means lie between $-1 \times \sigma_{\bar{y}}$ and $+1 \times \sigma_{\bar{y}}$ from the population mean μ. If we have sample size 4000, $\sigma_{\bar{y}}$ is equal to $\sqrt{\frac{0.14}{4000}} = 0.0059161$ and $\mu = 3.25$, so that 68% of the sample means lie between 3.2440839 and 3.2559161.

This means that we also know that $100 - 68 = 32\%$ of the sample means lie farther away from the mean: that it occurs in only 32% of the samples that a sample mean is smaller than 3.2440839 and larger than 3.2559161. Taking this a bit further, since we know that 95% of the values in a standard normal

distribution lie between -1.96 and +1.96 (see Section 1.19), we know that it happens in only 5% of the samples that the sample mean is smaller than $3.25 - 1.96 \times \sqrt{\frac{0.14}{4000}} = 3.2384045$ or larger than $3.25 + 1.96 \times \sqrt{\frac{0.14}{4000}} = 3.2615955$. Another way of putting this is that it happens in only 95% of the samples that a sample mean is at most $1.96 \times \sqrt{\frac{0.14}{4000}}$ away from the population mean 3.25. This distance of 1.96 times the standard error is called the *margin of error* (MoE). Here we focus on the margin of error that is based on 95% observations of the observations seen in the normal distribution:

$$MoE_{0.95} = z_{0.95} \times \sigma_{\bar{y}} = 1.96 \times \sigma_{\bar{y}} \tag{2.8}$$

where $z_{0.95}$ is the standardised value z for which holds that 95% of the values are between $\mu - z$ and $\mu + z$ (i.e., 1.96).

Knowing the population mean, we know that it is very improbable (5%) that a sample mean is farther away from the population mean than this margin of error. The next step is tricky, so pay close attention. If we know the population mean, we can construct an interval based on the margin of error for where we expect sample means to lie. In the above case, knowing that the population mean is 3.25, and we use an MoE based on 95%, we expect that 95% of the sample means will lie between 3.25 − MoE and 3.25 + MoE.

But what if we don't know the population mean, but do know the sample mean? We could use the same interval but centred around the sample mean instead of the population mean. Thus, we have a 95% interval if we take the sample mean as the centre and the MoE around it. Suppose that we randomly draw 4000 elephants and we obtain a sample mean of $\bar{y} = 3.26$, then we construct the 95% interval as running from $\bar{y} - MoE = 3.26 - MoE$ to $\bar{y} + MoE = 3.26 + MoE$. The margin of error is based on the standard error, which is in turn dependent on the population variance. If we don't know that, we have to estimate it from the sample. So suppose we find a sample variance $s^2 = 0.15$, we get the 95% interval from $\bar{y} - MoE = 3.26 - 1.96 \times \sqrt{\frac{0.15}{4000}}$ to $\bar{y} + MoE = 3.26 + 1.96 \times \sqrt{\frac{0.15}{4000}}$.

Such an interval, centred around the *sample* mean, is called a *confidence interval*. Because it is based on 95% of the sampling distribution (centred around the *population* mean) it is called a 95% confidence interval.

One way of thinking about this interval is that it represents 95% of the sample means *had the population mean been equal to the sample mean*. For example, a 95% interval around the sample mean of 3.26 represents 95% of the sample means that you would get if you would take many random samples from a population distribution with mean 3.26: the middle 95% of the sampling distribution for a population mean of 3.26.

A 95% confidence interval contains 95% of the sample means *had the population mean been equal to the sample mean*. Its construction is based on the estimated sampling distribution of the sample mean.

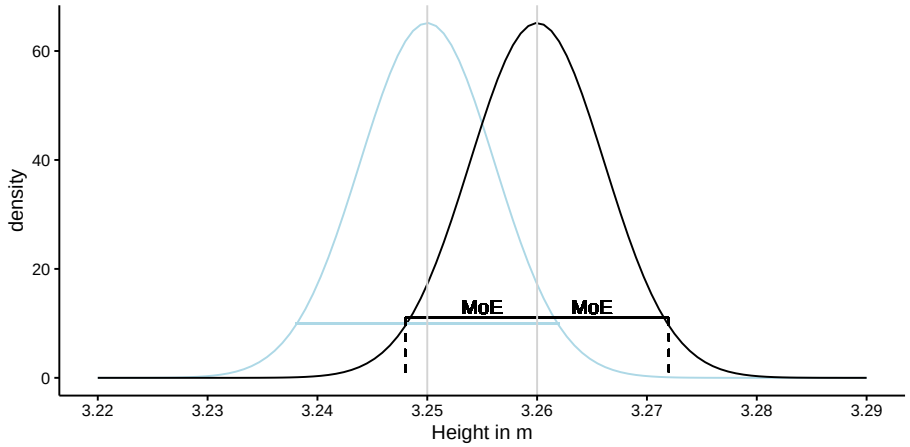


Figure 2.7: Illustration of the construction of a 95% confidence interval. Suppose we find a sample mean of 3.26 and a sample variance of 0.15, with $N = 4000$. The black curve represents the sampling distribution if the population mean would be 3.26 and a variance of 0.15. In reality, we don't know the population mean, it could be 3.25 or any other value. The sampling distribution for 3.25 is shown by the blue curve. Whatever the case, the length of an interval that contains 95% of the sample means is always the same: twice the margin of error. This interval centred around the sample mean, is called the 95% confidence interval.

The idea is illustrated in Figure 2.7. There you see two sampling distributions: one for if the population mean is 3.25 (blue) and one for if the population mean is 3.26 (black). Both are normal distributions because sample size is large, and both have the same standard error that can be estimated using the sample variance. Whatever the true population mean, we can estimate the margin of error that goes with 95% of the sampling distribution. We can then construct an interval that stretches the length of twice the margin of error around any value. We can do that for the real population mean (in blue), but the problem that we face in practice is that we don't know the population mean. We do know the sample mean, and if we centre the interval around that value, we get what is called the 95% confidence interval. We see that it ranges from 3.248 to 3.272. This we can use as a range of plausible values for the unknown population mean. With some level of 'confidence' we can say that the population mean is somewhere in this interval.

Note that when we say: the 95% confidence interval runs from 3.248 to 3.272, we cannot say, we are 95% sure that the population mean is in there. 'Confidence' is not the same as probability. We'll talk about this in a later section. First, we look at the situation where sample size is small so that we cannot use the Central Limit Theorem.

2.6 The t-statistic

In the previous section, we constructed a 95% confidence interval based on the standard normal distribution. We know from the standard normal distribution that 95% of the values are between -1.96 and +1.96. We used the standard normal distribution because the sampling distribution will look normal if sample size is large. We took the example of a sample size of 4000, and then this approach works fine, but remember that the actual sample size was 4. What if sample size is not large? Let's see what the sampling distribution looks like in that case.

Remember from the previous section that we standardised the sample means.

$$z_{\bar{y}} = \frac{\bar{y} - \mu}{\sigma_{\bar{y}}}$$

and that $z_{\bar{y}}$ has a standard normal distribution. But, this only works if we have a good estimate of $\sigma_{\bar{y}}$, the standard error. If sample size is limited, our estimate is not perfect. You can probably imagine that if you take one sample of 4 randomly selected elephants, you get one value for the estimated standard error ($\sqrt{\frac{s^2}{n}}$), and if you take another sample of 4 elephants, you get a slightly different value for the estimated standard error. Because we do not always have a good estimate for $\sigma_{\bar{y}}$, the standardisation becomes a bit more tricky. Let's call the standardised sample mean t instead of z:

$$t_{\bar{y}_i} = \frac{\bar{y}_i - \mu}{\sqrt{\frac{s_i^2}{n}}}$$

Thus, a standardised sample mean for sample i, will be constructed using an estimate for the standard error by computing the sample variance s^2 for sample i.

If you standardise every sample mean, each time using a slighly different standard deviation, and you plot a histogram of the t-values, you do not get a standard normal distribution, but a slightly different one.

In summary: if you know the standard error (because you know the population variance), the standardised sample means will show a normal distribution. If you don't know the standard error, you have to estimate it based on the sample variance. If sample size is really large, you can estimate the population variance pretty well, and the sample variances will be very similar to each other. In that case, the sampling distribution will look very much like a normal distribution. But if sample size is relatively small, each sample will show a different sample variance, resulting in different standard error estimates. If you standardise each sample mean with a different standard error, the sampling distribution will not look normal. This distribution is called a t-distribution. The difference between this distribution and the standard normal distribution is shown in Figure 2.8. The blue curve is the standard normal distribution, the

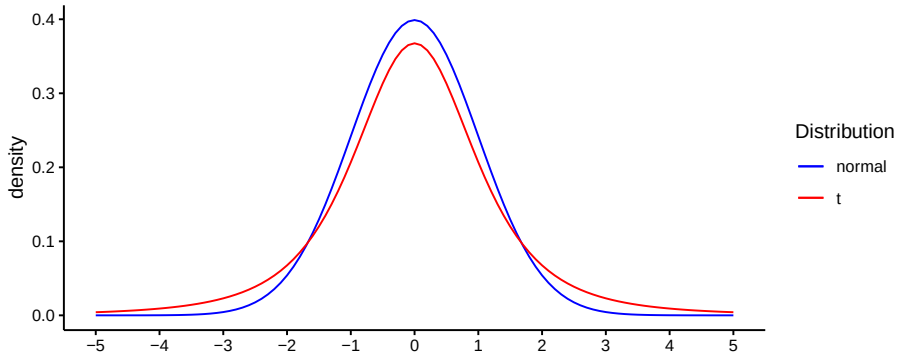


Figure 2.8: Distribution of t with sample size 4, compared with the standard normal distribution.

red curve is the distribution we get if we have sample size 4 and we compute $t_{\bar{y}_i} = \frac{\bar{y}_i - \mu}{\sqrt{\frac{s_i^2}{n}}}$ for many different samples.

When you compare the two distributions, you see that compared to the normal curve, there are fewer observations around 0 for the t-distribution: the density around 0 is lower for the red curve than for the blue curve. That's because there are more observations far away from 0: in the tails of the distributions, you see a higher density for the red curve (t) than for the blue curve (normal). They call this phenomenon 'heavy-tailed': relatively more observations in the tails than around the mean.

That the t-distribution is heavy-tailed has important implications. From the standard normal distribution, we know that 5% of the observations lie more than 1.96 away from the mean. But since there are relatively more observations in the tails of the t-distribution, 5% of the values lie farther away from the mean than 1.96. This is illustrated in Figure 2.9. If we want to construct a 95% confidence interval, we can therefore no longer use the 1.96 value.

With this t-distribution, 95% of the observations lie between -3.18 and +3.18. Of course, that is in the standardised situation. If we move back to our scale of elephant heights with a sample mean of 3.26, we have to transform this back to elephant heights. So -3.18 times the standard error away from the mean of 3.26, is equal to $3.26 - 3.18 \times \sqrt{\frac{0.15}{4}} = 2.6441956$, and +3.18 times the standard error away from the mean of 3.26, is equal to $3.26 + 3.18 \times \sqrt{\frac{0.15}{4}} = 3.8758044$. So the 95% interval runs from 2.64 to 3.88. This interval is called the 95% confidence interval, because 95% of the sample means will lie in this interval, if the population mean would be 3.26.

Notice that the interval includes the population mean of 3.25. If we would interpret this interval around 3.26 as containing plausible values for the population mean, we see that in this case, this is a fair conclusion, because the true

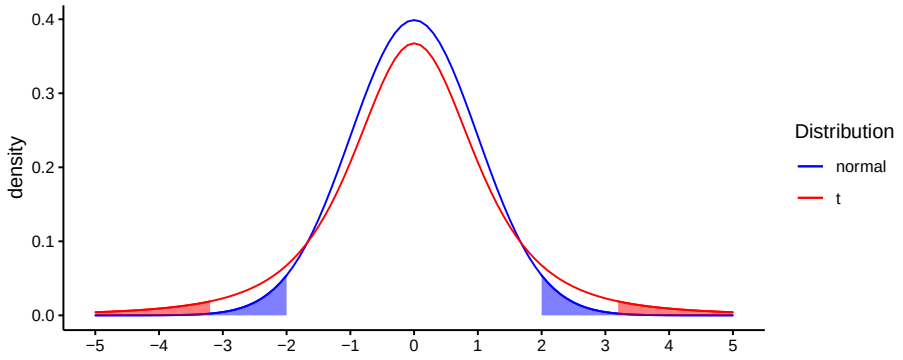


Figure 2.9: Distribution of t with sample size 4, compared with the standard normal distribution. Shaded areas represent 2.5% of the respective distribution.

value 3.25 lies within this interval.

2.7 Interpreting confidence intervals

The interpretation of confidence intervals is very difficult, and it often goes wrong, even in many textbooks on the matter.

One thing that should be very clear is that a confidence interval is constructed *as if you know the population mean and variance*, which, of course, you don't. We assume that the population is a certain value, say $\mu = m_0$, we assume that the standard error of the mean is equal to $\sigma_{\bar{y}} = \sqrt{\frac{s^2}{n}}$, and we know that if we would look at many many samples and compute standardised sample means, their distribution would be a t-distribution. Based on that t-distribution, we know in which interval 95% of the standardised sample means would lie and we use that to compute the margin of error and to construct an interval around the sample mean that we actually obtain. A lot of this reasoning is imagination: imagining that you know the population mean and that you have a good estimate for the population variance. Then you imagine what sample means would be reasonable to find. But of course , it's in fact the opposite: you only know the sample mean and sample variance and you want to know what are plausible values for the population mean.

You have to bear this reversal in mind when interpreting the 95% confidence interval around a sample mean. Many people state the following: with 95% probability, the 95% confidence interval contains the population mean. This is wrong. It is actually the opposite: the 95% interval around the population mean contains 95% of the sample means.

If you know the population mean μ, then 95% of the confidence intervals that you construct around the sample means that you get from random sampling will contain the mean μ. This is illustrated in Figure 2.10. Suppose we take $\mu = 3.25$.

Then if we imagine that we take 100 random samples from this population distribution, we can calculate 100 sample means and 100 sample variances. If we then construct 100 confidence intervals around these 100 sample means, we obtain the confidence intervals displayed in Figure 2.10. We see that 95 of these intervals contain the value 3.25, and 5 of them don't: only in samples 1, 15, 20, 28 and 36, the interval does not contain 3.25.

It can be mathematically shown that given a certain population mean, when taking many, many samples and constructing 95% confidence intervals, you can expect 95% of them will contain that population mean. That does *not* mean however that given a sample mean with a certain 95% interval, that interval contains the population mean with a probability of 95%. It only means that were this procedure of constructing confidence intervals to be repeated on numerous samples, the fraction of calculated confidence intervals that contain the true population mean would tend toward 95%. If you only do it once (you obtain a sample mean and you calculate the 95% confidence interval) it either contains the population mean or it doesn't: you cannot calculate a probability for this. In the statistical framework that we use in this book, one can only say something about the probability of data occurring given some population values:

Given that the population value is 3.25, and if you take many, many independent samples from the population, you can expect that 95% of the confidence intervals constructed based on resulting sample means will contain that population value of 3.25.

Using this insight, we therefore conclude that the fact we see the value of 3.25 in our 95% confidence interval around 2.9, gives us some reason to believe ('confidence') that 3.25 could also be a plausible candidate for the population mean.

Summarising, if we find a sample mean of say 2.9, we know that 2.9 is a reasonable guess for the population mean (it's an unbiased estimator). Moreover, if we construct a 95% confidence interval around this sample mean, this interval contains other plausable candidates for the population mean. However, it might be possible that the true population mean is not included.

2.8 t-distributions and degrees of freedom

The standardised deviation of a sample mean from a hypothesised population mean has a t-distribution. This happens when the population variance is not known, and we therefore have to estimate the standard error based on the sample variance. Because of this uncertainty about the population variance and consequently the standard error, the standardised score does not have a normal distribution but a t-distribution.

In the previous section we saw the distribution for the case that we had a sample size of 4. With such a small sample size, we have a very inaccurate estimate of the population variance. The sample variance s^2 will be very different

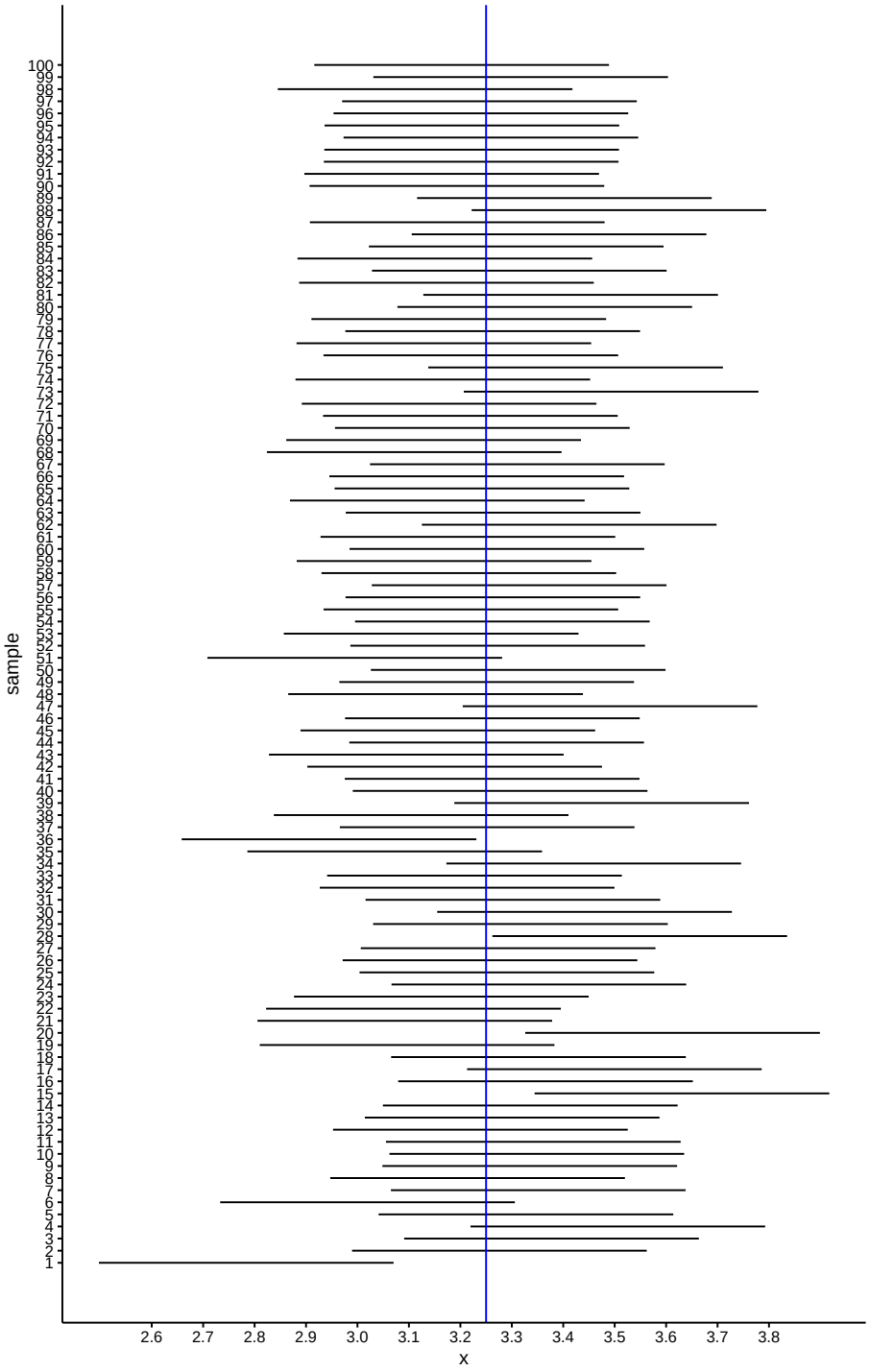


Figure 2.10: Confidence intervals

for every new sample of size 4. But if sample size increases, our estimates for the population variance will become more precise, and they will show less variability. This results in the sampling distribution to become less heavy-tailed, until it closely resembles the normal distribution for very large sample sizes.

This means that the shape of the sampling distribution is a t-distribution but that the shape of this t-distribution depends on sample size. More precisely, the shape of the t-distribution depends on its so-called *degrees of freedom* (explained below). Degrees of freedom are directly linked to the sample size. Degrees of freedom can be as small as 1, very large like 250, or infinitely larger. The t-distribution with a very large number like 2500, is practically indistinguishable from a normal distribution. However for a relatively low number of degrees of freedom, the shape is very different: relatively more observations are in the tails of the distribution and less so in the middle, compared to the normal distribution, see Figure 2.11.

The shape of the t-distribution is determined by its degrees of freedom: the higher the degrees of freedom, the more it resembles the normal distribution. So which t-distribution do we have to use when we are dealing with sample means and we want to infer something about the population mean, and what are degrees of freedom? As stated already above, the degrees of freedom is directly related to sample size: sample size determines the degrees of freedom of the t-distribution that we need. Degrees of freedom stands for the amount of information that we have and of course that depends on how many data values we have. In its most general case, the number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary. More specifically in our case, the degrees of freedom for a statistic like t are equal to the number of independent scores that go into the estimate, minus the number of parameters used as intermediate steps in the estimation of the parameter itself.

In the example above we had information about 4 elephants (4 values), so our information content is 4. However, remember that when we construct our t-value, we have to first compute the sample mean in order to compute the sample variance s^2. But, suppose you know the sample mean, you don't have to know all the 4 values anymore. Suppose the heights of the first three elephants are 3.24, 3.25 and 3.26, and someone computes the mean of all four elephants as 3.25, then you automatically know that the fourth elephant has a height of 3.25 (why?). Thus, once you know the mean of n elephants, you can give imaginary values for the heights of only $n-1$ elephants, because given the other heights and the mean, it is already determined.

The same is true for the t-statistic: once you know 3 elephant heights and statistic t, then you know the height of the fourth elephant automatically.

Because we assume the mean in our computation of s^2 (we fix it) we loose one information point, leaving 3. The shape of the standardised scores of fictitious new samples then looks like a t-distribution with 3 degrees of freedom.

Generally, if we have a sample size of n and the population variance is unknown, the shape of the standardised sample means (i.e., t-scores) of fictitious new samples is that of a t-distribution with $n-1$ degrees of freedom.

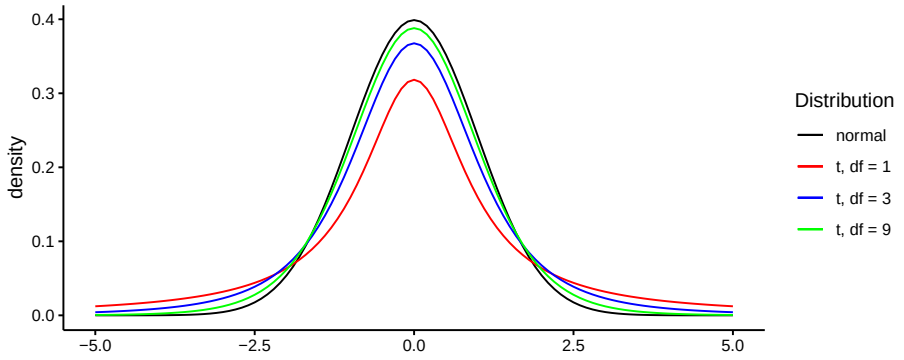


Figure 2.11: Difference in the shapes of the standard normal distribution and t-distributions with 1, 3 and 9 degrees of freedom.

2.9 Constructing confidence intervals

In previous sections we discussed the 95% confidence interval, because it is the most widely used interval. But other intervals are also seen, for instance 99% confidence intervals or 90% confidence intervals. A 99% confidence interval is wider than a 95% confidence interval, which in turn is wider than a 90% confidence interval. The width of the confidence interval also depends on the sample size. Here we show how to construct 90%, 99% and other intervals, for different sample sizes.

As we discussed for the 95% interval above, we looked at the t-distribution of 3 degrees of freedom because we had a sample size of 4 elephants. Suppose we have a sample size of 200, then we would have to look at a t-distribution of $200-1 = 199$ degrees of freedom. Table 2.2 shows information about a couple of t-distributions with different degrees of freedom. In the first column, cumulative probabilities are given, and the next column gives the respective quantiles. For instance, the column 'norm' shows that a cumulative proportion of 0.025 is associated with a quantile of -1.96 for the standard normal distribution. This means that for the normal distribution, 2.5% of the observations are smaller than -1.96. In the same column we see that the quantile 1.96 is associated with a cumulative probability of 0.975. This means that 97.5% of the observations in a normal distribution are smaller than 1.96. This implies that 100% - 97.5% = 2.5% of the observations are larger than 1.96. Thus, if 2.5% of the observations are larger than 1.96 and 2.5% of the observations are smaller than -1.96, then 5% of the observations are outside the interval (-1.96, 1.96), and 95% are inside this interval.

From Table 2.2, we see that for such a 95% interval, we have to use the values -1.96 and 1.96 for the normal distribution, but for the t-distribution we have to use other values, depending on the degrees of freedom. We see that for 3 degrees of freedom, we have to use the values -3.18 and 3.18, and for 199 degrees

of freedom the values -1.97 and +1.97. This means that for a t-distribution with 3 degrees of freedom, 95% of the observations lie in the interval from -3.18 to 3.18. Similarly, for a t-distribution with 199 degrees of freedom, the values for cumulative probabilities 0.025 and 0.975 are -1.97 and 1.97 respectively, so we can conclude that 95% of the observations lie in the interval from -1.97 to 1.97.

Now instead of looking at 95% intervals for the t-distribution, let's try to construct a 90% confidence interval around an observed sample mean. With a 90% confidence interval, 10% lies outside the interval. We can divide that equally to 5% on the low side and 5% on the high side. We therefore have to look at cumulative probabilities 0.05 and 0.95 in Table 2.2. The corresponding quantiles for the normal distribution are -1.64 and 1.64, so we can say that for the normal distribution, 90% of the values lie in the interval (-1.64, 1.64). For a t-distribution with 9 degrees of freedom, we see that the corresponding values are -1.83 and 1.83. Thus we conclude that with a t-distribution with 9 degrees of freedom, 90% of the observed values lie in the interval (-1.83, 1.83).

However, now note that we are not interested in the values of the t-distribution, but in likely values for the population mean. The standard normal and the t-distribution are standardised distributions. In order to get values for the confidence interval around the sample mean, we have to unstandardise the values. The value of 1.83 above means "1.83 standard errors away from the mean (the sample mean)". So suppose we find a sample mean of 3, with a standard error of 0.5, then we say that a 90% confidence interval for the population mean runs from $3 - 1.83 \times 0.5$ to $3 + 1.83 \times 0.5$, so from 1.375 to 4.625.

Follow these steps to compute a x% confidence interval:

Constructing confidence intervals

1. Compute the sample mean $\bar{y}$.
2. Estimate the population variance $s^2 = \frac{\Sigma_i (y_i - \bar{y})}{n-1}$.
3. Estimate the standard error $\hat{\sigma}_{\bar{y}} = \sqrt{\frac{s^2}{n}}$.
4. Compute degrees of freedom as $n - 1$.
5. Look up $t_{\frac{1-x}{2}}$. Take the t-distribution with the right number of degrees of freedom and look for the critical t-value for the confidence interval: if x is the confidence level you want, then look for quantile $\frac{1-x}{2}$. Then take its absolute value. That's your $t_{\frac{1-x}{2}}$.
6. Compute margin of error (MoE) as $\text{MoE} = t_{\frac{1-x}{2}} \times \hat{\sigma}_{\bar{y}}$.
7. Subtract and sum the sample mean with the margin of error: $(\bar{y} - \text{MoE}, \bar{y} + \text{MoE})$.

Note that for a large number of degrees of freedom, the values are very close

to those of the standard normal.

Table 2.2: Quantiles for the standard normal and several t-distributions.

probs	norm	t199	t99	t9	t5	t3
0.0005	-3.29	-3.34	-3.39	-4.78	-6.87	-12.92
0.0010	-3.09	-3.13	-3.17	-4.30	-5.89	-10.21
0.0050	-2.58	-2.60	-2.63	-3.25	-4.03	-5.84
0.0100	-2.33	-2.35	-2.36	-2.82	-3.36	-4.54
0.0250	-1.96	-1.97	-1.98	-2.26	-2.57	-3.18
0.0500	-1.64	-1.65	-1.66	-1.83	-2.02	-2.35
0.1000	-1.28	-1.29	-1.29	-1.38	-1.48	-1.64
0.9000	1.28	1.29	1.29	1.38	1.48	1.64
0.9500	1.64	1.65	1.66	1.83	2.02	2.35
0.9750	1.96	1.97	1.98	2.26	2.57	3.18
0.9900	2.33	2.35	2.36	2.82	3.36	4.54
0.9950	2.58	2.60	2.63	3.25	4.03	5.84
0.9990	3.09	3.13	3.17	4.30	5.89	10.21
0.9995	3.29	3.34	3.39	4.78	6.87	12.92

2.10 Obtaining a confidence interval for a population mean in R

Suppose we have values on miles per gallon (mpg) in a sample of cars, and we wish to construct a 99% confidence interval for the population mean. We can do that in the following manner. We take all the mpg values from the mtcars data set, and set our confidence level to 0.99 in the following manner:

```
t.test(mtcars$mpg, conf.level = 0.99)$conf.int

## [1] 17.16706 23.01419
## attr(,"conf.level")
## [1] 0.99
```

It shows that the 99% confidence interval runs from 17.2 to 23.0. The t.test() function does more than simply constructing confidence intervals. That is the topic of the next section.

2.11 Null-hypothesis testing

Suppose a professor of biology claims, based on years of measuring the height of elephants in Tanzania, that the mean height of elephants in Tanzania is 3.38 m. Suppose that you come up with data on a relatively small number of South-African elephants and the professor would like to know whether the two

groups of elephants have the same population mean. Do both the Tanzanian and South-African populations have the same mean of 3.38, or is there perhaps a difference in the means? A difference in means could indicate that there are genetic differences between the two elephant populations. The professor would like to base her conclusion on your sample of data, and you assume that the professor is right in that the population mean of Tanzanian elephants is 3.38 m.

One way of addressing a question like this is to look at the confidence interval for the South-African mean. Suppose you construct a 95% confidence interval. Based on a sample mean of 3.27, a sample variance s^2 of 0.14 and a sample size of 40, you calculate that the interval runs from 3.15 to 3.39. Based on that interval, you can conclude that 3.38 is a reasonable value for the population mean, and that it could well be that the both the Tanzanian and South-African populations have the same mean height of 3.38 m.

However, as we have seen in the previous section, there are many confidence intervals that we could compute. If instead of the 95% confidence interval, we would compute a 90% confidence interval, we would end up with an interval that runs from 3.17 to 3.37. In that case, the Tanzanian population mean is no longer included in the confidence interval for the South-African population mean, and we'd have to conclude that the populations have different means.

What interval to choose? Especially if you have questions like "Do the two populations have the same mean" and you want to have a clear yes or no answer, then *null-hypothesis testing* might be a solution. With null-hypothesis testing, a null-hypothesis is stated, after which you decide based on sample data whether or not the evidence is strong enough to reject that null-hypothesis. In our example, the null-hypothesis is that the South-African mean has the value 3.38 (the Tanzanian mean). We write that as follows:

$$H_0 : \mu_{SA} = 3.38 \tag{2.9}$$

We then look at the data on South-African elephants that could give us evidence that is either in line with this hypothesis or not. If it is not, we say that we reject the null-hypothesis.

The objective of null-hypothesis testing is that we either reject the null-hypothesis, or not. This is done using the data from a sample. In the null-hypothesis procedure, we simply assume that the null-hypothesis is true, and *compare the sample data with data that would result if the null-hypothesis were true.*

So, let's assume the null-hypothesis is true. In our case that means that the mean height of all South-African elephants is equal to that of all Tanzanian elephants, namely 3.38 m. Next, we compare our actual observed data with data that would *theoretically* result from a population mean of 3.38. What would sample data theoretically look like if the population mean is 3.38? In the previous sections, we learned what possible sample means would look like. Thus, let's focus on the sample mean.[2]

[2]The sample mean is called a *sufficient statistic* for the population mean. That means, if you want to know something about the population mean, the only information you need to

Based on what we learned about the sampling distribution of the sample mean, we know that possible values for the sample mean come from this distribution. It is more or less a normal distribution with mean 3.38, but what the variance is (the standard error), we don't know. We'd have to take a guess, based on the sample data that we have. Based on the sample data, we could compute the sample variance s^2, and then estimate the standard error as $\hat{\sigma}_{\bar{y}} = \sqrt{\frac{s^2}{n}}$. However, as we saw earlier, because we have to estimate the standard error, the sample means are no longer normally distributed, but t-distributed.

Suppose we observe 40 South-African elephants, and we obtain a sample mean of 3.27 and a sample variance s^2 of 0.14. The hypothesised population mean is 3.38. We know that the sampling distribution is a t-distribution because we do not know the population variance. To know the shape of the sampling distribution, we need three things: the mean of the sampling distribution (assuming the population mean is 3.38), the standard deviation (or variance) of the distribution, and the exact shape of the t-distribution (the degrees of freedom). The mean is easy: that is equal to the hypothesised population mean of 3.38 (why?). The standard deviation (standard error) is more difficult, but we can use the sample data to estimate it. We compute it using the sample variance: $\hat{\sigma}_{\bar{y}} = \sqrt{\frac{s^2}{n}} = 0.059$. And the last bit is easy: the degrees of freedom is simply sample size minus 1: $40 - 1 = 39$.

We plot this sampling distribution of the sample mean in Figure 2.12. This Figure tells us that if the null-hypothesis is really true and that the South-African mean height is 3.38, and we would take many different random samples of 40 elephants, we would see only sample means between 3.20 and 3.35. Other values are in fact possible, but very unlikely. But how likely is our observed sample mean of 3.27: do we feel that it is a likely value to find if the population mean is 3.38, or is it rather unlikely?

What do you think? Think this over for a bit before you continue to read.

In fact, every unique value for a sample mean is rather unlikely. If the population mean is 3.38, it will be very improbable that you will find a sample mean of exactly 3.38, because by sheer chance it could also be 3.39, or 3.40 or 3.37. But relatively speaking, those values are all more likely to find than more deviant values. The density curve tells you that values *around* 3.38 are more likely than values around 3.27 or 3.50, because the density is higher around the value of 3.38 than around those other values.

What to do?

The solution is to define *regions* for sample means where we think the sample mean is no longer probable under the null-hypothesis, and a region where it is probable enough to believe that the null-hypothesis could be true.

For example, we could define an *acceptance region* where 95% of the sample means would fall if the null-hypothesis is true, and a *rejection region* where only 5% of the sample means would fall if the null-hypothesis is true. Let's put the

get from the sample data is the mean of the sample values. Knowing the exact values does not give you extra information: the sample mean *suffices*. The proof for this is beyond this book.

rejection region in the tails of the distribution, where the most extreme values can be found (farthest away from the mean). We put half of the rejection region in the left tail and half of it in the right tail of the distribution, so that we have two regions that each covers 2.5% of the sampling distribution. These regions are displayed in Figure 2.13. The red ones are the rejection regions, and the green one is the acceptance region (covering 95% of the area).

Why 5%, why not 10% or 1%? Good question. It is just something that is accepted in a certain group of scientists. In the social and behavioural sciences, researchers feel that 5% is a small enough chance. In contrast, in quantum mechanics, researchers feel that 0.000057% is a small enough chance. Both values are completely arbitrary. We'll dive deeper into this arbitrary chance level in a later section. For now, we continue to use 5%.

From Figure 2.13 we see that the sample mean that we found for your 40 South-African elephants (3.27) does not lie in the red rejection region. We see that 3.27 lies well within the green section where we decide that sample means are likely to occur when the population is 3.38. Because this is likely, we think that the null-hypothesis is plausible: if the population mean is 3.28, it is plausible to expect a sample mean of 3.27, because in 95% of random samples we would see a sample mean between 3.255 and 3.500. The value 3.27 is a very reasonable value and we therefore do not reject the null-hypothesis. We conclude therefore that it could well be that both Tanzanian and South-African elephants have the same average height of 3.38, that is, we do not have any evidence that the population mean is *not* 3.38.

This is the core of null-hypothesis testing for a population mean: 1) you determine a null-hypothesis that states that the population mean has a certain value, 2) you figure out what kind of sample means you would get if the population mean would have that value, 3) you see if the sample mean that you actually have is far enough from the population mean to say that it is unlikely enough to result from the hypothesised population mean. If that is the case, then you reject the null-hypothesis, meaning you don't believe in it. If it is likely to result from the hypothesised population, you do not reject the null-hypothesis: there is no reason to suspect that the null-hypothesis is false.

2.12 Null-hypothesis testing with t-values

In the above example, we looked explicitly at the sampling distribution for a hypothesised value for the population mean. By determining what the distribution would look like (determining the mean, standard error and degrees of freedom), we could see whether a certain sample mean would give enough evidence to reject the null-hypothesis.

In this section we will show how to do this hypothesis testing more easily by first standardising the problem. The trick is that we do not have to make a picture of the sampling distribution every time we want to do a null-hypothesis test. We simply know that its shape is that of a t-distribution with degrees of freedom equal to $n - 1$. t-distributions are standardised distributions, always

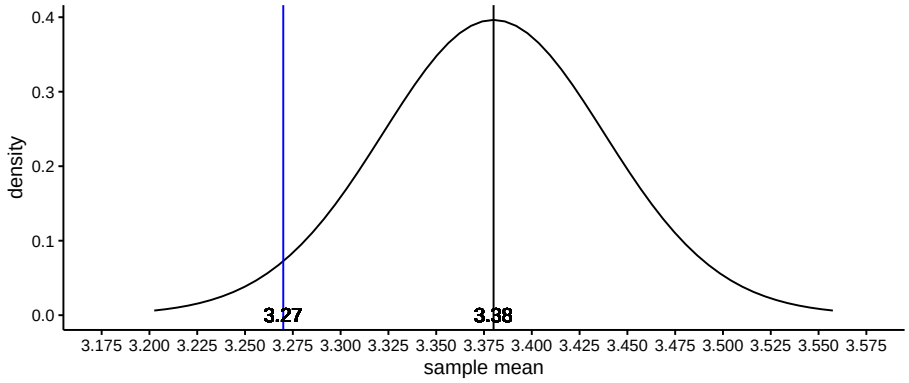


Figure 2.12: The sampling distribution under the null-hypothesis that the South-African population mean is 3.38. The blue line represents the sample mean for our observed sample mean of 3.27.

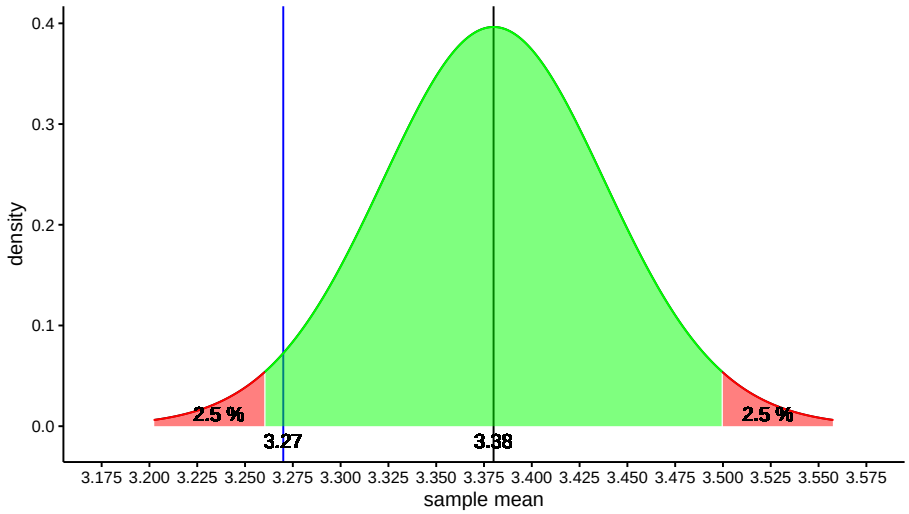


Figure 2.13: The sampling distribution under the null-hypothesis that the South-African population mean is 3.38. The red area denotes the range of values for which the null-hypothesis is rejected (rejection region), the green area denotes the range of values for which the null-hypothesis is not rejected (acceptance region).

with a mean of 0. They are the distribution of standardised t-statistics, where a sample mean is standardised by subtracting the population mean and dividing the result by the standard error.

Let's do this standardisation for our observed sample mean of 3.27. With a population mean of 3.38 and a standard error of $\hat{\sigma}_{\bar{y}} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{0.14}{40}} = 0.059$, we obtain:

$$t = \frac{3.27 - 3.38}{0.059} = -1.864 \tag{2.10}$$

We can then look at a t-distribution of $40 - 1 = 39$ degrees of freedom to see how likely it is that we find such a t-score if the null-hypothesis is true. The t-distribution with 39 degrees of freedom is depicted in Figure 2.14. Again we see the population mean represented, now standardised to a t-score of 0 (why?), and the observed sample mean, now standardised to a t-score of -1.864. As you can see, this graph gives you the same information as the sampling distribution in Figure 2.13. The advantage of using standardisation and using the t-distribution is that we can now easily determine whether or not an observed sample mean is somewhere in the red zone or in the green zone, without making a picture.

We have to find the point in the t-distribution where the red and green zones meet. These points in the graph are called *critical values*. From Figure 2.14 we can see that these critical values are around -2 and 2. But where exactly? This information can be looked up in the t-tables that were discussed earlier in this chapter. We plot such a table again in Table 2.3. A larger version is given in Appendix B.

In such a table, you can look up the 2.5th percentile. That is, the value for which 2.5% of the t-distribution is equal or smaller. Because we are dealing with a t-distribution with 39 degrees of freedom, we look in the column t39, and then in the row with cumulative probability 0.025 (equal to 2.5%), we see a value of -2.02. This is the critical value for the lower tail of the t-distribution. To find the critical value for the upper tail of the distribution, we have to know how much of the distribution is lower than the critical value. We know that 2.5% is higher, so it must be the case that the rest of the distribution, 100-2.5 = 97.5% is lower than that value. This is the same as a probability of 0.975. If we look for the critical value in the table, we see that it is 2.02. Of course this is the opposite of the other critical value, because the t-distribution is symmetrical.

Now that we know that the critical values are -2.02 and +2.02, we know that for our standardised t-score of -1.864 we are still in the green area, so we do not reject the null-hypothesis. We don't need to draw the distribution any more. For any value, we can directly compare it to the critical values. And not only for this example of 40 elephants and a sample mean of 3.27, but for any combination.

Suppose for example that we would have had a sample size of 10 elephants, and we would have found a sample mean of 3.28 with a slightly different sample variance, $s^2 = 0.15$. If we want to test the null-hypothesis again that the population mean is 3.38 based on these results, we would have to do the following

steps:

Null-hypothesis testing

1. Estimate the standard error $\hat{\sigma}_{\bar{y}} = \sqrt{\frac{s^2}{n}}$.
2. Calculate the t-statistic $t = \frac{\bar{y}-\mu}{\hat{\sigma}_{\bar{y}}}$, μ is the population mean under the null-hypothesis.
3. Determine the degrees of freedom, $n - 1$.
4. Determine the critical values for lower and upper tail of the appropriate t-distribution, using Appendix B.
5. If the t-statistic is between the two critical values, then we're in the green, we still believe the null-hypothesis is plausible.
6. If the t-statistic is not between the two critical values, we are in the red zone and we reject the null-hypothesis.

So let's do this for our hypothetical result:

1. Estimate the standard error: $\sqrt{\frac{0.15}{10}} = 0.1224745$
2. Calculate the t-statistic: $t = \frac{3.28-3.38}{0.1224745} = -0.8164966$
3. Determine the degrees of freedom: sample size minus 1 equals 9
4. In Table 2.3 we look for the row with probability 0.025 and the column for t9. We see a value of -2.26. The other critical value then must be 2.26.
5. The t-statistic of -0.8164966 lies between these two critical values, so these sample data do not lead to a rejection of the null-hypothesis that the population mean is 3.38. In other words, these data from 10 elephants do not give us reason to doubt that the population mean is 3.38.

2.13 The p-value

What we saw in the previous section was the classical null-hypothesis testing procedure: calculating a t-statistic and determine whether or not this t-score is in the red zone or green zone, by comparing them to critical values. In the old days, this was done by hand: the calculation of t and looking up the critical values in tables published in books.

These days we have the computer do the work for us. If you have a data set, a program can calculate the t-score for you. However, when you look at the output, you actually never see whether this t-score leads to a rejection of the

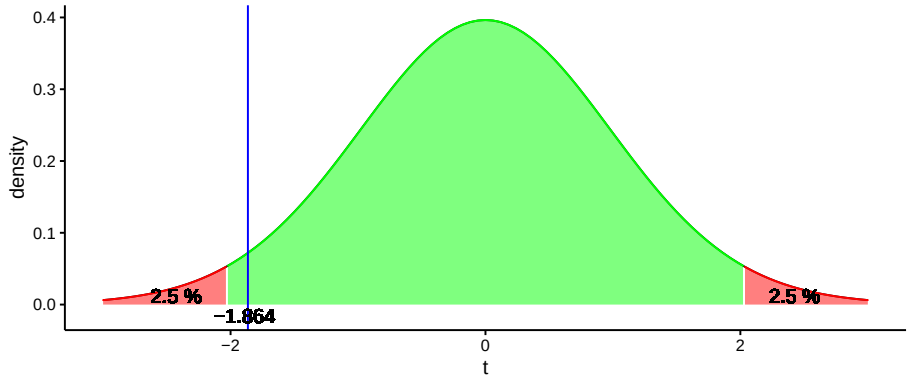


Figure 2.14: A t-distribution with 39 degrees of freedom to test the null-hypothesis that the South-African population mean is 3.38. The blue line represents the T-score for our observed sample mean of 3.27.

null-hypothesis or not. The only thing that a computer prints out is the t-score, the degrees of freedom, and a so-called p-value. In this section we explain what a p-value is and how you can use it for null-hypothesis testing.

Let's go back to our example in the previous section, where we found a sample mean height of 3.28 with only 10 elephants. We computed the t-score and obtained -0.82. We illustrate this result in Figure 2.15 where the red line indicates the t-score. By comparing this t-value with the critical values, we could decide that we do not reject the null-hypothesis. However, if you would do this calculation with a computer program like R, we would get the following result:

```
t = -0.82, df = 9, p-value = 0.434
```

Figure 2.15 shows what this p-value of 0.434 means. The green area in the middle represents the probability that a t-score lies between -0.82 and +0.82. That probability is shown in the figure as 0.567, so 56.7%. The left blue region represents the probability that if the null-hypothesis is true, the t-score will be less than -0.82. That probability is 0.217, so 21.7%. Because of symmetry, the probability that the t-score is more than 0.82 is also 0.217. The blue regions together therefore represent the probability that you find a t-score of less than -0.82 or more than 0.82, and that probability equals $0.217 + 0.217 = 0.434$. Therefore, the probability that you find a t-value of ± 0.82 or more extreme equals 0.434. This probability is called the p-value.

Why is this value useful?

Let's imagine that we find a t-score of exactly equal to one of the critical values. The critical value for a sample size of 10 animals related to a cumulative proportion of 0.025 equals -2.26 (see Table 2.3). Based on this table, we know that the probability of a t-value of -2.26 or lower equals 0.025. Because of symmetry, we also know that the probability of a t-value of -2.26 or higher also

Table 2.3: Quantiles for the standard normal and several t-distributions.

probs	norm	t199	t99	t47	t39	t9	t5	t3
0.0005	-3.29	-3.34	-3.39	-3.51	-3.56	-4.78	-6.87	-12.92
0.0010	-3.09	-3.13	-3.17	-3.27	-3.31	-4.30	-5.89	-10.21
0.0050	-2.58	-2.60	-2.63	-2.68	-2.71	-3.25	-4.03	-5.84
0.0100	-2.33	-2.35	-2.36	-2.41	-2.43	-2.82	-3.36	-4.54
0.0250	-1.96	-1.97	-1.98	-2.01	-2.02	-2.26	-2.57	-3.18
0.0500	-1.64	-1.65	-1.66	-1.68	-1.68	-1.83	-2.02	-2.35
0.1000	-1.28	-1.29	-1.29	-1.30	-1.30	-1.38	-1.48	-1.64
0.9000	1.28	1.29	1.29	1.30	1.30	1.38	1.48	1.64
0.9500	1.64	1.65	1.66	1.68	1.68	1.83	2.02	2.35
0.9750	1.96	1.97	1.98	2.01	2.02	2.26	2.57	3.18
0.9900	2.33	2.35	2.36	2.41	2.43	2.82	3.36	4.54
0.9950	2.58	2.60	2.63	2.68	2.71	3.25	4.03	5.84
0.9990	3.09	3.13	3.17	3.27	3.31	4.30	5.89	10.21
0.9995	3.29	3.34	3.39	3.51	3.56	4.78	6.87	12.92

equals 0.025. This brings us to the conclusion that the probability of a t-score of ± 2.26 or more extreme, is equal to $0.025 + 0.025 = 0.05 = 5\%$. Thus, when the t-score is equal to the critical value, then the p-value is equal to 5%. You can imagine that if the t-score becomes more extreme than the critical value the p-value will become less than 5%, and if the t-score becomes less extreme (closer to 0), the p-value becomes larger.

In the previous section, we said that if a t-score is more extreme than one of the critical values (when it doesn't have a value between them) then we reject the null-hypothesis. Thus, a p-value of 5% or less means that we have a t-score more extreme than the critical values, which in turn means we have to reject the null-hypothesis. Thus, based on the computer output, we see that the p-value is larger than 0.05, so we do not reject the null-hypothesis.

Overview

- critical value: the minimum (or maximum) value that a t-score should have to be in the red zone (the rejection region). If a t-value is more extreme than a critical value, then the null-hypothesis is rejected. The red zone is often chosen such that a t-score will be in that zone 5% of the time, assuming that the null-hypothesis is true.
- p-value: indicates the probability of finding a t-value equal or more extreme than the one found, assuming that the null-hypothesis is true. Often a p-value of less than 5% is used to support the conclusion that the null-hypothesis is not tenable. This is equivalent to a rejection region of 5% when using critical values.

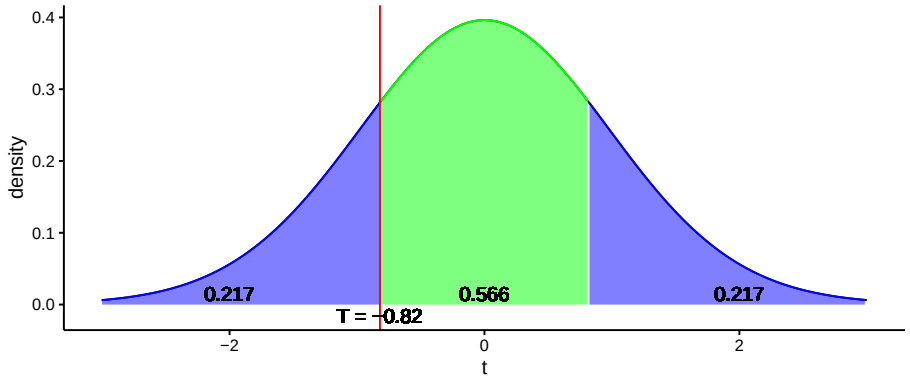


Figure 2.15: Illustration of what a p-value is. The total blue area represents the probability that under the null-hypothesis, you find a more extreme value than the t-score or its opposite. The blue area covers a proportion of .217 + .217 = 0.434 of the t-distribution. This amounts to a p-value of .434.

Let's apply this null-hypothesis testing to our luteinising hormone (LH) data. Based on the medical literature, we know that LH levels for women in their child-bearing years vary between 0.61 and 56.6 IU/L. Values vary during the menstrual period. If values are lower than normal, this can be an indication that the woman suffers from malnutrition, anorexia, stress or a pituitary disorder. If the values are higher, this is an indication that the woman has gone through menopause.

We're going to use the LH data presented earlier in this chapter to make a decision whether the woman has a healthy range of values for a woman in her child-bearing years by testing the null-hypothesis that the mean LH level in this woman is the same as the mean of LH levels in healthy non-menopausal women.

First we specify the null-hypothesis. Suppose we know that the mean LH level in this woman should be equal to 2.54, given her age and given the timing of her menstrual cycle. Thus our null-hypothesis is that the mean LH in our particular woman is equal to 2.54:

$$H_0 : \mu = 2.54 \tag{2.11}$$

Next, we look at our sample mean and see whether this is a likely or unlikely value to find under this null-hypothesis. The sample mean is 2.40. To know whether this is a likely value to find, we have to know the standard error of the sampling distribution, and we can estimate this by using the sample variance. The sample variance $s^2 = 0.3042553$ and we had 48 measures, so we estimate the standard error as $\hat{\sigma y} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{0.3042553}{48}} = 0.08$. We then apply

standardisation to get a t-value:

$$t = \frac{2.40 - 2.54}{0.08} = -1.75 \tag{2.12}$$

Next, we look up in a table whether this t-value is extreme enough to be considered unlikely under the null-hypothesis. In Table 2.3, we see that for 47 degrees of freedom, the critical value for the 0.025 quantile equals -2.01. For the 0.975 quantile it is 2.01. Our observed t-value of -1.75 lies within this range. This means that a sample mean of 2.40 is likely to be found when the population mean is 2.54, so we do not reject the null-hypothesis. We conclude that the LH levels are healthy for a woman her age.

We can do the null-hypothesis testing also with a computer. Let's analyse the data in R and do the computations with the following code. First we load the LH data:

```
data(lh)
```

Next, we test whether the population mean could be 2.54:

```
t.test(lh, mu = 2.54)

##
##  One Sample t-test
##
## data:  lh
## t = -1.7584, df = 47, p-value = 0.08518
## alternative hypothesis: true mean is not equal to 2.54
## 95 percent confidence interval:
##  2.239834 2.560166
## sample estimates:
## mean of x
##       2.4
```

In the output we see that the t-value is equal to -1.7584, similar to our -1.75. We see that the number of degrees of freedom is 47 $(n-1)$ and that the p-value equals 0.08518. This p-value is larger than 0.05, so we do *not* reject the null-hypothesis that the mean LH level in this woman equals 2.54. Her LH level is healthy.

2.14 One-sided versus two-sided testing

In the previous section, we tested a null-hypothesis in order to find evidence that an observed sample mean was either too large or too small to result from random sampling. For example, in the previous section we saw that the observed LH levels were not too low and we did not reject the null-hypothesis. But had

the LH levels been too high or too low, then we would have rejected the null-hypothesis.

In the reasoning that we followed, there were actually two hypotheses: the null-hypothesis that the population mean was exactly 2.54, and the *alternative hypothesis* that the population is not exactly 2.54:

$$H_0 : \mu = 2.54 \tag{2.13}$$
$$H_A : \mu \neq 2.54 \tag{2.14}$$

This kind of null-hypothesis testing is called *two-sided* or *two-tailed* testing: we look at two critical values, and if the computed t-score is outside this range (i.e., somewhere in the two tails of the distribution), we reject the null-hypothesis.

The alternative to two-sided testing is *one-sided* or *one-tailed* testing. Sometimes before an analysis you already have an idea of what direction the data will go. For instance, imagine a zoo where they have held elephants for years. These elephants always were of Tanzanian origin, with a mean height of 3.38. Lately however, the manager observes that the opening that connects the indoor housing with the outdoor housing gets increasingly damaged. Since the zoo recently acquired 4 new elephants of South-African origin, the manager wonders whether South-African elephants are on average taller than the Tanzanian elephants. To figure out whether South-African elephants are on average taller than the Tanzanian average of 3.38 or not, the manager decides to apply null-hypothesis testing. She has two hypotheses: null-hypothesis H_0 and *alternative hypothesis* H_A:

$$H_0 : \mu_{SA} = 3.38 \tag{2.15}$$
$$H_A : \mu_{SA} > 3.38 \tag{2.16}$$

This set of hypotheses leaves out one option: the South-African mean might be lower than the Tanzanian one. Therefore, one often writes the set of hypotheses like this:

$$H_0 : \mu_{SA} \leq 3.38 \tag{2.17}$$
$$H_A : \mu_{SA} > 3.38 \tag{2.18}$$

She next tests the null-hypothesis, more specifically the one where $\mu_{SA} = 3.38$. From the damaged doorway she expects the sample mean to be higher than 3.38, but is it high enough to serve as evidence that the population mean is also higher than 3.38? She decides that when the sample mean is in the rejection zone in the right tail of the sampling distribution, then she will decide that the null-hypothesis is not true, but that the alternative hypothesis must be true. This is illustrated in Figure 2.16.

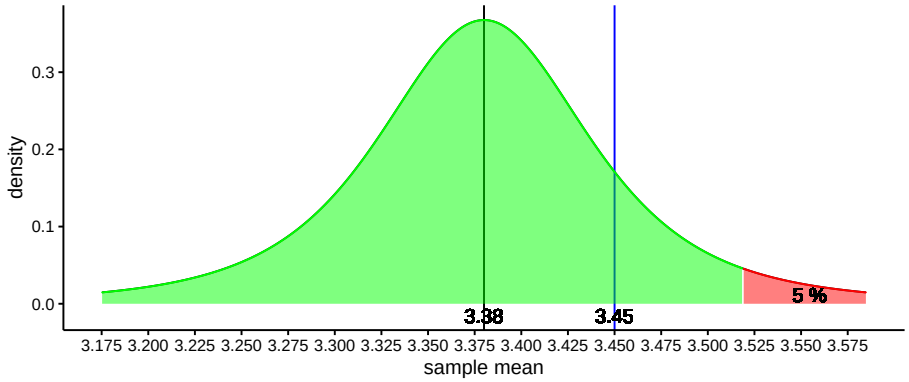


Figure 2.16: The sampling distribution under the null-hypothesis that the South-African population mean is 3.38. For one-tailed testing, the red area denotes the range of values for which the null-hypothesis is rejected (rejection region), the green area denotes the range of values for which the null-hypothesis is not rejected (acceptance region).

It shows the sampling distribution if we happen to have 4 new South-African elephants, with a sample mean of 3.45 and a standard error of 0.059. In red, we see the rejection region: if the sample mean happens to be in that zone we decide to reject the null-hypothesis. Similar to two-tailed testing, we decide that an area of 5% is small enough to suggest that the null-hypothesis is not true. Note that in two-tailed testing, this area of 5% was divided equally into the upper tail and the lower tail of the distribution, but with one-tailed testing we put it all in the tail where we expect to find the sample mean based on a theory or a hunch.

In this sampling distribution, based on 3 degrees of freedom, we see that the sample mean is not in the red zone – the rejection region – therefore we do not reject the null-hypothesis. We conclude that based on this random sample of 4 elephants, there is no evidence to suggest that South-African elephants are on average taller than Tanzanian elephants.

The same procedure can be done with standardisation. We compute the t-statistic as

$$t = \frac{3.45 - 3.38}{0.059} = 1.19 \tag{2.19}$$

In Table 2.3 we have to look up where the red zone starts: that is for the 0.95 quantile, because below that value lies 95% (green zone) and above it 5% (the red zone). We see that the 95th percentile for a t-distribution with 3 degrees of freedom is equal to 2. Our t-value 1.19 is less than that, so that we do not reject the null-hypothesis.

A third way is to compute a one-tailed p-value. This is illustrated in Figure 2.17. The one-tailed p-value for a t-statistic of 1.19 and 3 degrees of freedom

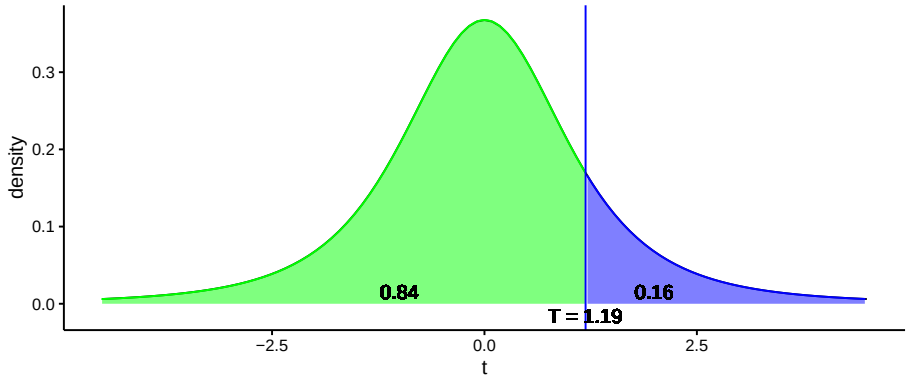


Figure 2.17: The sampling distribution under the null-hypothesis that the South-African population mean is 3.38. For one-tailed testing, the red area denotes the range of values for which the null-hypothesis is rejected (rejection region), the green area denotes the range of values for which the null-hypothesis is not rejected (acceptance region).

turns out to be 0.16. That is the proportion of the t-distribution that is blue. That means that if the null-hypothesis is true, you will find a t-value of 1.19 or larger in 16% of the cases. Because this proportion is more than 5%, we do not reject the null-hypothesis.

2.15 One-tailed testing applied to LH levels

As we have seen, LH levels that are too high are indicative of menopause, a normal transition for women. However, LH levels that are too low are indicative of an illness or malnutrition. In that case, it is important that the source of this malnutrition or the specific illness is diagnosed. You could therefore say that if LH levels are too low, a red flag should be put up, whereas if the LH levels are normal or higher, then there is usually no reason to worry.

LH levels can therefore be used to construct a diagnostic red flag decision system. If normal or high, then nothing happens, if too low, then something should be done. We could formulate these two alternative states of reality as two hypotheses:

$$H_0 : \mu_{LH} \geq 2.54 \tag{2.20}$$
$$H_A : \mu_{LH} < 2.54 \tag{2.21}$$

We decide beforehand that if a t-value is too far out in the left tail of the distribution, the LH levels are too low. We again use 5% of the area of the t-distribution. This decision process is illustrated in Figure 2.18 where we see a critical t-value of -1.68 when we we have 47 degrees of freedom (see Table 2.3).

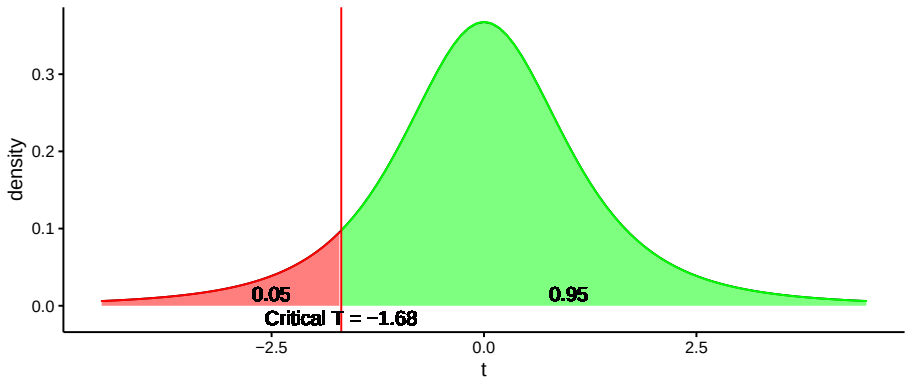


Figure 2.18: One-tailed decision process for deciding whether the average LH level in a woman is too low.

We calculate our t-value and find -1.75, see section 2.13. We see that this t-value is smaller than the critical value -1.68, so it is in the red rejection area. This is the area that we use for the rejection of the null-hypothesis, so based on these data we decide that the mean LH level in this woman is abnormally low.

Importantly, note that when we applied two-tailed hypothesis testing, we decided to *not* reject the null-hypothesis, whereas here with one-tailed testing, we decide to reject the null-hypothesis. All based on the same data, and the same null-hypothesis. The difference lies in the choice of the alternative hypothesis. When doing one-tailed testing, we put all of the critical region in only one tail of the t-distribution. This way, it becomes easier to reject a null-hypothesis, if the mean LH level is indeed lower than normal. However, it could also be easier to make a mistake: if the mean LH level is in fact normal, we could make a mistake in thinking that the sample mean is deviant, where it is actually not. Making mistakes in inference is the topic of the next section.

It is generally advised to use two-tailed testing rather than one-tailed testing. The reason is that in hypothesis testing, it is always the null-hypothesis that is being used as the starting point: what would the sample means (or their standardised versions: t-scores) look like if the null-hypothesis is true? Based on a certain null-hypothesis, say population mean μ equals 2.54, sample means could be as likely higher or lower than the population mean (since the sampling distribution is symmetrical). Even if you suspect that μ is actually lower, based on a very good theory, you would help yourself too much to falsify the null-hypothesis by putting the rejection area only in the left tail of the distribution. And what do you actually do if you find a sample mean that is in the far end of the right tail? Do you still accept the null-hypothesis? That would not make much sense. It is therefore better to just stick to the null-hypothesis, and see whether the sample mean is far enough removed to reject the null-hypothesis. If the sample mean is in the anticipated tail of the distribution, that supports the theory you had, and if the sample mean is in the opposite tail, it does not

support the theory you had.

Compare one-tailed and two-tailed testing in R using the LH data. By default, R applies two-tailed testing. R gives the following output:

```
t.test(lh, mu = 2.54)

##
##  One Sample t-test
##
## data:  lh
## t = -1.7584, df = 47, p-value = 0.08518
## alternative hypothesis: true mean is not equal to 2.54
## 95 percent confidence interval:
##  2.239834 2.560166
## sample estimates:
## mean of x
##       2.4
```

If you want one-tailed testing, where you expect that the mean LH level is lower than 2.54, you do that in the following manner[3]:

```
t.test(lh, mu = 2.54, alternative = "less")

##
##  One Sample t-test
##
## data:  lh
## t = -1.7584, df = 47, p-value = 0.04259
## alternative hypothesis: true mean is less than 2.54
## 95 percent confidence interval:
##      -Inf 2.533589
## sample estimates:
## mean of x
##       2.4
```

When you compare the p-values, you see that the p-value using one-tailed testing is half the size of the p-value using two-tailed testing (0.04 vs 0.08). Based on the previous sections, you should know why the p-value is halved! In the second output, using a critical p-value of 5% you would reject the null-hypothesis, whereas in the first output, you would not reject the null-hypothesis. Using one-tailed testing could lead to a big mistake: thinking that the sample mean is deviant enough to reject the null-hypothesis, while the null-hypothesis is actually true. We delve deeper into such mistakes in the next section.

[3]If you expect that the LH level will higher than 2.54, you use "greater" instead of "less".

2.16 Type I and type II errors

In the preceding sections, we have used the value of 5% a lot of times. We deemed that this was a fairly low probability, that allows us to take the decision to reject the null-hypothesis. We looked at the distribution of sample means, given that there was a certain population mean, and we looked at how often we can expect a sample mean that is smaller or larger than certain critical values. These critical values were based on 5% of the area of the sampling distribution. With two-tailed testing, this 5% was divided over the two extreme tails of the sampling distribution, and with one-tailed testing, this 5% rejection area was put in the tail end where we expected the population to be according to the alternative hypothesis (based on theory).

In this null-hypothesis testing procedure there is always the risk that we take the wrong decision. Let's return to our elephant example where we had the null-hypothesis that the population mean for South-African elephants equals 3.38. The alternative two-sided hypothesis was that the population mean was *not* equal to 3.38. After calculating the standard error, we calculated the t-score. We said that we reject the null-hypothesis when the obtained t-score was somewhere in the extreme ends of the tail: more specifically, in the rejection area that made up 5% of the area of the t-distribution. That means that if the null-hypothesis is true, there is a 5% probability that we find such a t-score. In that case we reject the null-hypothesis. But that could be the wrong decision: if the null-hypothesis is true it will happen in 5% of the cases that a t-score will be in the 5% rejection region. We then reject the null-hypothesis while it is actually true! Such a mistake is called a Type I error. In this case, type I error rate is 5%. It is a conditional probability. Conditional probabilities are probabilities that start from some given information. In this case, the given information is that the null-hypothesis is true: *given* that the null-hypothesis is true, it is the probability that we reject the null-hypothesis. Because we do not like to make mistakes, we want to have the probability of a mistake as low as possible.

In the social and behavioural sciences, one thinks that a probability of 5% is low enough to take the risk of making the wrong decision. As stated earlier, in quantum mechanics one is even more careful, using a probability of 0.000057 %. So why don't we also use a much lower probability of making a type I error? The answer is that we do not want to make another type of mistake: a type II error. A type II error is the mistake that we make when we do *not* reject the null-hypothesis, while it is not true. Taking the example of the elephants again, suppose that the population mean is *not* equal to 3.38, but the t-score is not in the rejection area, so we believe that the population mean is 3.38. This is then the wrong decision. The type of mistake we then make is a type II error.

Let's take this example further. Suppose we have a two-tailed decision process, where we compare two hypotheses about South-African elephants: either their mean height is equal to 3.38 (H_0), or it is not (H_A). We compute the t-statistic and determine the critical values based on 5% area in the tails of the t-distribution. This means that we allow ourselves to make a mistake in 5% of

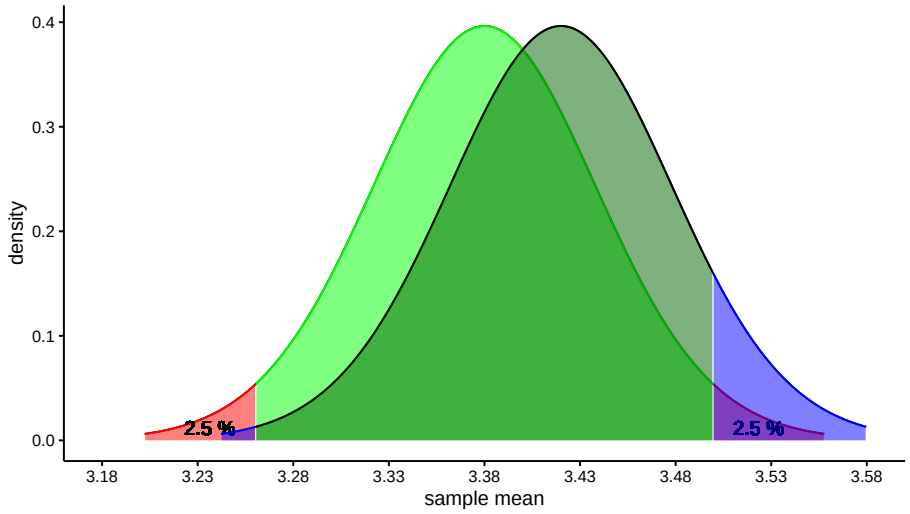


Figure 2.19: Two sampling distributions, one for a population mean of 3.38 (null-hypothesis) and one for a population mean of 3.42 (alternative hypothesis). The red areas represent the probability of a type I error, the dark green area the probability of a type II error. The blue area represents the probability of making the (correct) decision that the null-hypothesis is not true when it is indeed not true.

the cases: the probability that we find a t-score in one of the 2.5% tails equals $2.5\% + 2.5\% = 5\%$. This is the probability of a type I error. Note that we chose this value deliberately. This 5% we call α ('alpha'): it is the relative frequency we allow ourselves to make a type I error. We say then that our α is fixed to 0.05, or 5%. This means that if the null-hypothesis is true, the probability that the t-statistic will be in in the tails will be 5%.

Then what is the probability of a type II error? A type II error is based on the premise that the alternative hypothesis is true. That alternative hypothesis states that the population mean is *not* equal to 3.38. Given that, what is the probability that we do not reject the null-hypothesis?

This is impossible to compute, because the alternative hypothesis is very vaguely stated: it could be anything, as long as it is not 3.38. Let's make it a bit easier and state that the alternative hypothesis states that the population mean equals 3.42.

$$H_0 : \mu_{SA} = 3.38 \tag{2.22}$$
$$H_A : \mu_{SA} = 3.42 \tag{2.23}$$

If the population mean height is equal to 3.42, what would sample means look like? That's easy, that is the sampling distribution of the sample mean.

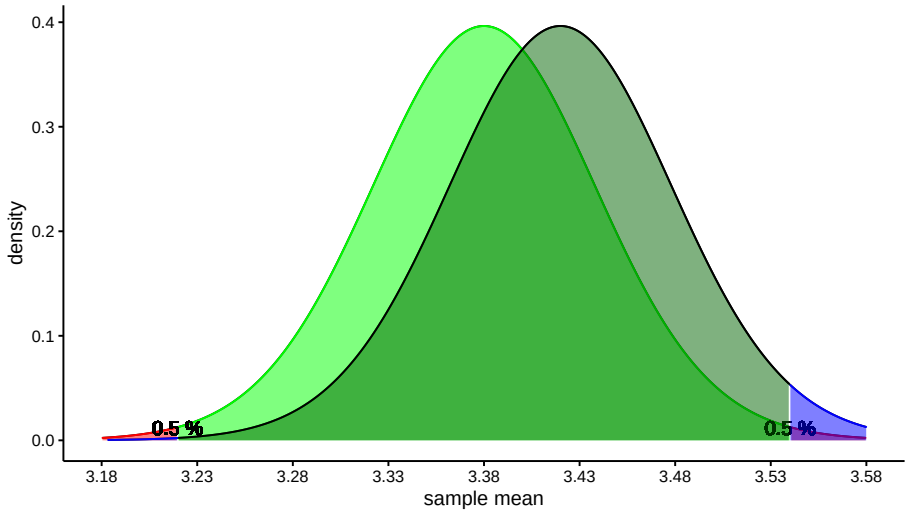


Figure 2.20: Two sampling distributions, one for a population mean of 3.38 (null-hypothesis) and one for a population mean of 3.42 (alternative hypothesis). The red areas represent the probability of a type I error, the dark green area the probability of a type II error. The blue area represents the probability of making the (correct) decision that the null-hypothesis is not true when it is indeed not true.

The mean of that sampling distribution would be 3.42. This is illustrated in Figure 2.19. The left curve is the sampling distribution for a population mean of 3.38. The red area represents the probability of a type I error. The right curve is the sampling distribution for a population mean of 3.42. The blue area represents the probability of rejecting the null-hypothesis. This is because if the sample mean is smaller than 3.260336 or larger than 3.499664, the sample mean is in the rejection area of the null-hypothesis testing and the null-hypothesis will therefore be rejected. The probability of this happening given that the *alternative* hypothesis is true ($H_A = 3.42$), is represented by the area under that curve: the blue area. If we determine the two blue areas in Figure 2.19, we end up with $0.004 + 0.097 = 0.101$. This is the probability of rejecting the null-hypothesis while it is not true, so this is no mistake at all. We would make a mistake when the alternative hypothesis is true, and we would *not* reject the null-hypothesis. This is represented by the dark green area. That area is equal to 1 minus the blue area: $1 - 0.101 = 0.899$.

When we have to make a definite decision about a population mean, the null-hypothesis framework can be used for that. Usually we don't want to make type I error mistakes, so we pick a low probability like 5% for the tails of the sampling distribution under the H_0. This value is called α: if the null-hypothesis is true, we don't want to reject it, so we allow this to happen in only 5% of the cases. One chooses α before collecting the data. You have to be careful with this

choice of α though because it directly affects the probability of making a type II error. This probability is denoted by β ('beta'): how often does it happen that if the alternative hypothesis is true, we do not reject the null-hypothesis. This relationship is illustrated in Figure 2.20. There, an α of 1% is chosen, using both tails (a two-tailed null-hypothesis test). You immediately see that the blue areas have also become smaller, and that by consequence the dark green area becomes larger: the probability of a type II error.

Thus, the α should be chosen wisely: if it is too large, you run a high risk of a type I error. But if it is too low, you run a high risk of a type II error. Let's think about this in the context of our luteinising hormone problem.

We saw that if the LH level is not normal, this is an indication of malnutrition or a disease and the patient should have further checks to see what the problem is. But if the LH level is normal or above, there is no disease and no further checks are required. Again we take the null-hypothesis that the mean LH level in this woman equals 2.54. What would be a type I error this case, and what would be type II error?

The type I error is the mistake of rejecting the null-hypothesis while it is in fact true. Thus, the woman's mean LH level is 2.54, but by coincidence, the mean of the 48 measurements that we have turns out to be in the rejection area of the sampling distribution. If this happens we make the mistake that we do a lot of tests with this woman to find out what's wrong with her, while in fact she is perfectly healthy! How bad would such a mistake be? It would certainly lead to extra costs, but also a lot of the woman's time. She would also probably start worrying that something is wrong with her. So we definitely don't want this to happen. We can minimize the risk of a type I error by choosing a low α.

The type II error is the mistake of *not* rejecting the null-hypothesis while it is in fact not true. Thus, the woman's mean LH level is lower than 2.54, but by coincidence, the sample mean of the 48 measurements that we have turns out to be in the acceptance area of the sampling distribution. This means that the woman's LH level does not seem to be abnormal, and the woman is sent home. How bad would such a mistake be? Well, pretty bad because the woman's hormone level is not normal, but everybody thinks that she is OK. She could be very ill but nothing is found in further tests, because there are no further tests. So we definitely don't want this to happen. We can minimize the risk of a type II error by choosing a higher α.

So here we have a conflict, and we have to make a balanced choice for α: too low we run the risk of type II errors, too high we run the risk of a type I error. Then you have to decide what is worse: a type I mistake or a type II mistake. In this case, you could say that sending the woman home while she is ill, is worse than spending money on tests that are actually not needed. Then you would choose a rather high α, say 10%. That means that if you have several women who are in fact healthy, 10% of them would receive extra testing. This is a fairly high percentage, but you are more sure that women with an illness will be detected and receive proper care.

But if you think it is most important that you don't spend too much money and that you don't want women to start worrying when it is not needed, you

can pick a low α like 1%: then when you have a lot of healthy women, only 1% of them will receive unnecessary testing.

Overview

- **Type I error**: the mistake of rejecting the null-hypothesis, while it is true
- **Type II error**: the mistake of not rejecting the null-hypothesis, while it is not true
- α: the relative frequency we allow ourselves to make a type I error
- β: the relative frequency of making a type II error

Chapter 3

Inference about a proportion

3.1 Sampling distribution of the sample proportion

So far, we focused on inference about a population mean: starting from a sample mean, what can we infer about the population mean? However, there are also other sample statistics we could focus on. We briefly touched on the variance in the sample and what it tells us about the population variance. In this section, we focus on inference regarding a proportion.

Let's go back to the example of the elephants in the zoo, and that the manager saw a damaged doorway. This is most likely caused by elephants that are taller than a certain height, making their heads bump the doorway when moving from one space to the other. Let's suppose the height of the doorway is 3.40 m and that the manager observes that of the 4 elephants in the zoo, 3 bump their head when passing the doorway. Suppose that the 4 elephants are randomly sampled from the entire population of elephants worldwide. What could we say based on these observations about the proportion of elephants worldwide that are taller than 3.40 m?

Let's again start from the population. Let's do the thought experiment that the population proportion of elephants taller than 3.40 m equals 0.6: 60% of all the elephants in the world are taller than 3.40 m. Let's randomly pick 4 elephants from this population. We might get 2 tall elephants and 2 less tall elephants. This means we get a sample proportion of $\frac{2}{4} = 0.5$. If we do this sampling a lot of times, we obtain the *sampling distribution of the sample proportion*. It is shown in Figure 3.1. It is a discrete (non-continuous) distribution that is clearly not a normal distribution. But, as we know from the Central Limit Theorem (Chapter 2), it will become a normal distribution when sample size increases.

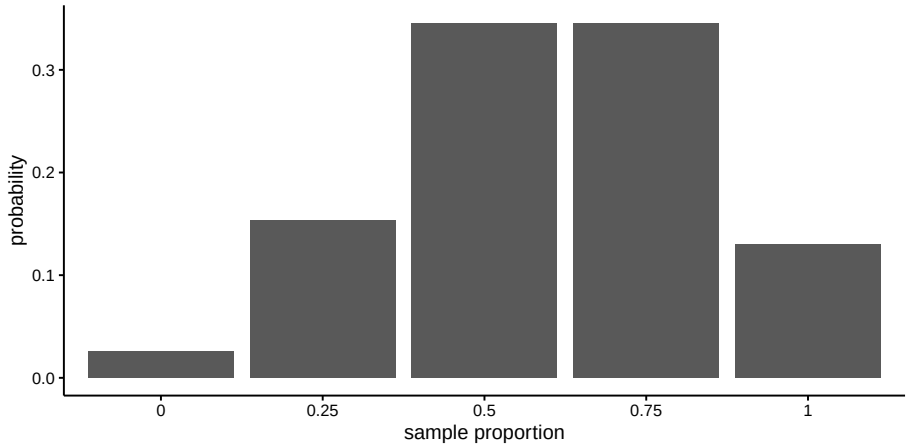


Figure 3.1: Sampling distribution of the sample proportion, when the population proportion is 0.60

Actually, the sampling distribution that we see in 3.1 is based on the *binomial distribution*. Using the binomial distribution, we can calculate the probabilities of getting various sample proportions in a straightforward manner, without relying on the normal distribution.

3.2 The binomial distribution

The binomial distribution gives us the probability of obtaining a certain number of elements, given how many elements there are in total and the population probability. In our case, the binomial distribution gives us the probability of having exactly 2 elephants taller than 3.40 m, given that there are 4 elephants in our sample and the population proportion equals 0.6. Let's go through the reasoning step by step.

The proportion of tall elephants in the population is $p = 0.6$. The sample size equals $n = 4$. Let's begin with randomly picking the first elephant: what's the probability that we select an elephant that is taller than 3.40 m? Well, that probability is equal to the proportion of 0.6. Next, what is the probability that the second elephant is taller than 3.40? Again, this is equal to 0.6.

Now something more complicated: what is the probability that both the first *and* the second elephant are taller than 3.40? This is equal to $0.6 \times 0.6 = 0.36$. What is the probability that *all* 4 elephants are taller than 3.40 m? That is equal to $0.6 \times 0.6 \times 0.6 \times 0.6 = 0.60^4 = 0.1296$. The probability that all 4 elephants are shorter than 3.40 m is equal to $(1 - 0.6)^4 = 0.4^4 = 0.0256$.

The probability for a mix of 2 tall elephants and 2 shorter elephants is more difficult to compute. You might remember from high school that it involves *combinations*. For example, the probability that the first 2 elephants are taller

than 3.40, and the last 2 elephants shorter, is equal to $0.6^2 \times (1-0.6)^2 = 0.0576$, but there are many other ways in which we can find 2 tall elephants and 2 shorter elephants when we randomly and sequentially pick 4 elephants. There are in fact 6 different ways of randomly selecting 4 elephants where only 2 are tall. When we use A for a tall elephant and B for a short elephant, the 6 possible orderings are in fact: AABB, BBAA, ABAB, BABA, ABBA, and BAAB.

This number of *permutations* is calculated using the *binomial coefficient*:

$$\binom{4}{2} = \frac{4!}{2!2!} = 6 \tag{3.1}$$

This number $\binom{4}{2}$ is called the binomial coefficient. It can be calculated using *factorials*: the exclamation mark ! stands for factorial. For instance, 5! ('five factorial') means $5 \times 4 \times 3 \times 2 \times 1$.

In its general form, the binomial coefficient looks like:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{3.2}$$

So suppose sample size n is equal to 4 and r equal to 2 (the number of tall elephants in the sample), we get:

$$\binom{4}{2} = \frac{4!}{2!(n-r)!} = \frac{4!}{2!2!} = \frac{4 \times 3 \times 2 \times 1}{2 \times 1 \times 2 \times 1} = 6 \tag{3.3}$$

Going back to the elephant example, there are $\binom{4}{2} = 6$ possible ways of getting 2 tall elephants and 2 short elephants when we sequentially pick 4 elephants. Each of these possibilities has a probability of $0.6^2 \times (1-0.6)^2 = 0.0576$. This is explained in Table 3.1. For instance, the probability of getting the ordering ABAB, is equal to the multiplication of the respective probabilities: $0.6 \times 0.4 \times 0.6 \times 0.4$. In the table you can see that the probability for any ordering is always 0.0576. Since any ordering will qualify as obtaining 2 tall elephants from a total of 4, we can sum these probabilities: the probability of getting the ordering AABB *or* BBAA *or* ABAB *or* BABA *or* ABBA *or* BAAB, is equal to $0.0576 + 0.0576 + 0.0576 + 0.0576 + 0.0576 + 0.0576 = 6 \times 0.0576 = 0.3456$. Here 6 is the number of permutations, calculated as the binomial coefficient $\binom{4}{2}$. We could therefore in general compute the probability of having 2 tall elephants in a sample of 4 as

$$p(\#A = 2 | n = 4, p = 0.6) = \binom{4}{2} \times 0.6^2 \times (1-0.6)^2 = 6 \times 0.0576 = 0.3456 \tag{3.4}$$

The probability of ending up with 2 tall elephants in a sample of 4 elephants, in any order, and where the proportion of tall elephants in the population is 0.6, is therefore equal to 0.3456.

In the more general case, if you have a population with a proportion p of As, a sample size of n, and you want to know the probability of finding r instances of A in your sample, it can be computed with the formula

$$p(\#A = r|n, p) = \binom{n}{r} \times p^r \times (1-p)^{(n-r)} \tag{3.5}$$

For example, the probability of obtaining 3 tall elephants when the total number of elephants is 4, is $\binom{4}{3} \times 0.6^3 \times (1-0.6)^1 = 4 \times 0.216 \times 0.4 = 0.3456$.

When we calculate the probabilities of finding 0, 1, 2, 3, or 4 tall elephants in sample of 4 when the population proportion is 0.6, we obtain the *binomial distribution* that is plotted in Figure 3.2. It is exactly the same as the sampling distribution in Figure 3.1, except that we plot the number of tall elephants in the sample on the horizontal axis, instead of the proportion. This means that we can use the binomial distribution to describe the sampling distribution of the sample proportion. To get the proportions, we simply divide the number of tall elephants in our sample by the total number of elephants (n) and we get Figure 3.1.

Table 3.1: Four possible ways of selecting 2 tall elephants (A) and 2 short elephants (B), together with the probability for each selection.

ordering	computation of probability	probability
AABB	0.6 x 0.6 x 0.4 x 0.4	0.0576
ABAB	0.6 x 0.4 x 0.6 x 0.4	0.0576
ABBA	0.6 x 0.4 x 0.4 x 0.6	0.0576
BAAB	0.4 x 0.6 x 0.6 x 0.4	0.0576
BABA	0.4 x 0.6 x 0.4 x 0.6	0.0576
BBAA	0.4 x 0.4 x 0.6 x 0.6	0.0576

Overview

- **sampling distribution of the sample proportion**: the distribution of proportions that you get when you randomly pick new samples from a population and for each sample compute the proportion.
- **binomial distribution**: a discrete distribution showing the probabilities of finding a certain number of successes (r), given sample size n and population proportion p.
- **binominal coefficient**: a coefficient used to calculate binomial probabilities. It represents the number of ways in which you can find r instances in a sample of size n. It is calculated as $\binom{n}{r} = \frac{n!}{r!(n-r)!}$.

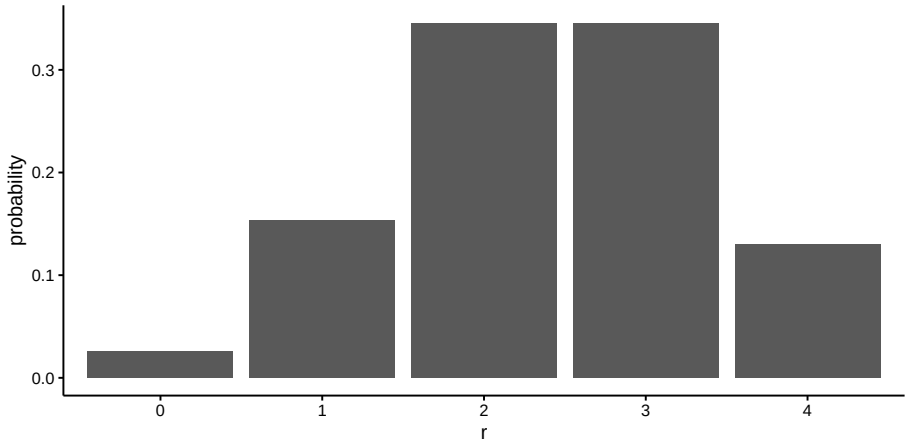


Figure 3.2: Binomial distribution with $N = 4$ and $p = 0.60$.

3.3 Confidence intervals

Based on what we know about the binomial distribution, we can perform inference on proportions. In Chapter 2 we saw that inference is very much based on the standard error (i.e., the standard deviation of the sampling distribution). We know from theory that the variance of the binomial distribution can be easily calculated as $n \times p \times (1 - p)$. Because we want to have the variance in proportions rather than in numbers, we have to divide this variance by n to get the variance of proportions: $\frac{n \times p \times (1-p)}{n} = p \times (1-p)$. Next, because the variance of a sampling distribution gets smaller with increasing n, we divide by n again, in a similar way as we did for the sampling distribution of the sample mean in Chapter 2. Taking the square root of this variance gives us the standard deviation of the sampling distribution (i.e., the standard error):

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{3.6}$$

This standard error makes it easy to construct confidence intervals. We know from the Central Limit Theorem that if n becomes infinitely large, the sampling distribution will become normal. When $n = 50$, the sampling distribution is already close to normal, as is shown in Figure 3.3. This fact, together with the standard error makes it easy to construct approximate confidence intervals.

Suppose that we had 50 elephants in our zoo, and the manager observed that 42 of them bump their head against the doorway. That is a sample proportion of $\frac{42}{50} = 0.84$. When we want to have a range of plausible values for the population proportion, we can construct a 95% confidence interval around this sample proportion. Because we know that for the standard normal distribution, 95% of the observations are between -1.96 and +1.96, we construct the 95% confidence

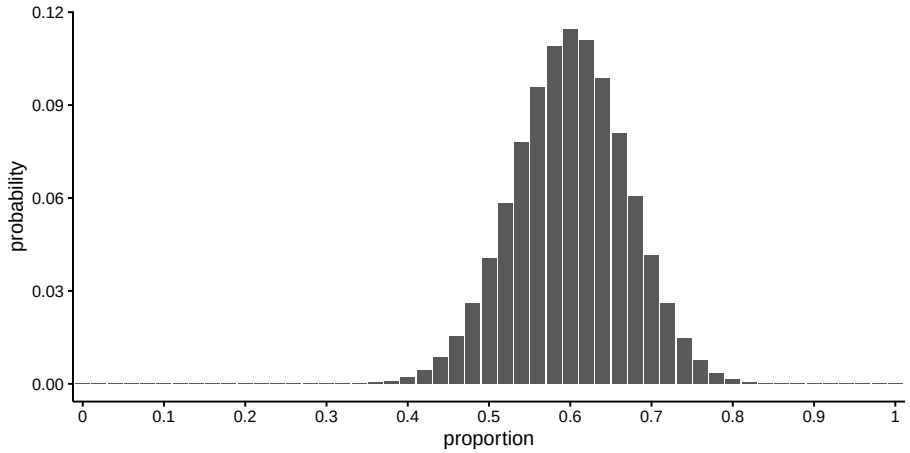


Figure 3.3: Sampling distribution with $N = 50$ and $p = 0.60$.

interval by multiplying 1.96 with the standard error, $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

However, since we do not know the population proportion p, we have to estimate it. From theory, we know that an unbiased estimator for the population proportion is the sample proportion: $\hat{p} = \frac{42}{50} = 0.84$. Our estimate for the standard error is then $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.0518459$.

If we use that value, we get the interval from $0.84 - 1.96 \times 0.0518459$ to $0.84 + 1.96 \times 0.0518459$: thus, our 95% confidence interval for the population proportion runs from 0.738382 to 0.941618.

3.4 Null-hypothesis concerning a proportion

Suppose that a researcher has measured all Tanzanian elephants and noted that a proportion of 0.60 was taller than 3.40 m. Suppose also that the manager in the zoo finds that 42 out of the 50 elephants bump their head and are therefore taller than 3.40. How can we know that the elephants could be a representative sample of Tanzanian elephants?

To answer this question with a yes or a no, we could apply the logic of null-hypothesis testing. Let the null-hypothesis be that the population proportion is equal to 0.60, and the alternative hypothesis that it is not equal to 0.60.

$$H_0 : p = 0.60 \tag{3.7}$$
$$H_A : p \neq 0.60 \tag{3.8}$$

Is the proportion of 0.84 that we observe in the sample (the zoo) a probable value to find if the proportion of all Tanzanian is equal to 0.60? If this is the

case, we do not reject the null-hypothesis, and believe that the zoo data could have been randomly selected from the Tanzanian population and are therefore representative. However, if the proportion of 0.84 is very unprobable given that the population proportion is 0.60, we reject the null-hypothesis and believe that the data are not representative.

With null-hypothesis testing we always have to fix our α first: the probability with which we are willing to accept a type I error. We feel it is really important that the sample is representative of the population, so we definitely do not want to make the mistake that we think the sample is representative (not rejecting the null-hypothesis) while it isn't (H_A is true). This would be a type II error (check this for yourself!). If we want to minimise the probability of a type II error (β), we have to pick a relatively high α (see Chapter 2), so let's choose our $\alpha = .10$.

Next, we have to choose a test statistic and determine critical values for it that go with an α of .10. Because we have a relatively large sample size of 50, we assume that the sampling distribution for a proportion of 0.60 is normal. From the standard normal distribution, we know that 90% ($1 - \alpha$!) of the values lie between -1.6448536 and 1.6448536 (see Table 2.3). If we therefore standardise our proportion, we have a measure that should show a standard normal distribution:

$$z_p = \frac{p_s - p_0}{sd} \tag{3.9}$$

where z_p is the z-score for a proportion, p_s is the sample proportion, p_0 is the population proportion assuming H_0, and sd is the standard deviation of the sampling distribution, which is the standard error. Note that we should take the standard error that we get when the null-hypothesis is true. We then get

$$z_p = \frac{0.84 - 0.6}{se} = \frac{0.24}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.24}{0.069282} = 3.4641016 \tag{3.10}$$

90% of the values in any normal distribution lie between ± 1.64 standard deviations away from the mean (see Table 2.3). Here we see a z-score that exceeds these critical values, and we therefore reject the null-hypothesis. We conclude that the proportion of tall elephants observed in the sample is larger than to be expected under the assumption that the population proportion is 0.6. We decide that the zoo data are not representative of the population data.

The decision process is illustrated in Figure 3.4.

3.5 Inference on proportions using R

Using the normal distribution is a nice trick when you have to do the calculations by hand. However, this approach is of course only valid when you have large sample sizes, so that you know that the shape of the normal distribution is a good approximation of the binomial distribution. In contrast, using the binomial

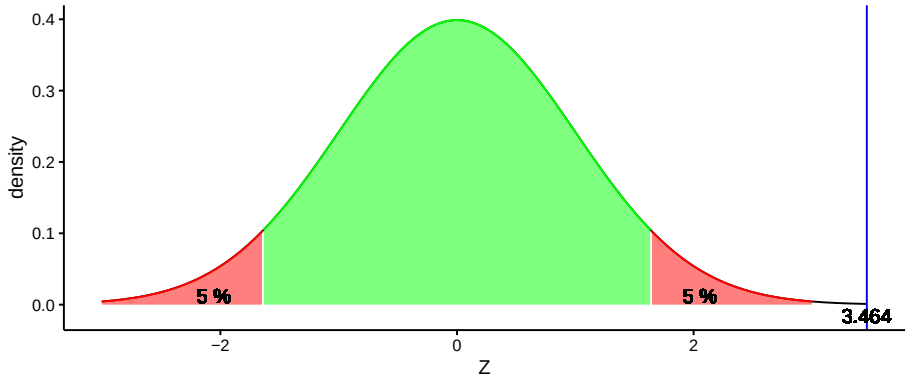


Figure 3.4: A normal distribution to test the null-hypothesis that the population proportion is 0.6. The blue line represents the z-score for our observed sample proportion of 0.84.

distribution always gives you the most exact answers. However it can be very tiresome to do all the computations by hand. In this section we discuss how to let R do the calculations for you.

Suppose we have a sample of 50 elephants, and we see that 42 of them bump their head against the doorway. What can we say about the population: what proportion of elephants in the entire population will bump their heads? In R, we use the `binom.test()` function to do inference on proportions. This function does all the calculations using the binomial distribution, so that the results are always trustworthy, even for small sample sizes. We state the number of oberved elephants that bump their head ($x = 42$), the sample size ($n = 50$), the kind of confidence interval (95%: conf.level = 0.95) and the proportion that we want to use for the null-hypothesis ($p = 0.6$):

```
binom.test(x = 42, n = 50, conf.level = 0.95, p = 0.6)

##
##  Exact binomial test
##
## data:  42 and 50
## number of successes = 42, number of trials = 50, p-value = 0.0004116
## alternative hypothesis: true probability of success is not equal to 0.6
## 95 percent confidence interval:
##  0.7088737 0.9282992
## sample estimates:
## probability of success
##                   0.84
```

The output shows the sample proportion: the probability of success is 0.84.

This is of course $\frac{42}{50}$. If we want to know what the population proportion is, we look at the 95% confidence interval that runs from 0.7088737 to 0.9282992. If you want to test the null-hypothesis that the population proportion is equal to 0.60, then we see that the p-value for that test is 0.0004116.

As said, the binomial test also works fine for small sample sizes. Let's go back to the very first example of this chapter: the zoo manager sees that of the 4 elephants they have, 3 bump their head and are therefore taller than 3.40 m. What does that tell us about the proportion of elephants worldwide that are taller than 3.40 m? If we assume that the 4 zoo elephants were randomly selected from the entire population of elephants, we can use the binomial distribution. In this case we type in R:

```
binom.test(x = 3, n = 4)

##
##  Exact binomial test
##
## data:  3 and 4
## number of successes = 3, number of trials = 4, p-value = 0.625
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1941204 0.9936905
## sample estimates:
## probability of success
##                   0.75
```

By default, `binom.test()` yields 95% confidence intervals, as can be seen in the output.[1] We see that the confidence interval for the population proportion runs from 0.1941204 to 0.9936905. Thus, based on this sample proportion of 0.75, we can see with some degree of confidence that the population proportion is somewhere between 0.19 and 0.99. That's of course not very informative, which makes sense considering we only observe 4 elephants.

[1] Note in the output that by default, `binom.test()` chooses the null-hypothesis that the population proportion is 0.5.

Chapter 4

Linear modelling: introduction

4.1 Dependent and independent variables

In the previous two chapters we discussed single variables. In Chapter 2 we discussed a numeric variable that had a certain mean, for instance we talked about the height of elephants. In Chapter 3 we talked about a dichotomous categorical variable: elephants being taller than 3.40 m or not, with a certain proportion of tall elephants. This chapter deals with the relationship between two variables, more specifically the relationship between two numeric variables.

In Chapter 1 we discussed the distinction between numeric, ordinal and categorical variables. In linear modelling, there is also another important distinction between variables: *dependent* and *independent* variables. Dependency of a variable is not really a property of a variable but it is the result of the data analyst's choice. Let's first think about relationships between two variables. Determining whether a variable is to be treated as independent or not, is often either a case of logic or a case of theory. When studying the relationship between the height of a mother and that of her child, the more logical it would be to see the height of the child *as dependent* on the height of the mother. This is because we assume that the genes are transferred from the mother to the child. The mother comes first, and the height of the child is partly the *result* of the mother's genes that were transmitted during fertilisation. The height of a child depends in part on the height of the mother. The variable that measures the result is usually taken as the *dependent* variable. The theoretical cause or antecedent is usually taken as the *independent* variable.

The dependent variable is often called the *response variable*. An independent variable is often called a *predictor variable* or simply *predictor*. Independent variables are also often called *explanatory* variables. We can explain a very tall child by the genes that it got from its very tall mother. The height of a child is then the response variable, and the height of the mother is the explanatory

variable. We can also predict the adult height of a child from the height of the mother.

The dependent variable is usually the most central variable. It is the variable that we'd like to understand better, or perhaps predict. The independent variable is usually an explanatory variable: it explains why some people have high values for the dependent variable and other people have low values. For instance, we'd like to know why some people are healthier than others. Health may then be our dependent variable. An explanatory variable might be age (older people tend to be less healthy), or perhaps occupation (being a dive instructor induces more health problems than being a university professor).

Sometimes we're interested to see whether we can predict a variable. For example, we might want to predict longevity. Age at death would then be our dependent variable and our independent (predictor) variables might concern lifestyle and genetic make-up.

Thus, we often see four types of relations:

- Variable A affects/influences another variable B.
- Variable A causes variable B.
- Variable A explains variable B.
- Variable A predicts variable B.

In all these four cases, variable A is the independent variable and variable B is the dependent variable.

Note that in general, dependent variables can be either numeric, ordinal, or categorical. Also independent variables can be numeric, ordinal, or categorical.

4.2 Linear equations

From secondary education you might remember linear equations. Suppose you have two quantities, X and Y, and there is a straight line that describes best their relationship. An example is given in Figure 4.1. We see that for every value of X, there is only one value of Y. Moreover, the larger the value of X, the larger the value of Y. If we look more closely, we see that for each increase of 1 unit in X, there is an increase of 2 units in Y. For instance, if $X = 1$, we see a Y-value of 2, and if $X = 2$ we see a Y-value of 4. So if we move from $X = 1$ to $X = 2$ (a step of one on the X-axis), we move from 2 to 4 on the Y-axis, which is an increase of 2 units. This increase of 2 units for every step of 1 unit in X is the same for all values of X and Y. For instance, if we move from 2 to 3 on the X-axis, we go from 4 to 6 on the Y-axis: an increase of again 2 units. This constant increase is typical for linear relationships. The increase in Y for every unit increase in X is called the *slope* of a straight line. In this figure, the slope is equal to 2.

The slope is one important characteristic of a straight line. The second important property of a straight line is the *intercept*. The intercept is the value

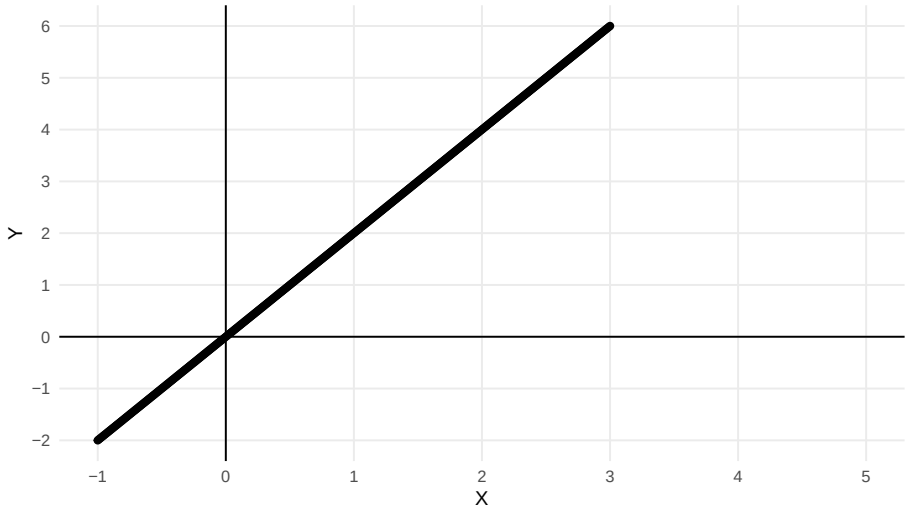


Figure 4.1: Straight line with intercept 0 and slope 2.

of Y, when $X = 0$. In Figure 4.1 we see that when $X = 0$, Y is 0, too. Therefore the intercept of this straight line is 0.

With the intercept and the slope, we completely describe this straight line: no other information is necessary. Such a straight line describes a *linear relationship* between X and Y. The linear relationship can be formalised using a linear equation. The general form of a linear equation for two variables X and Y is the following:

$$Y = \text{intercept} + \text{slope} \times X \tag{4.1}$$

For the linear relationship between X and Y in Figure 4.1 the linear equation is therefore

$$Y = 0 + 2X \tag{4.2}$$

which can be simplified to

$$Y = 2X \tag{4.3}$$

With this equation, we can find the Y-value for all values of X. For instance, if we want to know the Y-value for $X = 3.14$, then using the linear equation we know that $Y = 2 \times 3.14 = 6.28$. If we want to know the Y-value for $X = 49876.6$, we use the equation to obtain $Y = 2 \times 49876.6 = 99753.2$. In short, the linear equation is very helpful to quickly say what the Y-value is on the basis of the X-value, even if we don't have a graph of the relationship or if the graph does not extent to certain X-values.

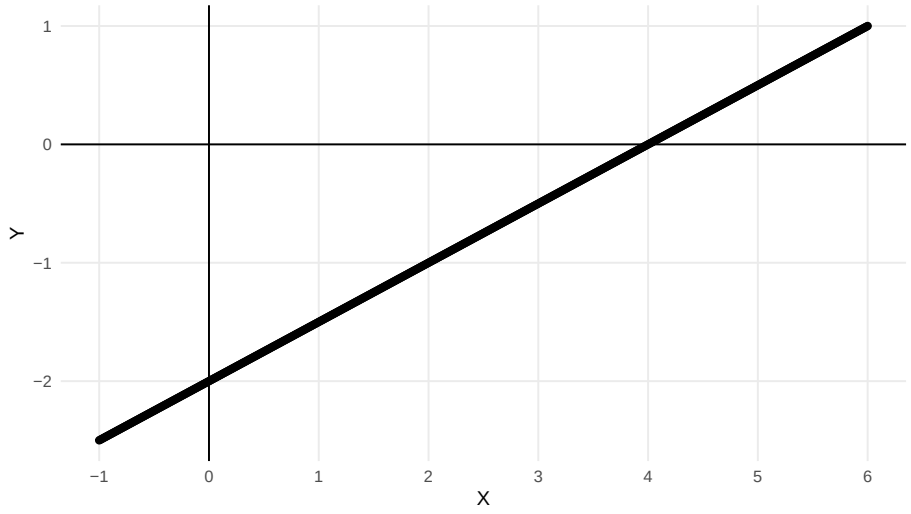


Figure 4.2: Straight line with intercept -2 and slope 0.5.

In the linear equation, we call Y the *dependent* variable, and X the *independent* variable. This is because the equation helps us determine or predict our value of Y on the basis of what we know about the value of X. When we graph the line that the equation represents, such as in Figure 4.1, the common way is to put the dependent variable on the vertical axis, and the independent variable on the horizontal axis.

Figure 4.2 shows a different linear relationship between X and Y. First we look at the slope: we see that for every unit increase in X (from 1 to 2, or from 4 to 5) we see an increase of 0.5 in Y. Therefore the slope is equal to 0.5. Second, we look at the intercept: we see that when $X = 0$, Y has the value -2. So the intercept is -2. Again, we can describe the linear relationship by a linear equation, which is now:

$$Y = -2 + 0.5X \tag{4.4}$$

Linear relationships can also be negative, see Figure 4.3. There, we see that if we move from 0 to 1, we see a *decrease* of 2 in Y (we move from $Y = -2$ to $Y = -4$), so -2 is our slope value. Because the slope is negative, we call the relationship between the two variables negative. Further, when $X = 0$, we see a Y-value of -2, and that is our intercept. The linear equation is therefore:

$$Y = -2 - 2X \tag{4.5}$$

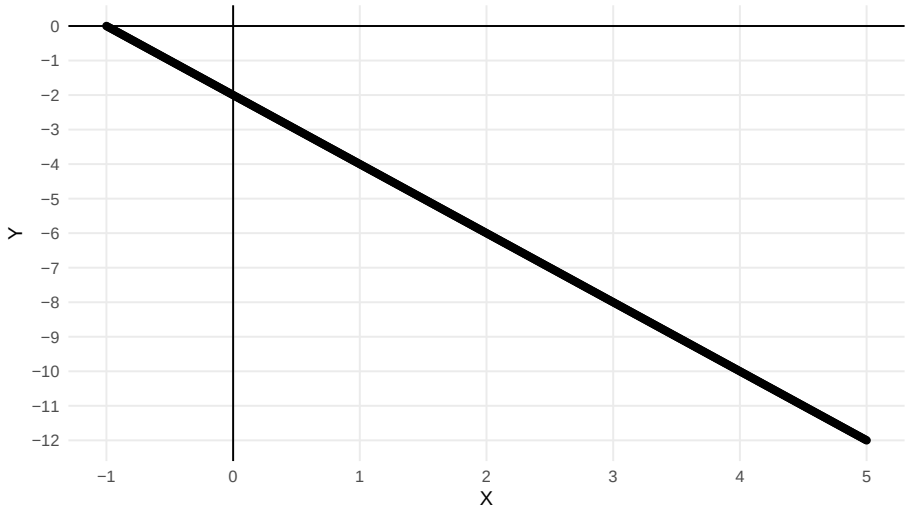


Figure 4.3: Straight line with intercept -2 and slope -2.

Overview

- **dependent variable**: the variable that we want to describe, understand, predict or explain. Usually denoted as Y.
- **independent variable**: the variable that we use in order to understand, predict or explain something. Usually denoted as X.
- **linear relationship**: two variables are said to be linearly related if their relationship can be described by a linear equation with an intercept and a slope.
- **intercept**: the value for Y (dependent variable) if $X = 0$ (independent variable).
- **slope**: the change in Y when we increase X by 1 unit.

4.3 Linear regression

In the previous section, we saw perfect linear relationships between quantities X and Y: for each X-value there was only one Y-value, and the values are all described by a straight line. Such relationships we hope to see in physics, but mostly see only in mathematics.

In social sciences we hardly ever see such perfectly linear relationships between quantities (variables). For instance, let us plot the relationship between

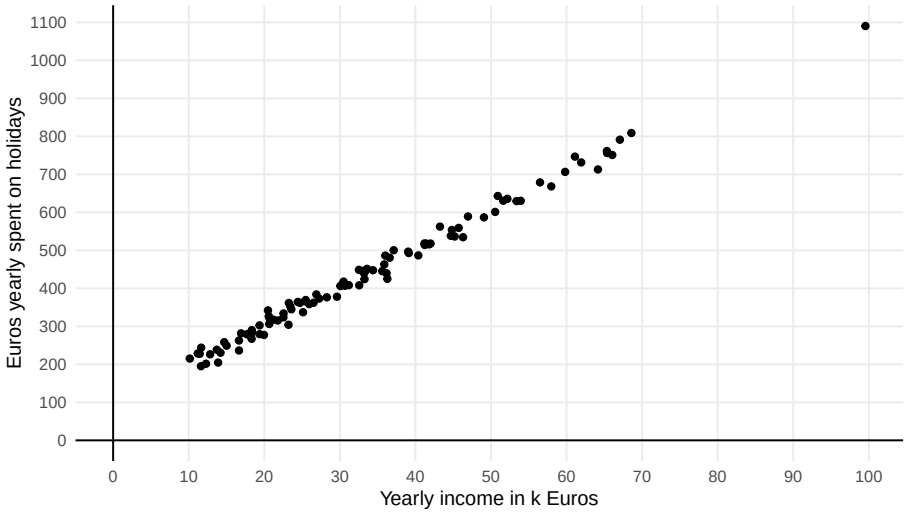


Figure 4.4: Data on holiday spending.

yearly income and the amount of Euros spent on holidays. Yearly income is measured in thousands of Euros (k Euros), and money yearly spent on holidays is measured in Euros. Let us regard money spent on holidays as our dependent variable and yearly income as our independent variable (we assume money needs to be saved before it can be spent). We therefore plot yearly income on the X-axis (horizontal axis) and holiday spendings on the Y-axis (vertical axis). Let's imagine we find the data from 100 women between 30 and 40 years of age that are plotted in Figure 4.4.

In the scatter plot, we see that one woman has a yearly income of 100,000 Euros, and that she spends almost 1100 Euros per year on holidays. We also see a couple of women who earn less, between 10,000 and 20,000 Euros a year, and they spend between 200 and 300 Euros per year on holiday.

The data obviously do not form a straight line. However, we tend to think that the relationship between yearly income and holiday spending is more or less linear: there is a general linear trend such that for every increase of 10,000 Euros in yearly income, there is an increase of about 100 Euros.

Let's plot such a straight line that represents that general trend, with a slope of 100 straight through the data points. The result is seen in Figure 4.5. We see that the line with a slope of 100 is a nice approximation of the relationship between yearly income and holiday spendings. We also see that the intercept of the line is 100.

Given the intercept and slope, the linear equation for the straight line approximating the relationship is

$$\texttt{HolidaySpendings} = 100 + 100 \times \texttt{YearlyIncome} \tag{4.6}$$

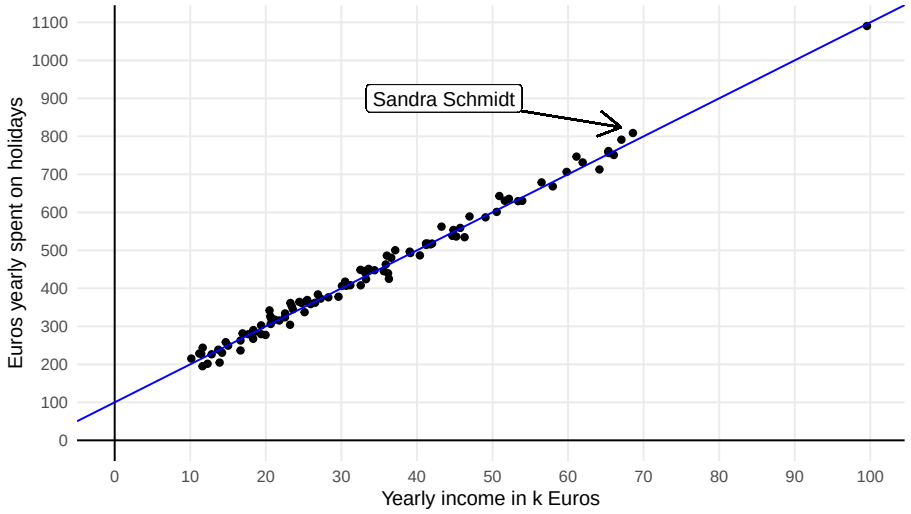


Figure 4.5: Data on holiday spending with an added straight line.

In summary, data on two variables may not show a perfect linear relationship, but in many cases, a perfect straight line can be a very reasonable approximation of the data. Another word for a reasonable approximation of the data is a *prediction model*. Finding such a straight line to approximate the data points is called *linear regression*. In this chapter we will see what method we can use to find a straight line. In linear regression we describe the behaviour of the dependent variable (the Y-variable on the vertical axis) on the basis of the independent variable (the X-value on the horizontal axis) using a linear equation. We say that *we regress variable Y on variable X*.

4.4 Residuals

Even though a straight line can be a good approximation of a data set consisting of two variables, it is hardly ever perfect: there are always discrepancies between what the straight line describes and what the data actually tell us.

For instance, in Figure 4.5, we see a woman, Sandra Schmidt, who makes 69 k Euros a year and who spends 809 Euros on holidays. According to the linear equation that describes the straight line, a woman that earns 69 k Euros a year would spend $100 + 100 \times 69 = 786$ Euros on holidays. The discrepancy between the actual amount spent and the amount prescribed by the linear equation equals $809 - 786 = 23$ Euros. This difference is rather small and the same holds for all the other women in this data set. Such discrepancies between the actual amount spent and the amount as prescribed or predicted by the straight line are called *residuals* or *errors*. The residual (or error) is the difference between a certain data point (the *actual* value) and what the linear equation predicts.

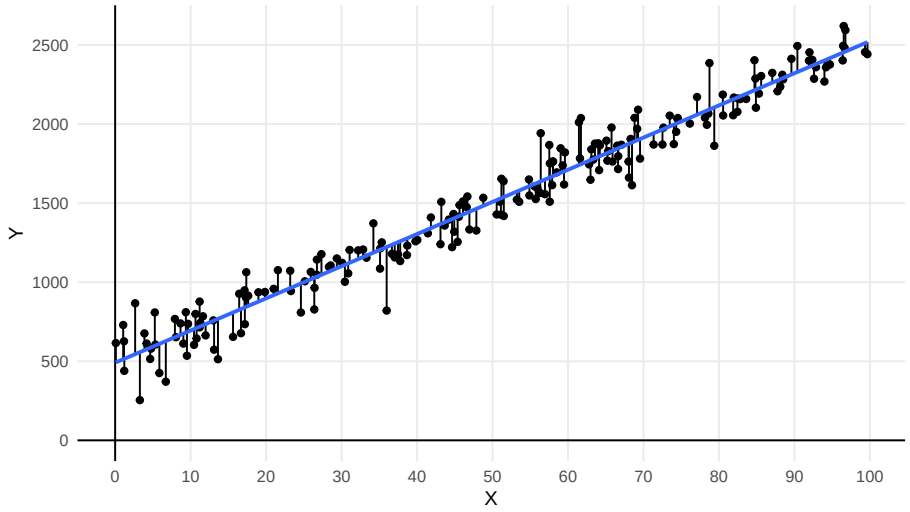


Figure 4.6: Data on variables X and Y with an added straight line.

Let us look at another fictitious data set where the residuals (errors) are a bit larger. Figure 4.6 shows the relationship between variables X and Y. The dots are the actual data points and the blue straight line is an approximation of the actual relationship. The residuals are also visualised: sometimes the observed Y-value is greater than the predicted Y-value (dots above the line) and sometimes the observed Y-value is smaller than the predicted Y-value (dots below the line). If we denote the ith predicted Y-value (predicted by the blue line) as $\widehat{Y}_i$ (pronounced as 'y-hat-i'), then we can define the residual or error as the discrepancy between the observed Y_i and the predicted $\widehat{Y}_i$:

$$e_i = Y_i - \widehat{Y}_i \tag{4.7}$$

where e_i stands for the error (residual) for the ith data point .

If we compute residual e_i for all Y-values in the data set, we can plot them using a histogram, as displayed in Figure 4.7. We see that the residuals are on average 0, and that the histogram resembles the shape of a normal distribution. We see that most of the residuals are around 0, and that means that most of the values Y-values are close to the line (where the predicted values are). We also see some large residuals but that there are not so many of these. Observing a more or less normal distribution of residuals happens often in research. Here, the residuals show a normal distribution with mean 0 and variance of 13336 (i.e., a standard deviation of 115).

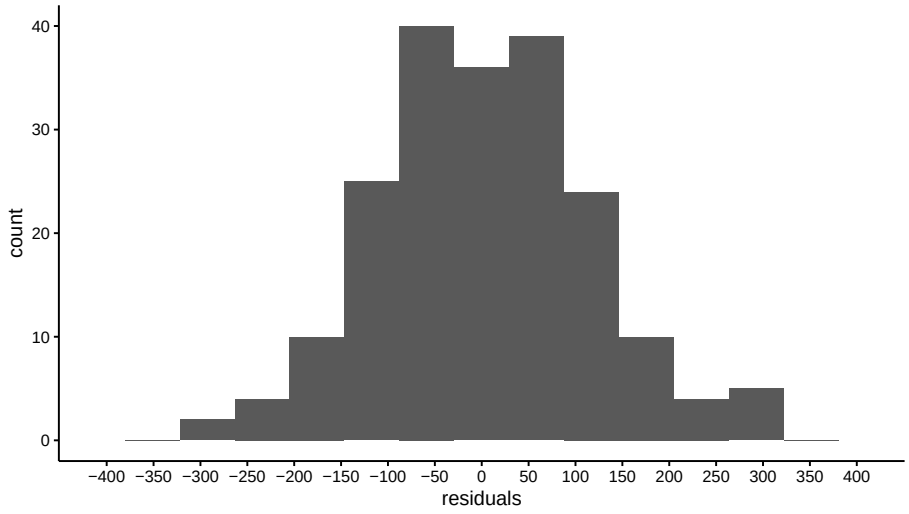


Figure 4.7: Histogram of the residuals (errors).

4.5 Least squares regression lines

You may ask yourself how to draw a straight line through the data points: How do you decide on the exact slope and the exact intercept? And what if you don't want to draw the data points and the straight line by hand? That can be quite cumbersome if you have more than 2000 data points to plot!

First, because we are lazy, we always use a computer to draw the data points and the line, that we call a *regression line*. Second, since we could draw many different straight lines through a scatter of points, we need a criterion to determine a nice combination of intercept and slope. With such a criterion we can then let the computer determine the regression line with its equation for us.

The criterion that we use in this chapter is called Least Squares, or Ordinary Least Squares (OLS). To explain the Least Squares principle, look again at Figure 4.6 where we see both small and large residuals. About half of them are positive (above the blue line) and half of them are negative (below the blue line).

The most reasonable idea is to draw a straight line that is more or less in the middle of the Y-values, in other words, with about half of the residuals positive and about half of them negative. Or perhaps we could say that on average, the residuals should be 0. A third way of saying the same thing is that the sum of the residuals should be equal to 0.

However, the criterion that all residuals should sum to 0 is not sufficient. In Figure 4.8 we see a straight line with a slope of 0 where the residuals sum to 0. However, this regression line does not make intuitive sense: it does not describe the structure in the data very well. Moreover, we see that the residuals are generally much larger than in Figure 4.6.

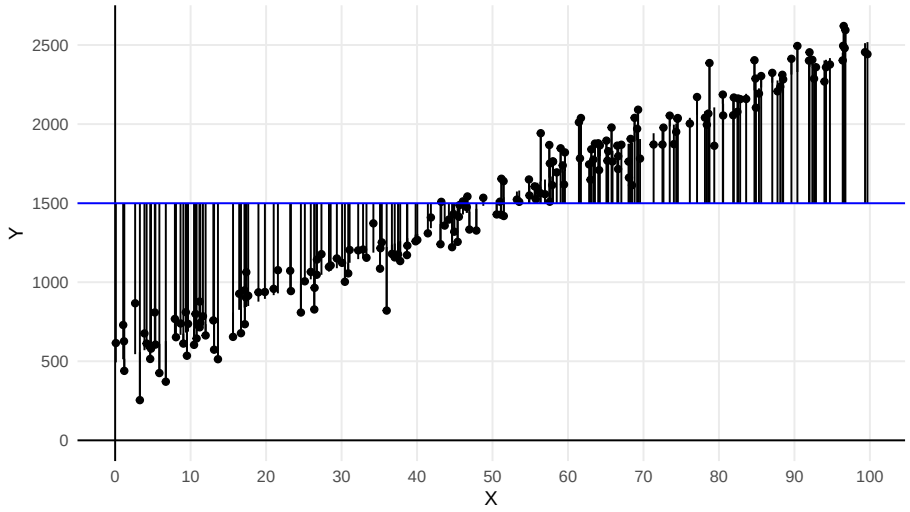


Figure 4.8: Data on variables X and Y with an added straight line. The sum of the residuals equals 0.

We therefore need a second criterion to find a nice straight line. We want the residuals to sum to 0, but also want the residuals to be as small as possible: the discrepancies between what the linear equation predicts (the $\widehat{Y}$-values) and the actual Y-values should be as small as possible.

So now we have two criteria: we want the sum of the residuals to be 0 (about half of them negative, half of them positive), and we want the residuals to be as small as possible. We can achieve both of these when we use as our criterion the idea that the sum of the *squared* residuals be as small as possible. Recall from Chapter 1 that the sum of the squared deviations from the mean is closely related to the variance. So if the sum of the squared residuals is as small as possible, we know that the *variance* of the residuals is as small as possible. Thus, as our criterion we can use the regression line for which the sum of the squared differences between predicted and observed Y-values is as small as possible.

Figure 4.9 shows three different regression lines for the same data set. Figure 4.10 shows the respective distributions of the residuals. For the first line, we see that the residuals sum to 0, for the residuals are on average 0 (the red vertical line). However, we see quite large residuals. The residuals for the second line are smaller: we see very small positive residuals, but the negative residuals are still quite large. We also see that the residuals do not sum to 0. For the third line, we see both criteria optimised: the sum of the residuals is zero and the residuals are all very small. We see that for regression line 3, the sum of squared residuals is at its minimum value. It can also be mathematically shown that if we minimise the sum of squared differences between the predicted and observed Y-values, they automatically show a mean of 0, satisfying the first criterion.

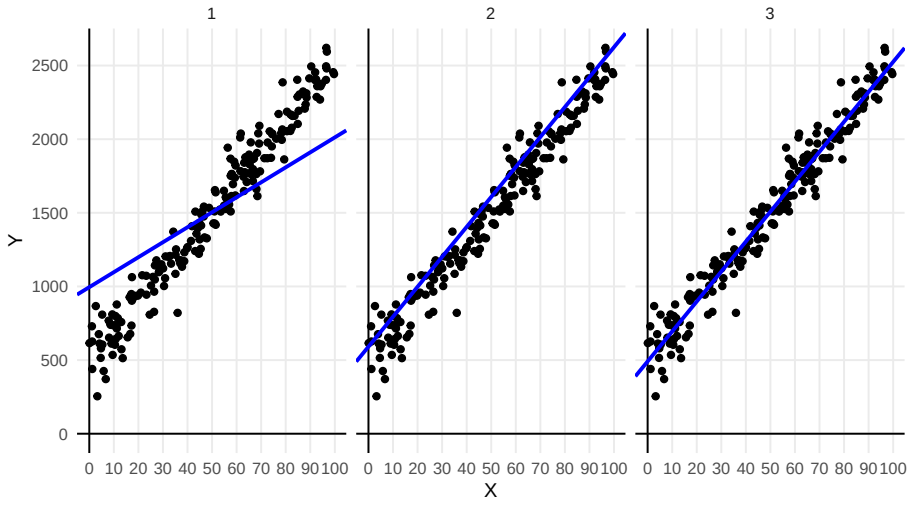


Figure 4.9: Three times the same data set, but with different regression lines.

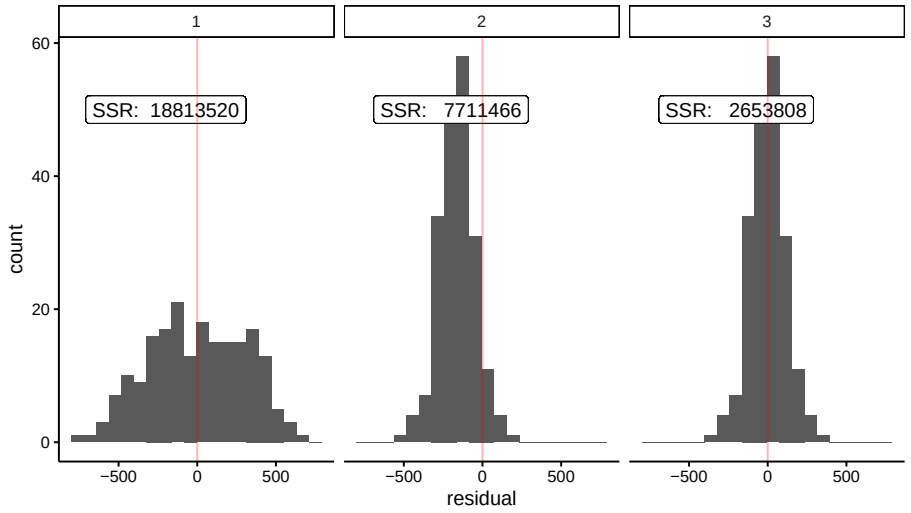


Figure 4.10: Histogram of the residuals (errors) for three different regression lines, and the respective sums of squared residuals (SSR).

In summary, when we want to have a straight line that describes our data best (i.e., the regression line), we'd like a line such that the residuals are on average 0 (i.e, sum to 0), and where we see the smallest residuals possible. We reach these criteria when we use the line in such a way that we have the lowest value for the sum of the squared residuals possible. This line is therefore called the least squares or OLS regression line.

There are generally two ways of finding the intercept and the slope values that satisfy the Least Squares principle.

1. **Numerical search** Try some reasonable combinations of values for the intercept and slope, and for each combination, calculate the sum of the squared residuals. For the combination that shows the lowest value, try to tweak the values of the intercept and slope a bit to find even lower values for the sum of the squared residuals. Use some stopping rule otherwise you keep looking forever.

2. **Analytical approach** For problems that are not too complex, like this linear regression problem, there are simple mathematical equations to find the combination of intercept and slope that gives the lowest sum of squared residuals.

Using the analytical approach, it can be shown that the Least Squares slope can be found by solving:

$$\text{slope} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2} \tag{4.8}$$

and the Least Squares intercept can be found by:

$$\text{intercept} = \overline{Y} - \text{slope} \times \overline{X} \tag{4.9}$$

where $\overline{X}$ and $\overline{Y}$ are the means of the independent X_i and dependent Y_i observations, respectively.

In daily life, we do not compute this by hand but let computers do it for us, with software like for instance R.

Overview

- **residual**: the difference between a certain data point (the *actual* value) and what the linear equation predicts.
- **linear regression**: When we want to describe the behaviour of the dependent variable (the Y-variable on the vertical axis) on the basis of the independent variable (the X-value on the horizontal axis) by a straight line, linear regression is the process of finding such a straight line.
- **Least Squares principle**: In order to find the best regression line, you need a criterion. The Least Squares principle is such a criterion and specifies that the sum of the squares of the residuals should be as small as possible.

4.6 Linear models

By performing a regression analysis of Y on X, we try to predict the Y-value from a given X on the basis of a linear equation. We try to find an intercept and a slope for that linear equation such that our prediction is 'best'. We define 'best' as the linear equation for which we see the lowest possible value for the sum of the squared residuals (least squares principle).

Thus, the prediction for the ith value of Y ($\widehat{Y}_i$) can be computed by the linear equation

$$\widehat{Y}_i = b_0 + b_1 X_i \tag{4.10}$$

where we use b_0 to denote the intercept, b_1 to denote the slope and X_i as the ith value of X.

In reality, the predicted values for Y always deviate from the observed values of Y: there is practically always an error e that is the difference between $\widehat{Y}_i$ and Y_i. Thus we have for the observed values of Y

$$Y_i = \widehat{Y}_i + e_i = b_0 + b_1 X_i + e_i \tag{4.11}$$

Typically, we assume that the residuals e have a normal distribution with a mean of 0 and a variance that is often unknown but that we denote by σ_e^2. Such a normal distribution is denoted by $N(0, \sigma_e^2)$. Taking the linear equation and the normally distributed residuals together, we have a *model* for the variables X and Y.

$$Y_i = b_0 + b_1 X_i + e_i \tag{4.12}$$

$$e_i \sim N(0, \sigma_e^2) \tag{4.13}$$

A model is a specification of how a set of variables relate to each other. Note that the model for the residuals, the normal distribution, is an essential part of the model. The linear equation only gives you *predictions* of the dependent variable, not the variable itself. Together, the linear equation and the distribution of the residuals give a full description of how the dependent variable *depends* on the independent variable.

A model may be an adequate description of how variables relate to each other or it may not, that is for the data analyst to decide. If it is an adequate description, it may be used to predict yet unseen data on variable Y (because we can't see into the future), or it may be used to draw some inferences on data that can't be seen, perhaps because of limitations in data collection. Remember Chapter 2 where we made a distinction between sample data and population data. We could use the linear equation that we obtain using a sample of data to make predictions for data in the population. We delve deeper into that issue in Chapter **??**.

The model that we see in Equations 4.12 and 4.13 is a very simple form of the *linear model*. The linear model that we see here is generally known as the *simple regression model*: the simple regression model is a linear model for one numeric dependent variable, an intercept, a slope for only one (hence 'simple') numeric independent variable, and normally distributed residuals. In the remainder of this book, we will see a great variety of linear models: with one or more independent variables, with numeric or with categorical independent variables, and with numeric or categorical dependent variables. All these models can be seen as extensions of this simple regression model. What they all have in common is that they aim to predict one dependent variable from one or more independent variables using a linear equation.

4.7 Finding the OLS intercept and slope using R

Figure 4.11 shows a data set on the relationship between the number of cylinders (`cyl`) and miles per gallon (`mpg`) in 1 cars. The blue line is the least squares regression line. The coefficients for this line can be found with R using the following code:

```
model <- lm(mpg ~ cyl, data = mtcars)
model
```

In the syntax we use the `lm()` function to indicate that we want to use the linear model. Next, we say that we want to model the variable `mpg`. The ~ ('tilde') sign means "is modelled by" or "is predicted by", and next we plug in the independent variable `cyl`. Thus, this code says we want to model the `mpg` variable by the `cyl` variable, or predict `mpg` scores by `cyl`. Next, we indicate that the `mpg` and `cyl` variables can be found in the data frame called `mtcars`. Finally, we store the results in the object `model`.

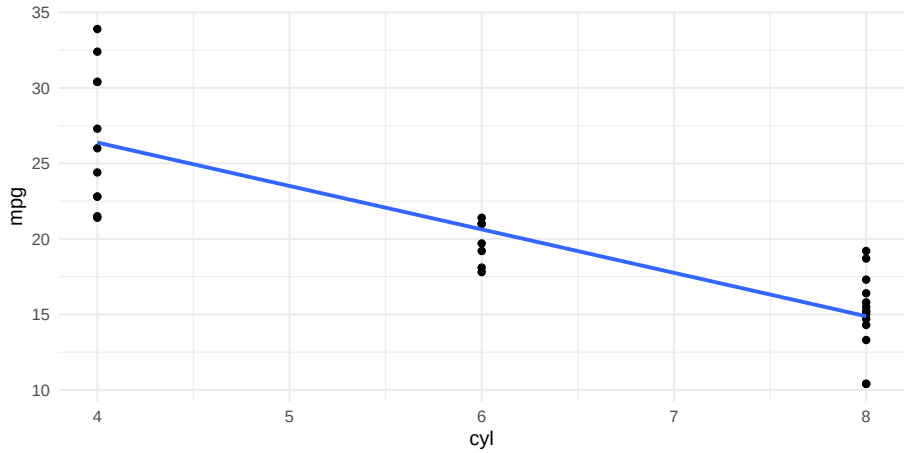


Figure 4.11: Data set on number of cylinders (`cyl`) and miles per gallon (`mpg`) in 32 cars.

In the second line of code we indicate that we want to see the results, that we stored in `model`.

```
model

##
## Call:
## lm(formula = mpg ~ cyl, data = mtcars)
##
## Coefficients:
## (Intercept)          cyl
##      37.885       -2.876
```

The output above shows us a repetition of the `lm()` analysis, and then two coefficients. These are the *regression coefficients* that we wanted: the first is the intercept, and the second is the slope. These coefficients are the *parameters* of the regression model. Parameters are parts of a model that can vary from data set to data set, but that are not variables (variables vary within a data set, parameters do not). Here we use the linear model from Equation 4.12 where b_0, b_1 and σ_e^2 are parameters since they are different for different data sets.

The output does not look very pretty. Using the `broom` package, we can get the same information about the analysis, and more:

```
library(broom)
results <- lm(mpg ~ cyl, data = mtcars)
results %>% tidy()

## # A tibble: 2 x 5
```

```
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    37.9      2.07      18.3  8.37e-18
## 2 cyl            -2.88     0.322     -8.92 6.11e-10
```

R then shows two rows of values, one for the intercept and one for the slope parameter for `cyl`. For now, we only look at the first two columns. In these columns we find the least squares values for these parameters for this data set on 32 cars that we are analysing here.

In the second column, called *estimate*, we see that the intercept parameter has the value 37.9 (when rounded to 1 decimal) and the slope has the value -2.88. Thus, with this output, the linear equation for the regression equation can be filled in:

$$\texttt{mpg} = 37.9 - 2.88 \times \texttt{cyl} + e \tag{4.14}$$

With this equation we can predict values for `mpg` for number of cylinders that are not even in the data set displayed in Figure 4.11. For instance, that plot does not show a car with 2 cylinders, but on the basis of the linear equation, the best bet would be that such a car would run $37.9 - 2.88 \times 2 = 32.14$ miles per gallon.

The OLS linear model parameters are in the *estimate* column of the R output, but there are also a number of other columns: standard error, statistic (t), and p-value, terms that we encountered earlier in Chapter 2. These columns will be discussed further in Chapters **??** and **??**.

4.8 Pearson correlation

For any set of two numeric variables, we can determine the least squares regression line. However, it depends on the data set how well that regression line describes the data. Figure 4.12 shows two different data sets on variables X and Y. Both plots also show the least squares regression line, and they both turn out to be exactly the same: $Y = 100 + 10X$.

We see that the regression line describes data set A very well (left panel): the observed dots are very close to the line, which means that the residuals are very small. The regression line does a worse job for data set B (right panel) since there are quite large discrepancies between the observed Y-values and the predicted Y-values. Put differently, the regression equation can be used to predict Y-values in data set A very well, almost without error, whereas the regression line cannot be used to predict Y-values in data set B very precisely. The regression line is also the least squares regression line for data set B, so any improvement by choosing another slope or intercept is not possible.

Francis Galton was the first to think about how to quantify this difference in the ability of a regression line to predict the dependent variable. Karl Pearson later worked on this measure and therefore it came to be called Pearson's

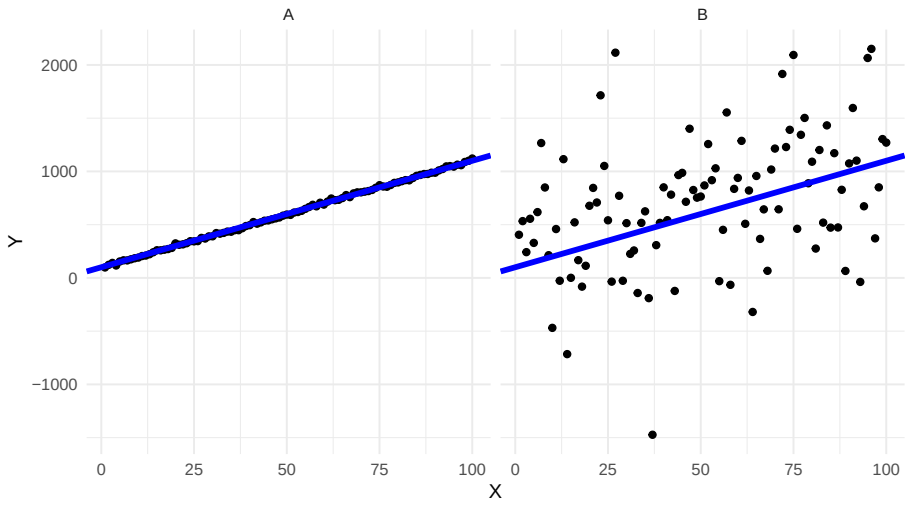


Figure 4.12: Two data sets with the same regression line.

correlation coefficient. It is a standardised measure, so that it can be used to compare different data sets.

In order to get to Pearson's correlation coefficient, you first need to standardise both the independent variable, X, and the dependent variable, Y. You standardise scores by taking their values, subtract the mean from them, and divide by the standard deviation (see Chapter 1). So, in order to obtain a standardised value for $X = x$ we compute z_X,

$$z_X = \frac{x - \overline{X}}{\sigma_X} \tag{4.15}$$

and in order to obtain a standardised value for $Y = y$ we compute z_Y,

$$z_Y = \frac{y - \overline{Y}}{\sigma_Y}. \tag{4.16}$$

Let's do this both for data set A and data set B, and plot the standardised scores, see Figure 4.13. If we then plot the least squares regression lines for the standardised values, we obtain different equations. For both data sets, the intercept is 0 because by standardising the scores, the means become 0. But the slopes are different: in data set A, the slope is 0.997 and in data set B, the slope is 0.376.

$$Z_Y = 0 + 0.997 \times Z_X = 0.997 \times Z_X \tag{4.17}$$

$$Z_Y = 0 + 0.376 \times Z_X = 0.376 \times Z_X \tag{4.18}$$

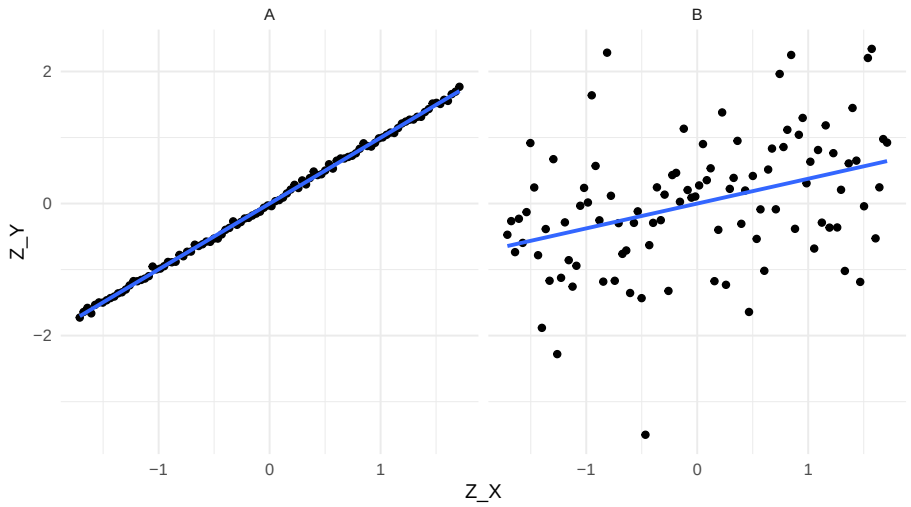


Figure 4.13: Two data sets, with different regression lines after standardisation.

These two slopes, the slope for the regression of standardized Y-values on standardized X-values, are the correlation coefficients for data sets A and B, respectively. For obvious reasons, the correlation is sometimes also referred to as the *standardised slope coefficient*.

Correlation stands for the *co-relation* between two variables. It tells you how well one variable can be predicted from the other. The correlation is bi-directional: the correlation between Y and X is the same as the correlation between X and Y. For instance in Figure 4.13, if we would have put the Z_X-variable on the Z_Y-axis, and the Z_Y-variable on the Z_X-axis, the slopes would be exactly the same. This is true because the variances of the Y- and X-variables are equal after standardisation (both variances equal to 1).

Since a slope can be negative, a correlation can be negative too. Furthermore, a correlation is always between -1 and 1. Look at Figure 4.13: the correlation between X and Y is 0.997. The dots are almost on a straight line. If the dots would all be exactly on the straight line, the correlation would be 1.

Figure **??** shows a number of scatter plots of X and Y with different correlations. Note that if dots are very close to the regression line, the correlation can still be close to 0: if the slope is 0 (bottom-left panel), then one variable cannot be predicted from the other variable, hence the correlation is 0, too.

In summary, the correlation coefficient indicates how well one variable can be predicted from the other variable. It is the slope of the regression line if both variables are standardised. If prediction is not possible (when the regression slope is 0), the correlation is 0, too. If the prediction is perfect, without errors (no residuals) and with a slope unequal to 0, then the correlation is either -1 or +1, depending on the sign of the slope. The correlation coefficient between

variables X and Y is usually denoted by r_{XY} for the sample correlation and ρ_{XY} (pronounced 'rho') for the population correlation.

4.9 Covariance

The correlation ρ_{XY} as defined above is a standardised measure for how much two variables co-relate. There exists also an unstandardised measure for how much two variables co-relate: the *covariance*. The correlation ρ_{XY} is the slope when X and Y each have variance 1. When you multiply correlation ρ_{XY} by a quantity indicating the variation of the two variables, you get the covariance. This quantity is the product of the two respective standard deviations.

The covariance between variables X and Y, denoted by σ_{XY}, can be computed as:

$$\sigma_{XY} = \rho_{XY} \times \sigma_X \times \sigma_Y \tag{4.19}$$

For example, if the variance of X equals 49 and the variance of Y equals 25, then the respective standard deviations are 7 and 5. If the correlation between X and Y equals 0.5, then the covariance between X and Y is equal to $0.5 \times 7 \times 5 = 17.5$.

Similar to the correlation, the covariance of two variables indicates by how much they co-vary. For instance, if the variance of X is 3 and the variance of Y is 5, then a covariance of 2 indicates that X and Y co-vary: if X increases by a certain amount, Y also increases. If you want to know how many standard deviations Y increases if X increases with one standard deviation, you can turn the covariance into a correlation by dividing the covariance by the respective standard deviations.

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{2}{\sqrt{3}\sqrt{5}} = 0.52 \tag{4.20}$$

Similar to correlations and slopes, covariances can also be negative.

Instead of computing the covariance on the basis of the correlation, you can also compute the covariance using the data directly. The formula for the covariance is

$$\sigma_{XY} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n} \tag{4.21}$$

so it is the mean of the squared cross-products of two variables.[1] Note that the numerator bears close resemblance to the numerator of the equation that

[1] Again, similar to what was said about the formula for the variance of a variable, on-line you will often find the formula $\frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$. The difference is that here we are talking about the definition of the covariance of two observed variables, and that elsewhere one talks about trying to estimate the covariance between two variables in the population. Similar to the variance, the covariance in a sample is a biased estimator of the covariance in the population. To remedy this bias, we divide the cross-products not by n but by $n - 1$

we use to find the least squares slope, see Equation 4.8. This is not strange since both the slope and the covariance say something about the relationship between two variables. Also note that in the equation that we use to find the least squares slope the denominator bears close relationship to the formula for the variance, since $\sigma_X^2 = \frac{\sum(X_i - \overline{X})^2}{n}$ (see Chapter 1). We could therefore rewrite Equation 4.8 that finds the least squares or OLS slope as:

$$\begin{aligned} \text{slope}_{OLS} &= \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2} \\ &= \frac{\sigma_{XY} \times n}{\sigma_X^2 \times n} \\ &= \frac{\sigma_{XY}}{\sigma_X^2} \end{aligned} \tag{4.22}$$

This shows how all three quantities slope, correlation and covariance say something about the linear relationship between two variables. The slope says how much the dependent variable increases if the independent variable increases by 1, the correlation says how much of a standard deviation the dependent variable increases if the independent variable increases by one standard deviation (alternatively: the slope after standardisation), and the covariance is the mean cross-product of two variables (alternatively: the unstandardised correlation).

4.10 Numerical example of covariance, correlation and least square slope

Table 4.1: Computing cross-products for the covariance of two variables.

X	Y	X - meanX	Y - meanY	Crossproduct
-1	2	-0.60	2.20	-1.32
0	-1	0.40	-0.80	-0.32
1	-2	1.40	-1.80	-2.52
-2	1	-1.60	1.20	-1.92
0	-1	0.40	-0.80	-0.32

Table 4.1 shows a small data set on two variables X and Y with 5 observations. The mean value of X is -0.4 and the mean value of Y is -0.2. If we subtract the respective mean from each observed value and multiply, we get a column of cross-products. For example, take the first row: $X - \overline{X} = -1 - (-0.4) = -0.6$ and $Y - \overline{Y} = 2 - (-0.2) = 2.20$. If we multiply these numbers we get the cross-product $-0.6 \times 2.20 = -1.32$. If we compute all cross-products and sum them, we get -6.40. Dividing this by the number of observations (5), yields the covariance: -1.28.

If we compute the variances of X and Y (see Chapter 1), we obtain 1.04 and 2.16, respectively. Taking the square roots we obtain the standard deviations:

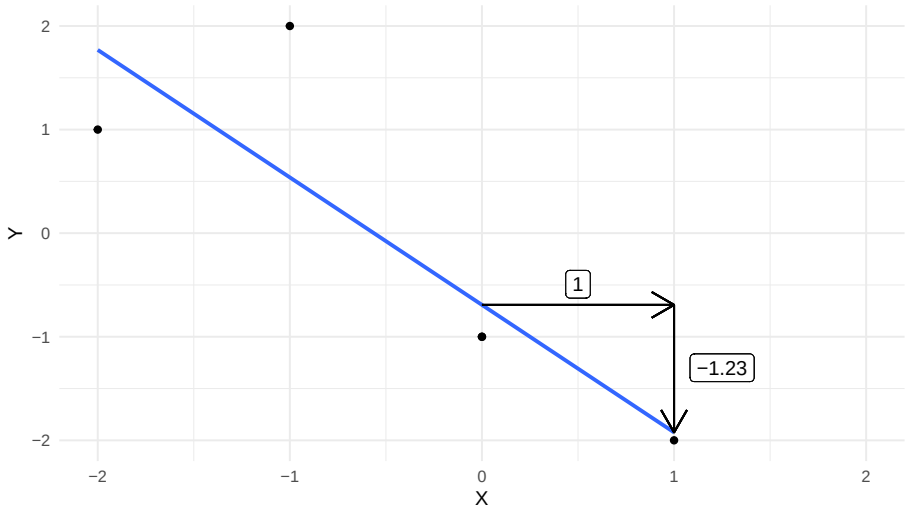


Figure 4.14: Data example and the regression line.

1.0198039 and 1.4696938. Now we can calculate the correlation on the basis of the covariance as $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{-1.28}{1.0198039 \times 1.4696938} = -0.85$.

We can also calculate the least squares slope as $\frac{\sigma_{XY}}{\sigma_X^2} = \frac{-1.28}{1.04} = -1.23$.

The original data are plotted in Figure 4.14 together with the regression line. The standardised data and the corresponding regression line are plotted in Figure 4.15. Note that the slopes are different, and that the slope of the regression line for the standardised data is equal to the correlation.

4.11 Correlation, covariance and slopes in R

Let's use the `mtcars` dataframe and compute the correlation between the number of cylinders (`cyl`) and miles per gallon (`mpg`). We do that with the function `cor`:

```
cor(mtcars$cyl, mtcars$mpg)

## [1] -0.852162
```

There is a strong negative correlation. This means that generally, the more cylinders a car has, the lower the mileage. We can also compute the covariance, with the function `cov`:

```
cov(mtcars$cyl, mtcars$mpg)

## [1] -9.172379
```

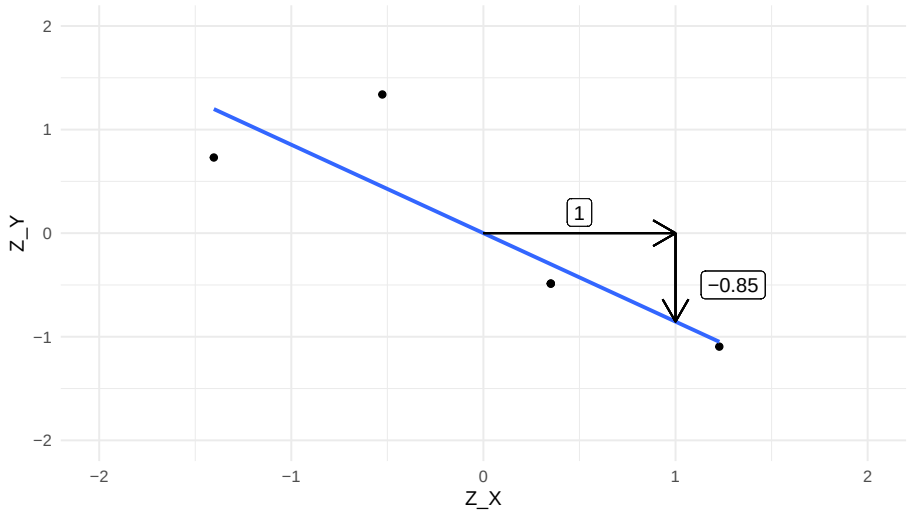


Figure 4.15: Data example (standardised values) and the regression line.

Note that R uses the formula with $n-1$ in the denominator. If we want R to compute the covariance using n in the denominator, we have to write an alternative function ourselves:

```
cov_alt <- function(x, y){
  X <- (x - mean(x)) # deviations from mean x
  Y <- (y - mean(y)) # deviations from mean y
  XY <- X %*% Y      # multiply each X with each Y and sum them
  return(XY / length(x)) # divide by n
}
cov_alt(mtcars$cyl, mtcars$mpg)

##           [,1]
## [1,] -8.885742
```

To determine the least squares slope for the regression line of `mpg` on `cyl`, we divide the covariance by the variance of `cyl` (Equation 4.22):

```
cov(mtcars$cyl, mtcars$mpg) / var(mtcars$cyl)

## [1] -2.87579
```

If we first standardise the data with the function `scale` and then compute the least squares slope, we get

```
z_mpg <- mtcars$mp %>% scale()   # standardise mpg

## Warning:  Unknown or uninitialised column:  'mp'.
## Error in array(x, c(length(x), 1L), if (!is.null(names(x))) list(names(x),
:  'data' must be of a vector type, was 'NULL'

z_cyl <- mtcars$cyl %>% scale()  # standardise cyl

cov(z_mpg, z_cyl) / var(z_cyl)

## Error in is.data.frame(x):  object 'z_mpg' not found

cor(z_mpg, z_cyl)

## Error in is.data.frame(x):  object 'z_mpg' not found

cov(z_mpg, z_cyl)

## Error in is.data.frame(x):  object 'z_mpg' not found
```

We see from the output that the slope coefficient for the standardised situation is equal to both the correlation and the covariance of the standardised values.

4.12 Explained and unexplained variance

So far in this chapter, we have seen relationships between two variables: one dependent variable and one independent variable. The dependent variable we usually denote as Y, and the independent variable we denote by X. The relationship was modelled by a linear equation: an equation with an intercept b_0 and a slope parameter b_1:

$$Y = b_0 + b_1 X \tag{4.23}$$

Further, we argued that in most cases, the relationship between X and Y cannot be completely described by a straight line. Not all of the variation in Y can be explained by the variation in X. Therefore, we have *residuals* e, defined as the difference between the observed Y-value and the Y-value that is predicted by the straight line, (denoted by $\widehat{Y}$):

$$e = Y - \widehat{Y} \tag{4.24}$$

Therefore, the relationship between X and Y is denoted by a regression equation, where the relationship is approached by a linear equation, plus a residual part e:

$$Y = b_0 + b_1 X + e \tag{4.25}$$

The linear equation gives us only the predicted Y-value, $\widehat{Y}$:

$$\widehat{Y} = b_0 + b_1 X \tag{4.26}$$

We've also seen that the residual e is assumed to have a normal distribution, with mean 0 and variance σ_e^2:

$$e \sim N(0, \sigma_e^2) \tag{4.27}$$

Remember that linear models are used to explain (or predict) the variation in Y: why are there both high values and low values for Y? Where does the variance in Y come from? Well, the linear model tells us that the variation is in part explained by the variation in X. If b_1 is positive, we predict a relatively high value for Y for a high value of X, and we predict a relatively low value for Y if we have a low value for X. If b_1 is negative, it is of course in the opposite direction. Thus, the variance in Y is in part explained by the variance in X, and the rest of the variance can only be explained by the residuals e.

$$\text{Var}(Y) = \text{Var}(\widehat{Y}) + \text{Var}(e) = \text{Var}(b_0 + b_1 X) + \sigma_e^2 \tag{4.28}$$

Because the residuals do not explain anything (we don't know where these residuals come from), we say that the *explained* variance of Y is only that part of the variance that is explained by independent variable X: $\text{Var}(b_0 + b_1 X)$. The *unexplained* variance of Y is the variance of the residuals, σ_e^2. The explained variance is often denoted by a ratio: the explained variance divided by the total variance of Y:

$$\text{Var}_{explained} = \frac{\text{Var}(b_0 + b_1 X)}{\text{Var}(Y)} = \frac{\text{Var}(b_0 + b_1 X)}{\text{Var}(b_0 + b_1 X) + \sigma_e^2} \tag{4.29}$$

From this equation we see that if the variance of the residuals is large, then the explained variance is small. If the variance of the residuals is small, the variance explained is large.

4.13 More than one predictor

In regression analysis, and in linear models in general, we try to make the explained variance as large as possible. In other words, we try to minimise the residual variance, σ_e^2. One way to do that is to use more than one independent variable. If not all of the variance in Y is explained by X, then why not include multiple independent variables?

Let's use an example with data on the weight of books, the size of books (area), and the volume of books. Let's try first to predict the weight of a book, `weight`, on the basis of the volume of the book, `volume`. Suppose we find the following regression equation and a value for σ_e^2:

$$\texttt{weight} = 107.7 + 0.71 \times \texttt{volume} + e \tag{4.30}$$
$$e \sim N(0, 15362) \tag{4.31}$$

In the data set, we see that the variance of the weight, Var(`weight`) is equal to 72274. Since we also know the variance of the residuals, we can solve for the variance explained by `volume`:

$$\text{Var}(\texttt{weight}) = 72274 = \text{Var}(107.7 + 0.7 \times \texttt{volume}) + 15362$$
$$\text{Var}(107.7 + 0.7 \times \texttt{volume}) = 72274 - 15362 = 56912$$

So the proportion of explained variance is equal to $\frac{56912}{72274} = 0.7874478$. This is quite a high proportion: nearly all of the variation in the weight of books is explained by the variation in volume.

But let's see if we can explain even more variance if we add an extra independent variable. Suppose we know the area of each book. We expect that books with a large surface area weigh more. Our linear equation then looks like this:

$$\texttt{weight} = 22.4 + 0.71 \times \texttt{volume} + 0.5 \times \texttt{area} + e \tag{4.32}$$
$$e \sim N(0, 6031) \tag{4.33}$$

How much of the variance in weight does this equation explain? The amount of explained variance equals the variance of `weight` minus the residual variance: $72274 - 6031 = 66243$. The proportion of explained variance is then equal to $\frac{66243}{72274} = 0.9165537$. So the proportion of explained variance has increased!

Note that the variance of the residuals has decreased; this is the main reason why the proportion of explained variance has increased. By adding the extra independent variable, we can explain some of the variance that without this variable could not be explained! In summary, by adding independent variables to a regression equation, we can explain more of the variance of the dependent variable. A regression analysis with more than one independent variable we call *multiple regression*. Regression with only one independent variable is called *simple regression*.

4.14 R-squared

With regression analysis, we try to explain the variance of the dependent variable. With multiple regression, we use more than one independent variable to try to explain this variance. In regression analysis, we use the term *R-squared* to refer to the proportion of explained variance, usually denoted with the symbol R^2. The unexplained variance is of course the variance of the residuals, Var(e),

usually denoted as σ_e^2. So suppose the variance of dependent variable Y equals 200, and the residual variance in a regression equation equals say 80, then R^2 or the proportion of explained variance is $(200 - 80)/200 = 0.60$.

$$R^2 = \sigma^2_{explained}/\sigma^2_Y = (\sigma^2_Y - \sigma^2_{unexplained})/\sigma^2_Y = (\sigma^2_Y - \sigma^2_e)/\sigma^2_Y \tag{4.34}$$

This is the definition of R-squared at the population level, where we know the exact values of the variances. However, we do not know these variances, since we only have a sample of all values.

We know from Chapter 2 that we can take estimators of the variances σ^2_Y and σ^2_e. We should not use the variance of Y observed in the sample, but the unbiased estimator of the variance of Y in the population

$$\widehat{\sigma^2_Y} = \frac{\Sigma_i (Y_i - \overline{Y})^2}{n-1} \tag{4.35}$$

where n is sample size (see Section 2.3).

For σ^2_e we take the unbiased estimator of the variance of the residuals e in the population

$$\widehat{\sigma^2_e} = \frac{\Sigma_i (e_i - \bar{e})^2}{n-1} = \frac{\Sigma_i e_i^2}{n-1} \tag{4.36}$$

Here we do not have to subtract the mean from the residuals, because the mean is 0 by definition.

If we plug these estimators into Equation 4.34, we get

$$\begin{aligned}
\widehat{R^2} &= \frac{\widehat{\sigma^2_Y} - \widehat{\sigma^2_e}}{\widehat{\sigma^2_Y}} = \frac{\frac{\Sigma(Y_i - \overline{Y})^2}{n-1} - \frac{\Sigma e_i^2}{n-1}}{\frac{\Sigma(Y_i - \overline{Y})^2}{n-1}} \\
&= \frac{\Sigma(Y_i - \overline{Y})^2 - \Sigma e_i^2}{\Sigma(Y_i - \overline{Y})^2} = \frac{\Sigma(Y_i - \overline{Y})^2}{\Sigma(Y_i - \overline{Y})^2} - \frac{\Sigma e_i^2}{\Sigma(Y_i - \overline{Y})^2} \\
&= 1 - \frac{SSR}{SST}
\end{aligned} \tag{4.37}$$

where SSR refers to the sum of the squared residuals (errors)[2], and SST refers to the total sum of squares (the sum of the squared deviations from the mean for variable Y).

As we saw in Section 4.5, in a regression analysis, the intercept and slope parameters are found by minimising the sum of squares of the residuals, SSR. Since the variance of the residuals is based on this sum of squares, in any regression analysis, the variance of the residuals is always as small as possible. The

[2]In the literature and online, sometimes you see SSR and sometimes you see SSE, both referring to the sum of the squared residuals

values of the parameters for which the SSR (and by consequence the variance) is smallest, are the least squares regression parameters. And if the variance of the residuals is always minimised in a regression analysis, the explained variance is always maximised!

Because in any least squares regression analysis based on a sample of data, the explained variance is always maximised, we may overestimate the variance explained in the population data. In regression analysis, we therefore very often use an *adjusted R-squared* that takes this possible overestimation (*inflation*) into account. The adjustment is based on the number of independent variables and sample size.

The formula is

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where n is sample size and p is the number of independent variables. For example, if R^2 equals 0.10 and we have a sample size of 100, and 2 independent variables, the adjusted R^2 is equal to $1 - (1 - 0.10)\frac{100-1}{100-2-1} = 1 - (0.90)\frac{99}{97} = 0.08$. Thus, the estimated proportion of variance explained at population level, corrected for inflation, equals 0.08. Because R^2 is inflated, the adjusted R^2 is never larger than the unadjusted R-squared.

$$R^2_{adj} \leq R^2$$

4.15 Multiple regression in R

Let's use the book data and run a multiple regression in R. The data set is called `allbacks` and is available in the R package `DAAG` (you may need to install that package first). The syntax looks very similar to simple regression, except that we now specify two independent variables, `volume` and `area`, instead of one. We combine these two independent variables using the +-sign.

```
library(DAAG)
model <- lm(weight ~ volume + area, data = allbacks)
model %>% tidy()
```

Below we see the output:

```
library(DAAG)
library(broom)
model <- lm(weight ~ volume + area, data = allbacks)
model %>%
  tidy()

## # A tibble: 3 x 5
##   term        estimate std.error statistic      p.value
```

```
##   <chr>          <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)   22.4     58.4       0.384 0.708
## 2 volume         0.708    0.0611   11.6   0.0000000707
## 3 area           0.468    0.102     4.59  0.000616
```

There we see an intercept, a slope parameter for `volume` and a slope parameter for `area`. Remember from Section 4.2 that the intercept is the predicted value when the independent variable has value 0. This extends to multiple regression: the intercept is the predicted value when the independent variables all have value 0. Thus, the output tells us that the predicted weight of a book that has a volume of 0 and an area of 0, is 22.413. The slopes tell us that for every unit increase in `volume`, the predicted `weight` increases by 0.708, and for every unit increase in `area`, the predicted `weight` increases by 0.468.

So the linear model looks like:

$$\texttt{weight} = 22.413 + 0.708 \times \texttt{volume} + 0.468 \times \texttt{area} + e \tag{4.38}$$

Thus, the predicted weight of a book that has a volume of 10 and an area of 5, the expected weight is equal to $22.413 + 0.708 \times 10 + 0.468 \times 5 = 31.833$.

In R, the R-squared and the adjusted R-squared can be obtained by first making a summary of the results, and then accessing these statistics directly.

```
sum <- model %>% summary()
sum$r.squared
sum$adj.r.squared
```

```
sum$r.squared

## [1] 0.9284738

sum$adj.r.squared

## [1] 0.9165527
```

The output tells you that the R-squared equals 0.93 and the adjusted R-squared 0.92. The variance of the residuals can also be found in the summary object:

```
sum$sigma^2

## [1] 6031.052
```

4.16 Multicollinearity

In general, if you add independent variables to a regression equation, the proportion explained variance, R^2, increases. Suppose you have the following three regression equations:

$$\texttt{weight} = b_0 + b_1 \times \texttt{volume} + e \tag{4.39}$$
$$\texttt{weight} = b_0 + b_1 \times \texttt{area} + e \tag{4.40}$$
$$\texttt{weight} = b_0 + b_1 \times \texttt{volume} + b_2 \times \texttt{area} + e \tag{4.41}$$

If we carry out these three analyses, we obtain an R^2 of 0.8026346 if we only use `volume` as predictor, and an R^2 of 0.1268163 if we only use `area` as predictor. So perhaps you'd think that if we take both `volume` and `area` as predictors in the model, we would get an R^2 of $0.8026346 + 0.1268163 = 0.9294509$. However, if we carry out the multiple regression with `volume` and `area`, we obtain an R^2 of 0.9284738, which is slightly less! This is not a rounding error, but results from the fact that there is a correlation between the volume of a book and the area of a book. Here it is a tiny correlation of 0.002, but nevertheless it affects the proportion of variance explained when you use both these variables.

Let's look at what happens when independent variables are strongly correlated. Table 4.2 shows measurements on a breed of seals (only measurements on the first 6 seals are shown). These data are in the dataframe `cfseals` in the package `DAAG`. Often, the age of an animal is gauged from its weight: we assume that heavier seals are older than lighter seals. If we carry out a simple regression of `age` on `weight`, we get the output

```
data(cfseal) # available in package DAAG
out1 <- lm(age ~ weight , data = cfseal)
out1 %>% tidy()

## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   11.4      4.70        2.44 2.15e- 2
## 2 weight         0.817    0.0716     11.4  4.88e-12

var(cfseal$age)  # total variance of age

## [1] 1090.855

summary(out1)$sigma^2  # variance of residuals

## [1] 200.0776
```

resulting in the equation:

$$\texttt{age} = 11.4 + 0.82 \times \texttt{weight} + e \tag{4.42}$$
$$e \sim N(0, 200) \tag{4.43}$$

From the data we calculate the variance of `age`, and we find that it is 1090.8551724. The variance of the residuals is 200, so that the proportion of explained variance is $(1090.8551724 - 200)/1090.8551724 = 0.8166576$.

Table 4.2: Part of Cape Fur Seal Data.

age	weight	heart
33.00	27.50	127.70
10.00	24.30	93.20
10.00	22.00	84.50
10.00	18.50	85.40
12.00	28.00	182.00
18.00	23.80	130.00

Since we also have data on the weight of the heart alone, we could try to predict the age from the weight of the heart. Then we get output

```
out2 <- lm(age ~ heart , data = cfseal)
out2 %>% tidy()

## # A tibble: 2 x 5
##   term        estimate std.error statistic       p.value
##   <chr>          <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept)   20.6      5.21        3.95 0.000481
## 2 heart          0.113    0.0130      8.66 0.00000000209

summary(out2)$sigma^2  # variance of residuals

## [1] 307.1985
```

that leads to the equation:

$$\texttt{age} = 20.6 + 0.11 \times \texttt{heart} + e \tag{4.44}$$

$$e \sim N(0, 307) \tag{4.45}$$

Here the variance of the residuals is 307, so the proportion of explained variance is $(1090.8551724 - 370)/1090.8551724 = 0.6608166$.

Now let's see what happens if we include both total weight and weight of the heart into the linear model. This results in the following output

```
out3 <- lm(age ~ heart + weight , data = cfseal)
out3 %>% tidy()

## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  10.3       4.99       2.06  0.0487
## 2 heart        -0.0269    0.0373    -0.723 0.476
## 3 weight        0.993     0.254      3.91  0.000567
```

```
summary(out3)$sigma^2  # variance of residuals

## [1] 203.55
```

with model equation:

$$\begin{aligned} \texttt{age} &= 10.3 - 0.03 \times \texttt{heart} + 0.99 \times \texttt{weight} + e && (4.46) \\ e &\sim N(0, 204) && (4.47) \end{aligned}$$

Here we see that the regression parameter for `weight` has increased from 0.82 to 0.99. At the same time, the regression parameter for `heart` has decreased, has even become negative, from 0.11 to -0.03. From this equation we see that there is a strong relationship between the total weight and the age of a seal, but on top of that, for every unit increase in the weight of the heart, there is a very small decrease in the expected age. The slope for `heart` has become practically negligible, so we could say that on top of the effect of total weight, there is no remaining relationship between the weight of the heart and age. In other words, once we can use the total weight of a seal, there is no more information coming from the weight of the heart.

This is because the total weight of a seal and the weight of its heart are strongly correlated: heavy seals generally have heavy hearts. Here the correlation turns out to be 0.96, almost perfect! This means that if you know the total weight of a seal, you practically know the weight of its heart. This is logical of course, since the total weight is a composite of all the weights of all the parts of the animal: the total weight variable *includes* the weight of the heart.

Here we have seen, that if we use multiple regression, we should be aware of how strongly the independent variables are correlated. Highly correlated predictor variables do not add extra predictive power. Worse: they can cause problems in obtaining regression parameters because it becomes hard to tell which variable is more important: if they are strongly correlated (positive or negative), then they measure almost the same thing!

When two predictor variables are perfectly correlated, either 1 or -1, regression is no longer possible, the software stops and you get a warning. We call such a situation *multicollinearity*. But also if the correlation is close to 1 or -1, you should be very careful interpreting the regression parameters. If this happens, try to find out what variables are highly correlated, and select the variable that makes most sense.

In our seal data, there is a very high correlation between the variables `heart` and `weight` that can cause computational and interpretation problems. It makes more sense to use only the total weight variable, since when seals get older, *all* their organs and limbs grow larger, not just their heart.

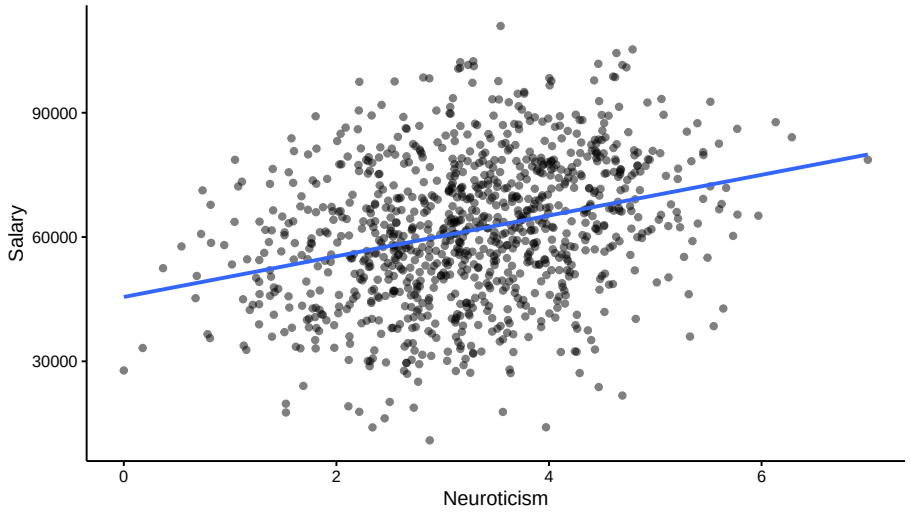


Figure 4.16: Simulated HR data set.

4.17 Simpson's paradox

With multiple regression, you may uncover very surprising relationships between two variables, that can never be found using simple regression. Here's an example from Paul van der Laken[3], who simulated a data set on the topic of Human Resources (HR).

Assume you run a company with 1000 employees and you have asked all of them to fill out a Big Five personality survey. Per individual, you therefore have a score depicting their personality characteristic `Neuroticism`, which can run from 0 (not at all neurotic) to 7 (very neurotic). Now you are interested in the extent to which this `Neuroticism` of employees relates to their `salary` (measured in Euros per year).

We carry out a simple regression, with `salary` as our dependent variable and `Neuroticism` as our independent variable. We then find the following regression equation:

$$\texttt{salary} = 45543 + 4912 \times \texttt{Neuroticism} + e \tag{4.48}$$

Figure 4.16 shows the data and the regression line. From this visualisation it looks like Neuroticism relates *positively* to their yearly salary: more neurotic people earn more salary than less neurotic people. More precisely, we see in the equation that for every unit increase on the `Neuroticism` scale, the predicted salary increases with 4912 Euros a year.

[3] https://paulvanderlaken.com/2017/09/27/simpsons-paradox-two-hr-examples-with-r-code/

Next we run a multiple regression analysis. We suspect that one other very important predictor for how much people earn is their educational background. The `Education` variable has three levels: 0, 1 and 2. If we include both `Education` and `Neuroticism` as independent variables and run the analysis, we obtain the following regression equation:

$$\texttt{salary} = 50935 - 3176 \times \texttt{Neuroticism} + 20979 \times \texttt{Education} + e \qquad (4.49)$$

Note that we now find a *negative* slope parameter for the effect of `Neuroticism`! This implies there is a relationship in the data where neurotic employees earn *less* than their less neurotic colleagues! How can we reconcile this seeming paradox? Which result should we trust: the one from the simple regression, or the one from the multiple regression?

The answer is: neither. Or better: both! Both analyses give us different information.

Let's look at the last equation more closely. Suppose we make a prediction for a person with a low educational background (`Education` = 0). Then the equation tells us that the expected salary of a person with a neuroticism score of 0 is around 50935, and of a person with a neuroticism score of 1 is around 47759. That's an increase of -3176, which is the slope for `Neuroticism` in the multiple regression. So for employees with low education, the more neurotic employees earn less! If we do the same exercise for average education and high education employees, we find exactly the same pattern: for each unit increase in neuroticism, the predicted yearly salary drops by 3176 Euros.

It is true that in this company, the more neurotic persons generally earn a higher salary. But if we take into account educational background, the relationship flips around. This can be seen from Figure 4.17: looking only at the people with a low educational background (`Education` = 0, the red data points), then the more neurotic people earn less than their less neurotic colleagues with a similar educational background. And the same is true for people with an average education (`Education` = 1, the green data points) and a high education (`Education` = 2, the blue data points). Only when you put all employees together in one group, you see a positive relationship between `Neuroticism` and `salary`.

Simpson's paradox tells us that we should always be careful when interpreting positive and negative correlations between two variables: what might be true at the total group level, might not be true at the level of smaller subgroups. Multiple linear regression helps us investigate correlations more deeply and uncover exciting relationships between multiple variables.

Simpson's paradox helps us in interpreting the slope coefficients in multiple regression. In simple regression, when we only have one independent variable, we saw that the slope for an independent variable A is the increase in the dependent variable if we increase variable A by one unit. In multiple regression, we have multiple independent variables, say A, B and C. The interpretation for the slope coefficient for variable A is then the increase in the dependent variable

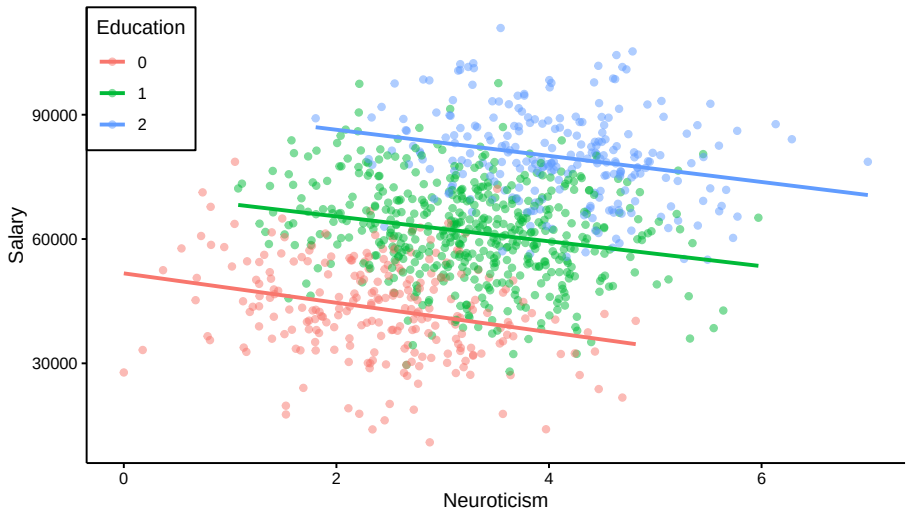


Figure 4.17: Same HR data, now with markers for different education levels.

if we increase variable A by one unit, *with the other independent variables B and C held constant*. For example, the slope for variable A is the increase when we take particular values for variables B and C, say $B = 5$ and $C = 7$.

Multiple regression therefore plays an important part in studying causation. Suppose that a researcher finds in South-African beach data that on days with high ice cream sales there are also more shark attacks. Might this indicate that there is a causal relationship between ice cream sales and shark attacks? Might bellies full of ice cream be more attractive to sharks? Or when there are many shark attacks, might people prefer eating ice cream over swimming? Alternatively, there might be a third variable that explains both the shark attacks and the ice cream sales: temperature! Sharks attack during the summer when temperature is high, and that's also the time people eat more ice cream. There is no causal relationship, since if you only look at data from sunny summer days (holding temperature constant), you don't see a relationship between shark attacks and ice cream sales (just many shark attacks and high ice cream sales). And if you only look at cold wintry days, you also see no relationship (no shark attacks and no ice cream sales). But if you take *all* days into account, you see a relationship between shark attacks and ice cream sales. Because this correlation is non-causal and explained by the third variable temperature, we call this correlation a *spurious* correlation.

This spurious correlation is plotted in Figure 4.18. If you look at all the data points at once, you see a linear relationship between shark attacks and ice cream sales. However, if you hold temperature constant by looking at only the light blue data points (high temperatures), there is no linear relationship. Neither is there a linear relationship when you only look at the dark blue data

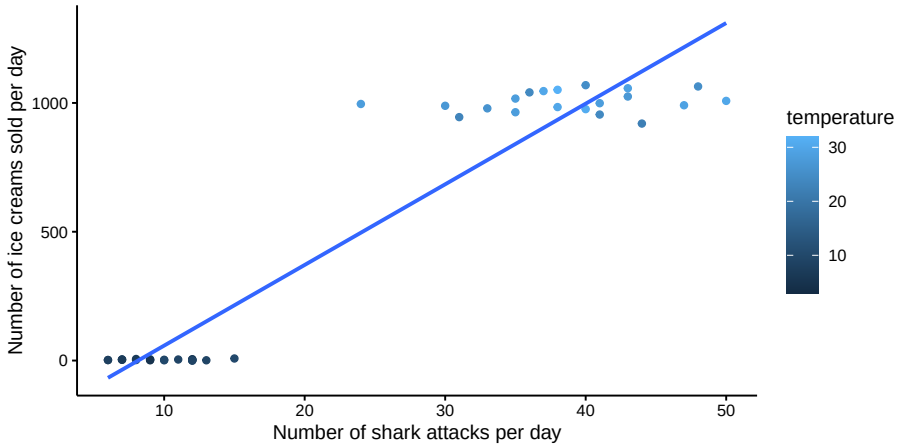


Figure 4.18: A spurious correlation between the number of shark attacks and ice cream sales.

points (low temperatures).

Appendices

Appendix A

Cumulative probabilities for the standard normal distribution

Table A.1: Cumulative proportions for the standard normal distribution.

z	cum. proportion	z	cum. proportion	z	cum. proportion
0.00	0.500000	0.90	0.815940	1.80	0.964070
0.01	0.503989	0.91	0.818589	1.81	0.964852
0.02	0.507978	0.92	0.821214	1.82	0.965620
0.03	0.511966	0.93	0.823814	1.83	0.966375
0.04	0.515953	0.94	0.826391	1.84	0.967116
0.05	0.519939	0.95	0.828944	1.85	0.967843
0.06	0.523922	0.96	0.831472	1.86	0.968557
0.07	0.527903	0.97	0.833977	1.87	0.969258
0.08	0.531881	0.98	0.836457	1.88	0.969946
0.09	0.535856	0.99	0.838913	1.89	0.970621
0.10	0.539828	1.00	0.841345	1.90	0.971283
0.11	0.543795	1.01	0.843752	1.91	0.971933
0.12	0.547758	1.02	0.846136	1.92	0.972571
0.13	0.551717	1.03	0.848495	1.93	0.973197
0.14	0.555670	1.04	0.850830	1.94	0.973810
0.15	0.559618	1.05	0.853141	1.95	0.974412
0.16	0.563559	1.06	0.855428	1.96	0.975002
0.17	0.567495	1.07	0.857690	1.97	0.975581
0.18	0.571424	1.08	0.859929	1.98	0.976148
0.19	0.575345	1.09	0.862143	1.99	0.976705
0.20	0.579260	1.10	0.864334	2.00	0.977250

Continued on next page

Table A.1: Cumulative proportions for the standard normal distribution.

z	cum. proportion	z	cum. proportion	z	cum. proportion
0.21	0.583166	1.11	0.866500	2.01	0.977784
0.22	0.587064	1.12	0.868643	2.02	0.978308
0.23	0.590954	1.13	0.870762	2.03	0.978822
0.24	0.594835	1.14	0.872857	2.04	0.979325
0.25	0.598706	1.15	0.874928	2.05	0.979818
0.26	0.602568	1.16	0.876976	2.06	0.980301
0.27	0.606420	1.17	0.879000	2.07	0.980774
0.28	0.610261	1.18	0.881000	2.08	0.981237
0.29	0.614092	1.19	0.882977	2.09	0.981691
0.30	0.617911	1.20	0.884930	2.10	0.982136
0.31	0.621720	1.21	0.886861	2.11	0.982571
0.32	0.625516	1.22	0.888768	2.12	0.982997
0.33	0.629300	1.23	0.890651	2.13	0.983414
0.34	0.633072	1.24	0.892512	2.14	0.983823
0.35	0.636831	1.25	0.894350	2.15	0.984222
0.36	0.640576	1.26	0.896165	2.16	0.984614
0.37	0.644309	1.27	0.897958	2.17	0.984997
0.38	0.648027	1.28	0.899727	2.18	0.985371
0.39	0.651732	1.29	0.901475	2.19	0.985738
0.40	0.655422	1.30	0.903200	2.20	0.986097
0.41	0.659097	1.31	0.904902	2.21	0.986447
0.42	0.662757	1.32	0.906582	2.22	0.986791
0.43	0.666402	1.33	0.908241	2.23	0.987126
0.44	0.670031	1.34	0.909877	2.24	0.987455
0.45	0.673645	1.35	0.911492	2.25	0.987776
0.46	0.677242	1.36	0.913085	2.26	0.988089
0.47	0.680822	1.37	0.914657	2.27	0.988396
0.48	0.684386	1.38	0.916207	2.28	0.988696
0.49	0.687933	1.39	0.917736	2.29	0.988989
0.50	0.691462	1.40	0.919243	2.30	0.989276
0.51	0.694974	1.41	0.920730	2.31	0.989556
0.52	0.698468	1.42	0.922196	2.32	0.989830
0.53	0.701944	1.43	0.923641	2.33	0.990097
0.54	0.705401	1.44	0.925066	2.34	0.990358
0.55	0.708840	1.45	0.926471	2.35	0.990613
0.56	0.712260	1.46	0.927855	2.36	0.990863
0.57	0.715661	1.47	0.929219	2.37	0.991106
0.58	0.719043	1.48	0.930563	2.38	0.991344
0.59	0.722405	1.49	0.931888	2.39	0.991576
0.60	0.725747	1.50	0.933193	2.40	0.991802
0.61	0.729069	1.51	0.934478	2.41	0.992024

Continued on next page

Table A.1: Cumulative proportions for the standard normal distribution.

z	cum. proportion	z	cum. proportion	z	cum. proportion
0.62	0.732371	1.52	0.935745	2.42	0.992240
0.63	0.735653	1.53	0.936992	2.43	0.992451
0.64	0.738914	1.54	0.938220	2.44	0.992656
0.65	0.742154	1.55	0.939429	2.45	0.992857
0.66	0.745373	1.56	0.940620	2.46	0.993053
0.67	0.748571	1.57	0.941792	2.47	0.993244
0.68	0.751748	1.58	0.942947	2.48	0.993431
0.69	0.754903	1.59	0.944083	2.49	0.993613
0.70	0.758036	1.60	0.945201	2.50	0.993790
0.71	0.761148	1.61	0.946301	2.51	0.993963
0.72	0.764238	1.62	0.947384	2.52	0.994132
0.73	0.767305	1.63	0.948449	2.53	0.994297
0.74	0.770350	1.64	0.949497	2.54	0.994457
0.75	0.773373	1.65	0.950529	2.55	0.994614
0.76	0.776373	1.66	0.951543	2.56	0.994766
0.77	0.779350	1.67	0.952540	2.57	0.994915
0.78	0.782305	1.68	0.953521	2.58	0.995060
0.79	0.785236	1.69	0.954486	2.59	0.995201
0.80	0.788145	1.70	0.955435	2.60	0.995339
0.81	0.791030	1.71	0.956367	2.70	0.996533
0.82	0.793892	1.72	0.957284	2.80	0.997445
0.83	0.796731	1.73	0.958185	2.90	0.998134
0.84	0.799546	1.74	0.959070	3.00	0.998650
0.85	0.802337	1.75	0.959941	3.20	0.999313
0.86	0.805105	1.76	0.960796	3.40	0.999663
0.87	0.807850	1.77	0.961636	3.60	0.999841
0.88	0.810570	1.78	0.962462	3.80	0.999928
0.89	0.813267	1.79	0.963273	4.00	0.999968

Appendix B

Critical values for the t-distribution

Table B.1: Two-tailed critical values for the t-distribution, given the degrees of freedom (rows) and type I error rate α (columns).

df	0.2	0.1	0.05	0.02	0.01	0.005	0.002	0.001
1	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
10000	1.282	1.645	1.960	2.327	2.576	2.808	3.091	3.291