Analysing data using linear models

Stéphanie M. van den Berg

Fifth edition (R)
(November 6, 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:	November 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 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 nearly 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 variables (column names), so we can supply them to the function argument `values_from` to make new variables:

```
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, positive 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 6). 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 7.

```
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. Variable `shirtsize` is stored as an ordinal factor and will be

treated accordingly. Variable `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).

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

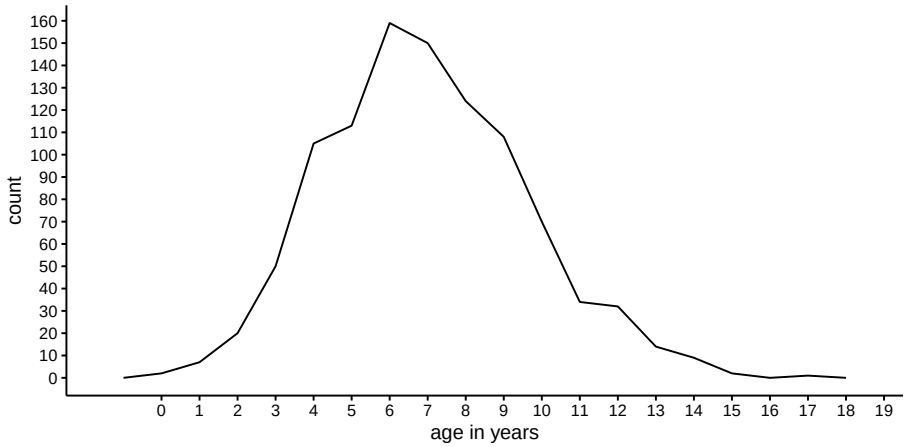


Figure 1.1: A frequency plot

and 8, but not so many children with ages below 3 or above 14. The advantage 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.

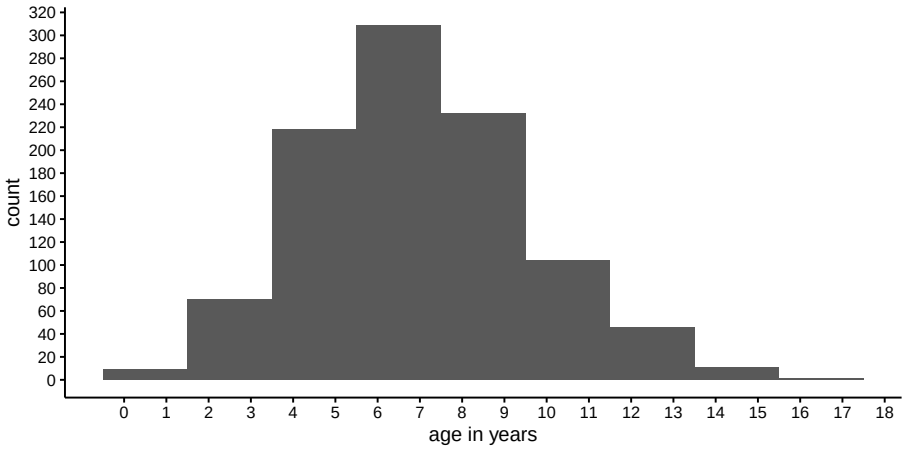


Figure 1.2: A histogram

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.

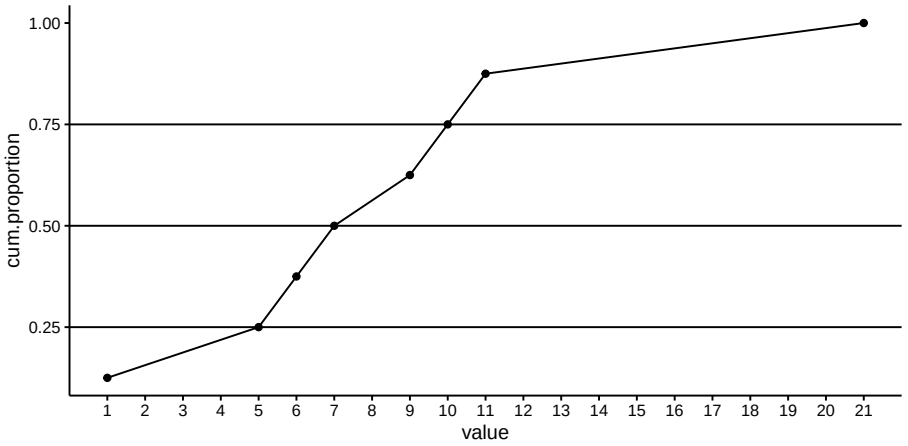


Figure 1.3: Cumulative proportions.

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.

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

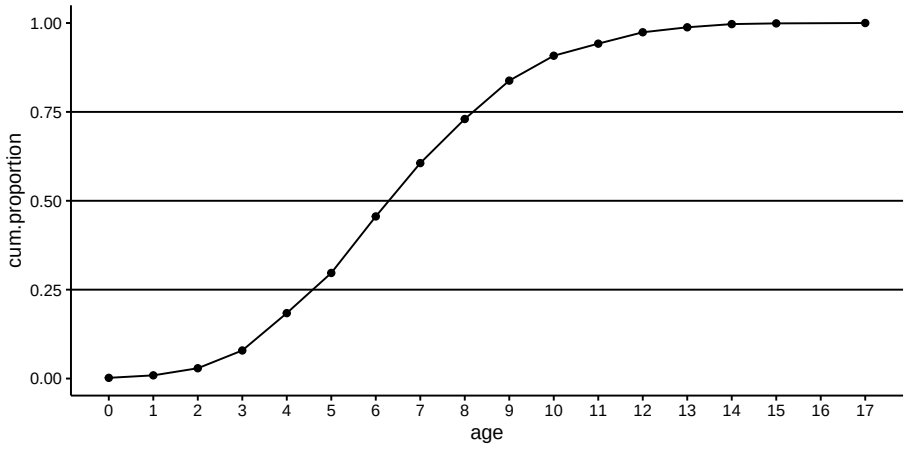


Figure 1.4: Cumulative proportions.

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]:

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

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

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

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 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

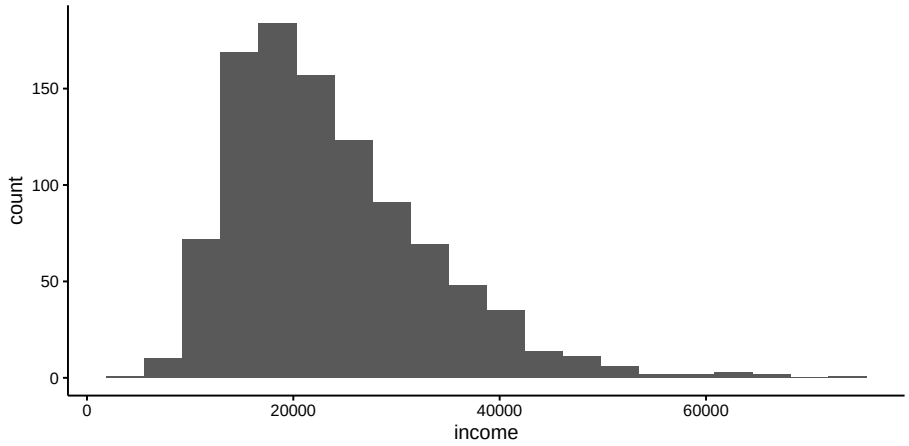


Figure 1.5: Distribution of yearly income.

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 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

[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!

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

"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))

## # 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 range

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 each other.

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 range* (IQR) or 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 range is equal to $147 - 120 = 27$.

Thus, we can measure variation using the range or the interquartile range. 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. This is a sum of squared differences, or sum of squares.

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]

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

[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.

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, 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, respectively.

```
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 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.

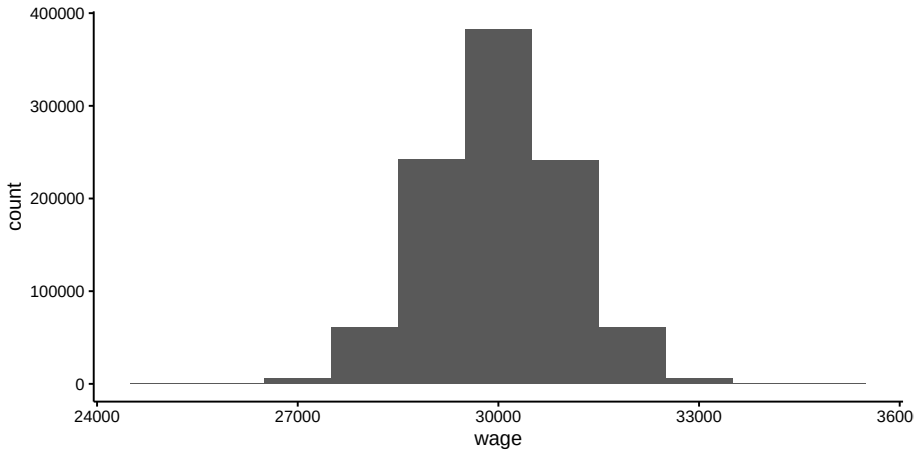


Figure 1.6: A histogram of wages with bin size 1000.

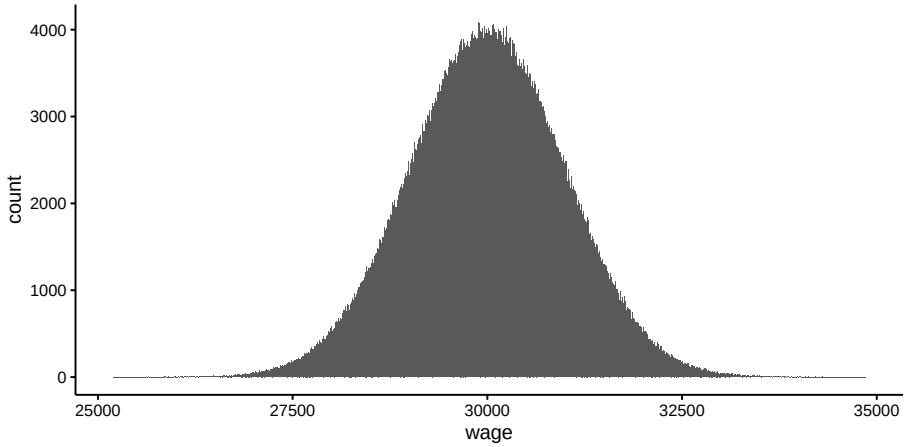


Figure 1.7: A histogram of wages with bin size 10.

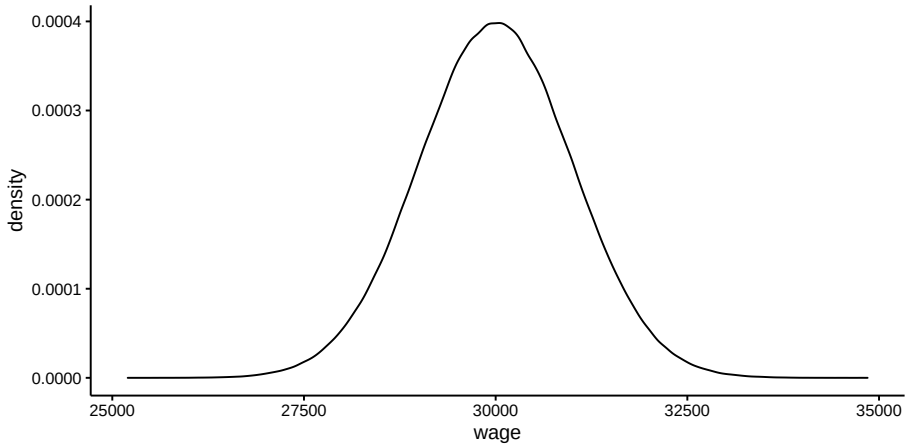


Figure 1.8: A density plot of the wage variable.

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. 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 distri-

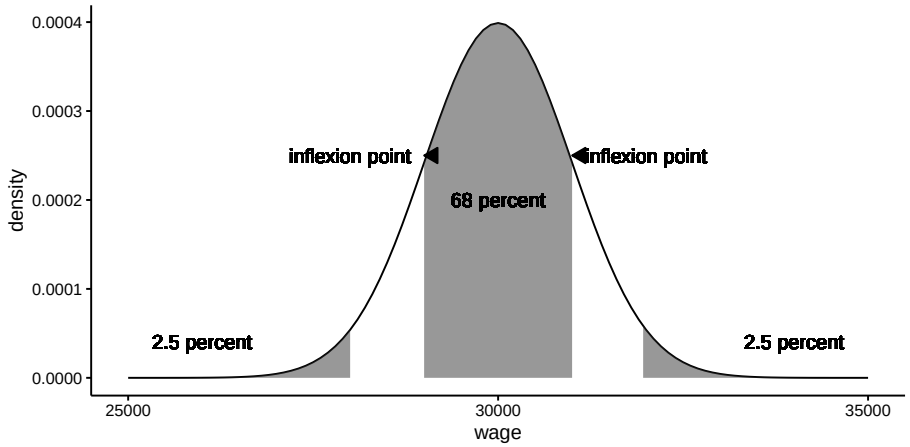


Figure 1.9: The theoretical normal distribution with mean 30,000 and standard deviation 1000.

bution, 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 $1, 1.1, 1.2 \ldots, 3.8, 3.9, \ldots$ 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:

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

[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.

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.

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.

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

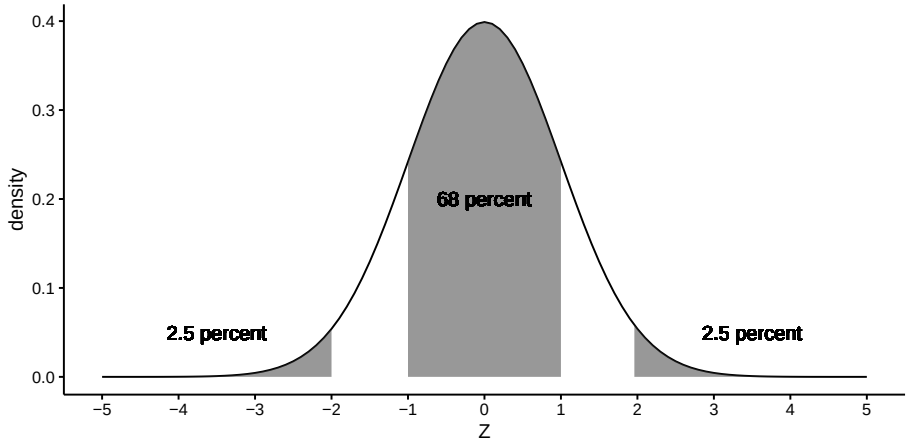


Figure 1.10: The standard normal distribution.

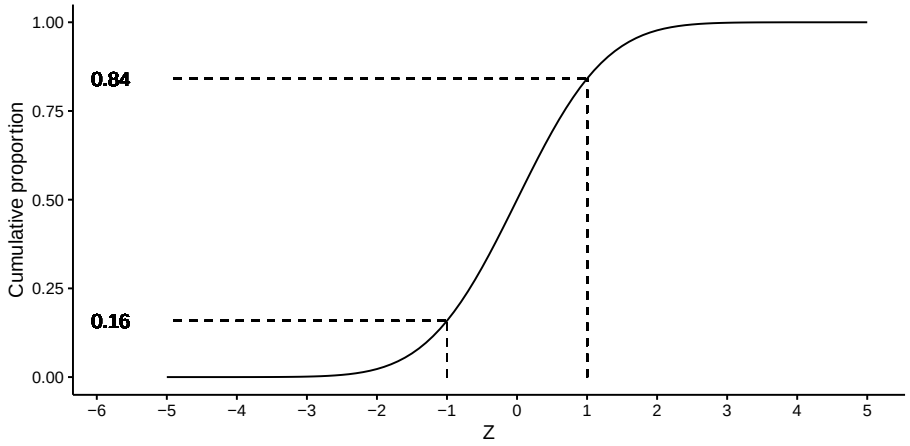


Figure 1.11: The cumulative standard normal distribution.

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*, also called *the empirical 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 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), and 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. The-

oretical 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 6), the chi-square distribution (Chapters 2, 8, **??**, 10 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("")
```

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.

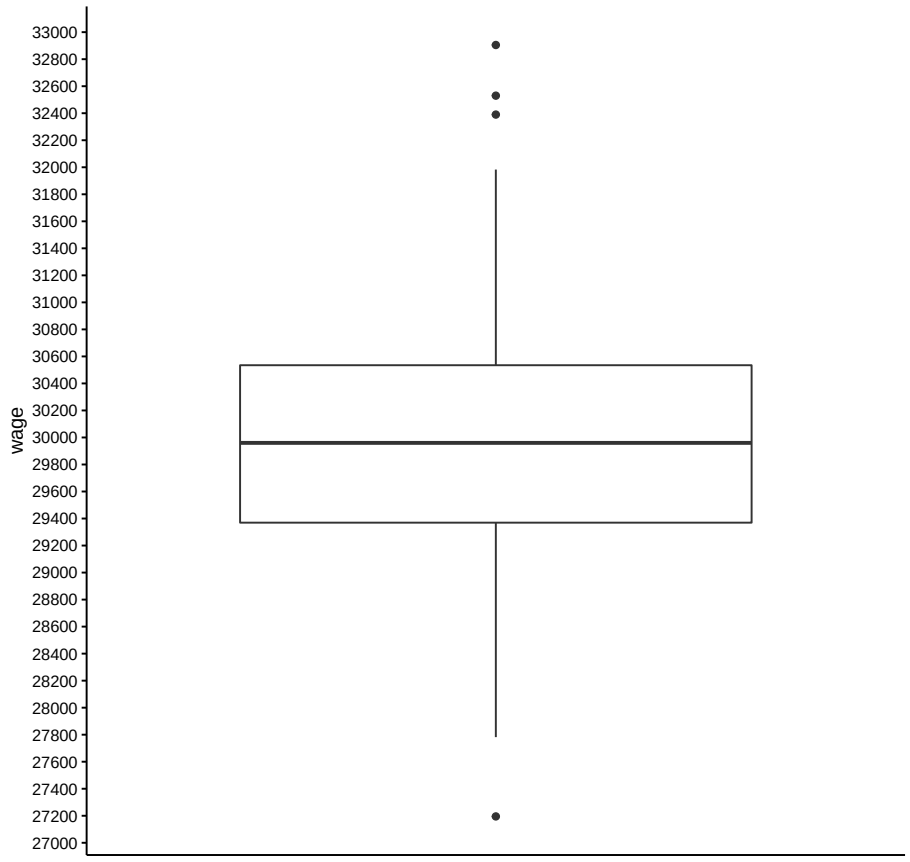


Figure 1.12: A boxplot of the wages earned by a sample of 150 administrative clerks

Table 1.14: A frequency table of nationalities.

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

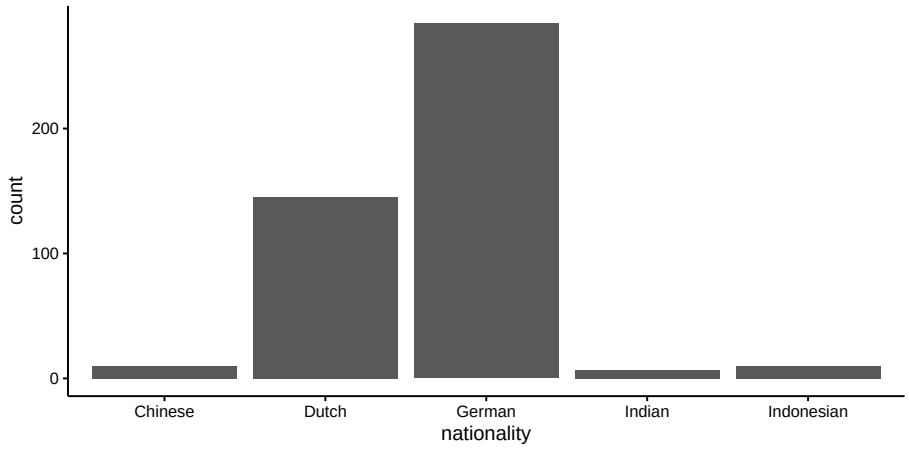


Figure 1.13: A bar chart of the observed nationalities in a lecture hall.

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 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.

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.

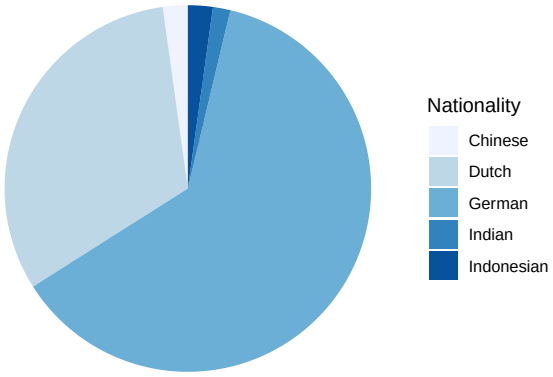


Figure 1.14: A pie chart of nationalities.

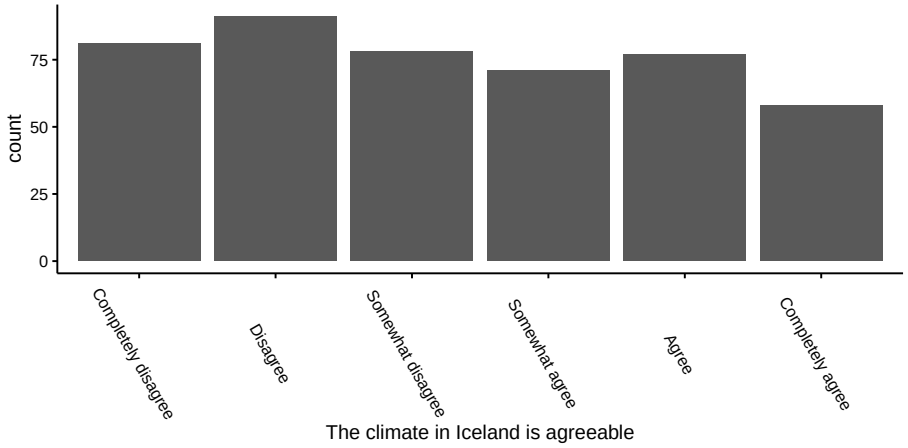
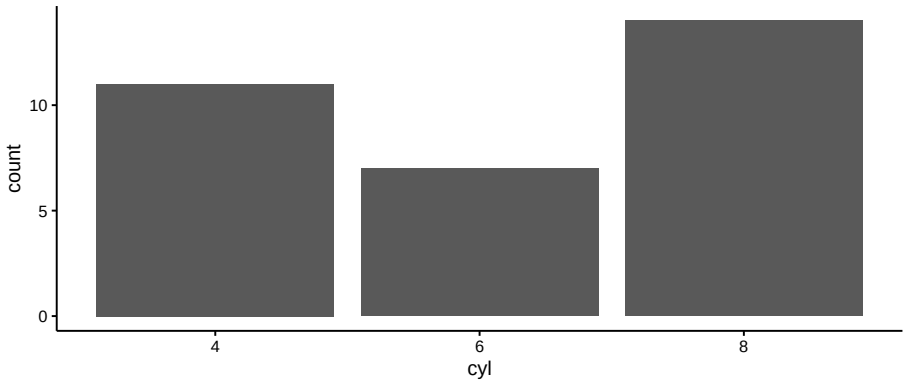


Figure 1.15: Opinions on the climate in Iceland.

1.24 Visualising categorical and ordinal variables in R

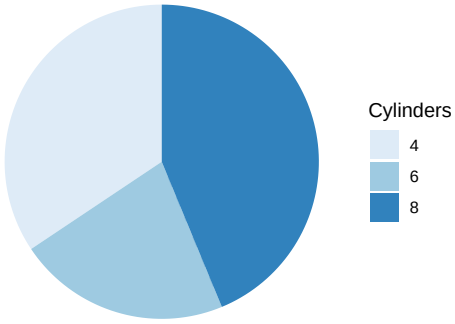
If a categorical variable is stored as numeric, turn it into a factor first. Then R will treat it as categorical. A bar plot with the frequencies on the y-axis can be made with `geom_bar()`:

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



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 length 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 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.

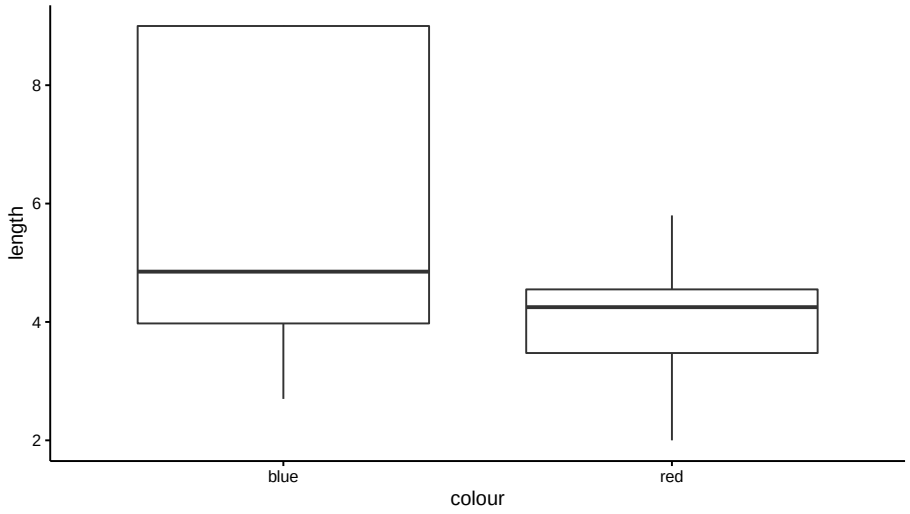


Figure 1.16: A boxplot of the pencil data.

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 `weight` and `length` is easily understood: it appears there is a *linear* relationship between `weight` and `length`. 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 line is shown in Figure 1.19.

You see that by visualising two variables, important patterns may emerge

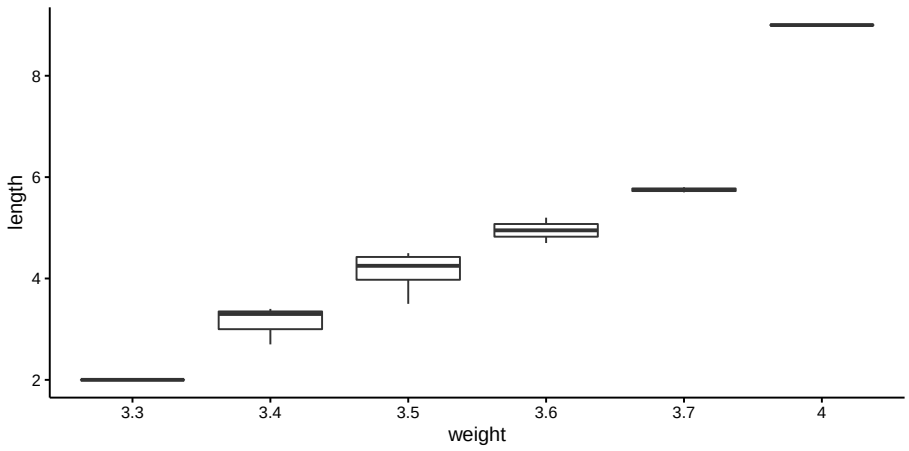


Figure 1.17: A boxplot of the pencil data.

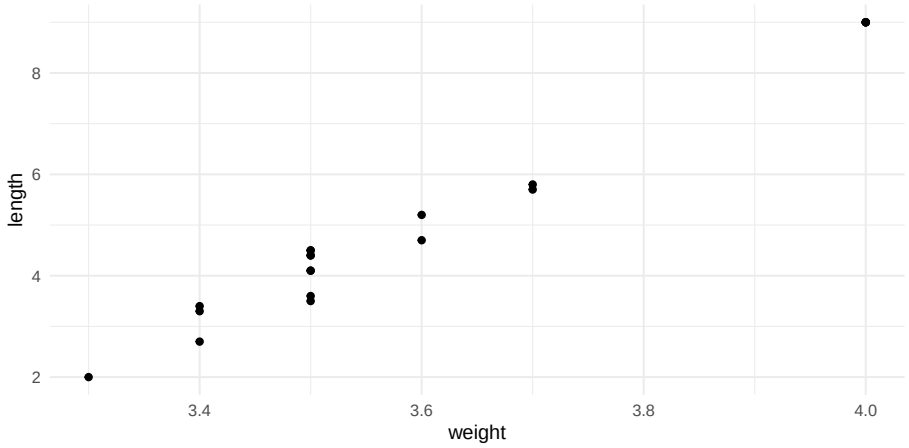


Figure 1.18: A scatterplot of length and weight.

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

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()`:

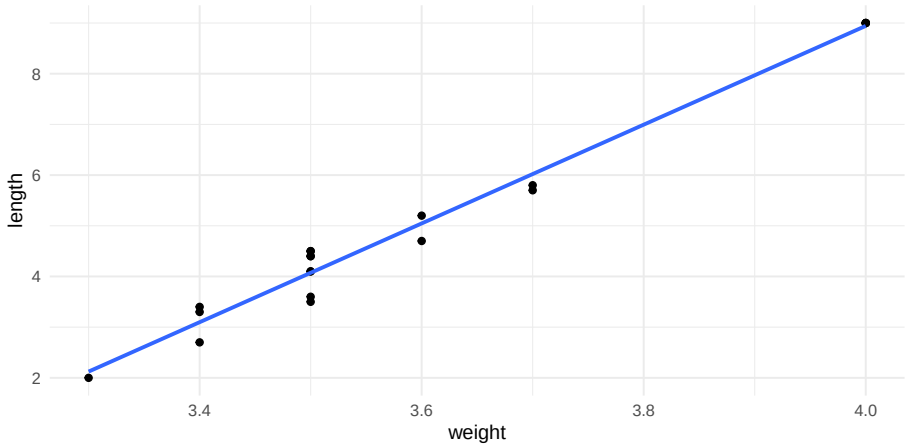


Figure 1.19: A scatterplot of length and weight, with a straight line that summarises the relationship.

```
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), where one variable is the *outcome* variable, and the other variable is a *predictor* variable, that predicts the value on the outcome variable. 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 one outcome variable and more than two

numeric predictor variables (multiple regression). In Chapter 5 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 length 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 6 we will show how we can use straight lines to summarise relationships with predictor variables that we want to treat as categorical.

Chapter 7 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 8 then discusses alternative methods that you can use when linear models are not appropriate.

Chapter 9 focuses on moderation: how one predictor variable can affect the effect that a second predictor 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 predictor variables is a categorical variable.

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 for some data sets. Therefore, Chapter **??** discusses alternative methods to study variables that are repeatedly measured in the same research unit.

Chapters 10 and **??** discuss *generalised linear models*. These are models where the outcome variable is not numeric and continuous. Chapter 10 discusses a method that is appropriate when the outcome variable has only two values, say "yes" and "no", or "pass" and "fail". Chapter **??** discusses a method that can be used when the outcome 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 11 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

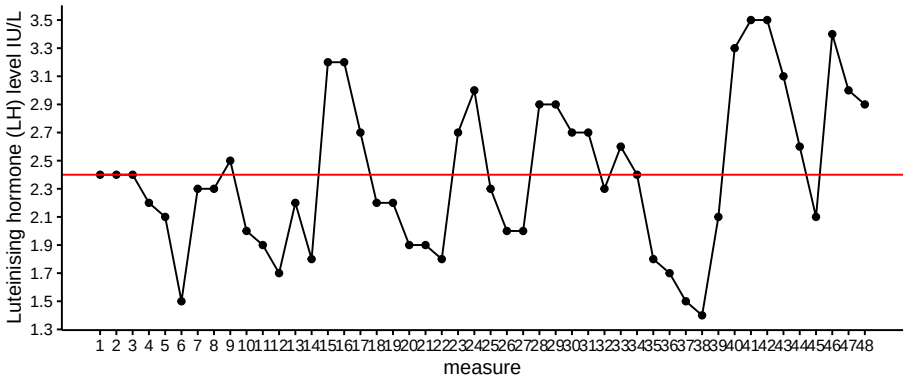


Figure 2.1: Luteinising hormone levels measured in one female, 48 measures taken at 10-minute intervals.

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

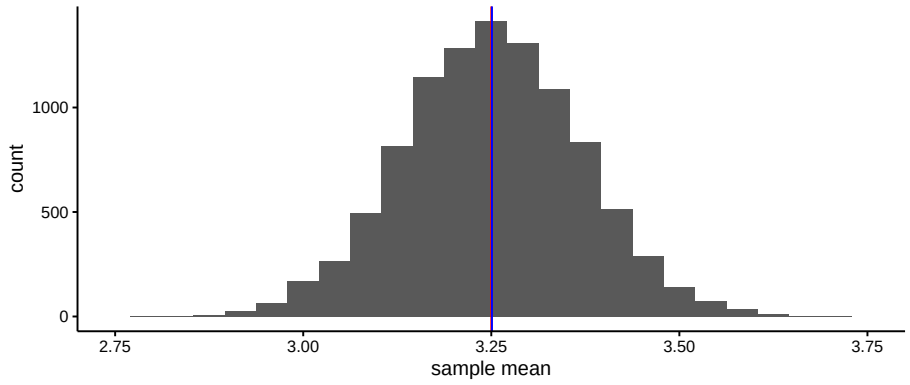


Figure 2.2: A histogram of 10,000 sample means when the sample size equals 10.

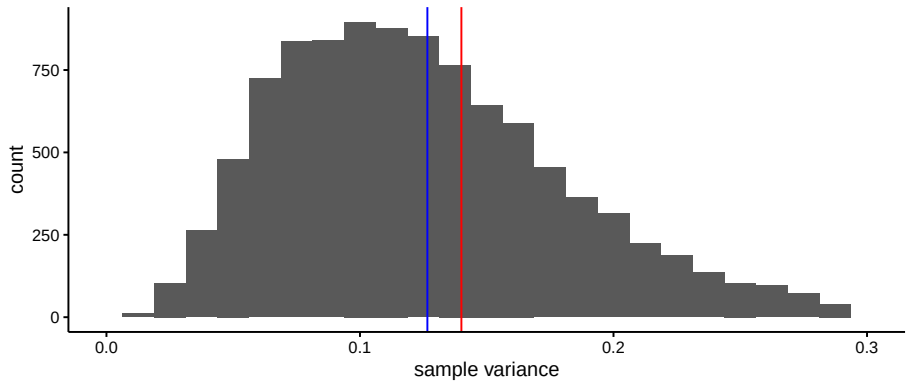


Figure 2.3: A histogram of 10,000 sample variances when the sample size equals 10. The red line indicates the population variance. The blue line indicates the mean of all variances observed in the 10,000 samples.

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 8, **??**, 10 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.

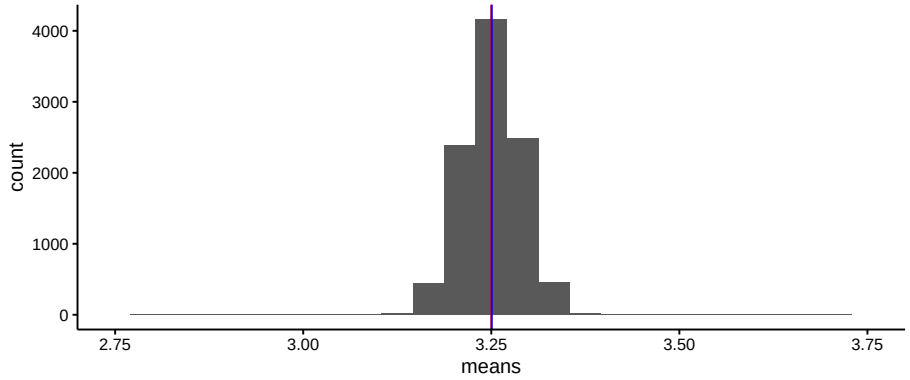


Figure 2.4: A histogram of 10,000 sample means when the sample size equals 100.

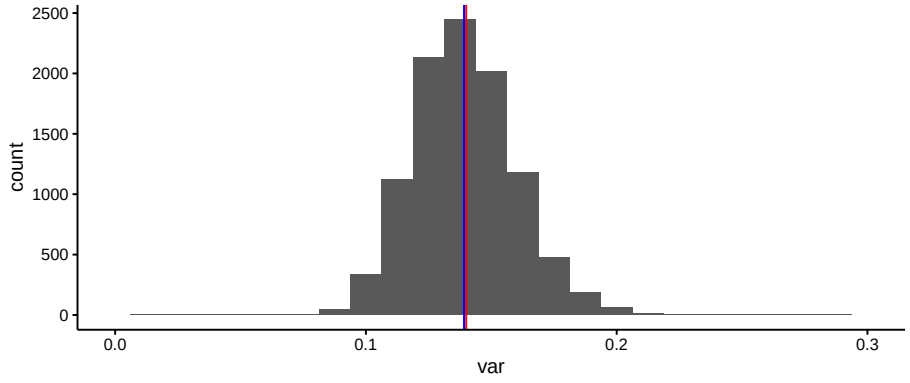


Figure 2.5: A histogram of 10,000 sample variances when the sample size equals 100.

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?

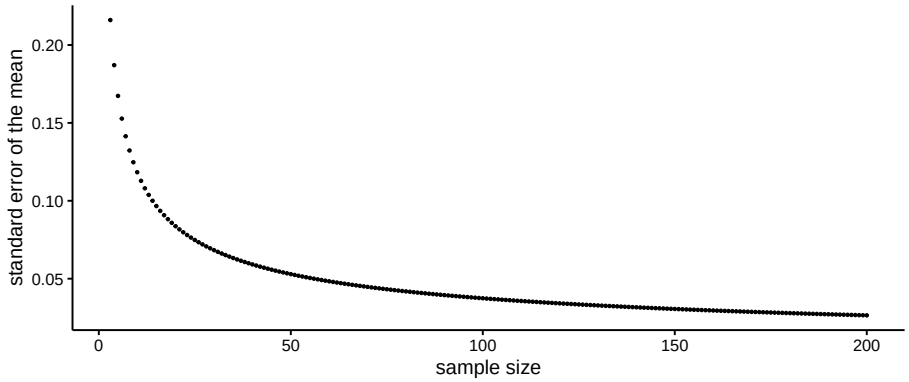


Figure 2.6: Relationship between sample size and the standard error of the mean, when the population variance equals 0.14.

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.

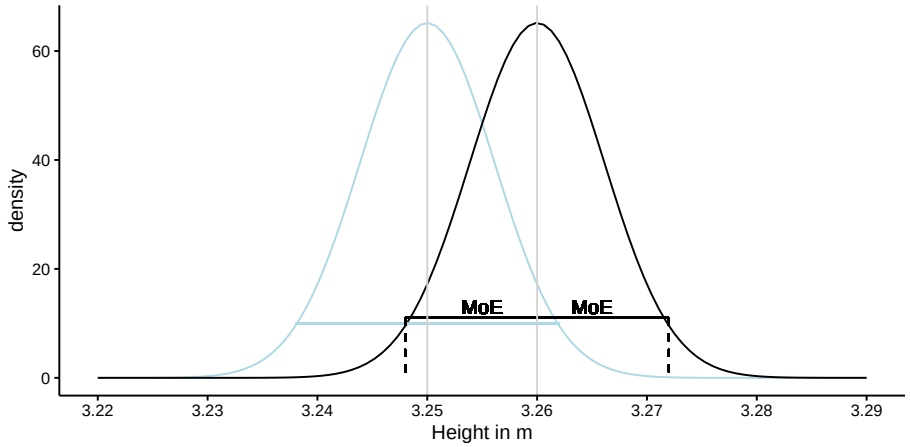


Figure 2.7: Illustration of the construction of a 95% confidence interval. Suppose we find a sample mean of 3.26 and a sample variance of 0.15, with $N = 4000$. The black curve represents the sampling distribution if the population mean would be 3.26 and a variance of 0.15. In reality, we don't know 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 about twice (i.e., 1.96) 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

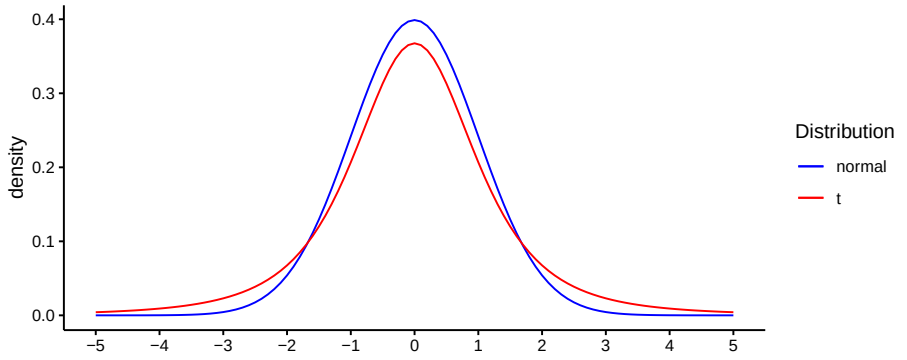


Figure 2.8: Distribution of t with sample size 4, compared with the standard normal distribution.

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

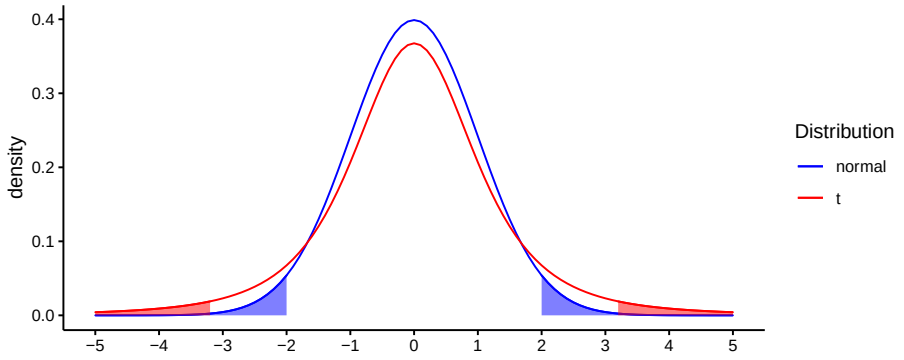


Figure 2.9: Distribution of t with sample size 4, compared with the standard normal distribution. Shaded areas represent 2.5% of the respective distribution.

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

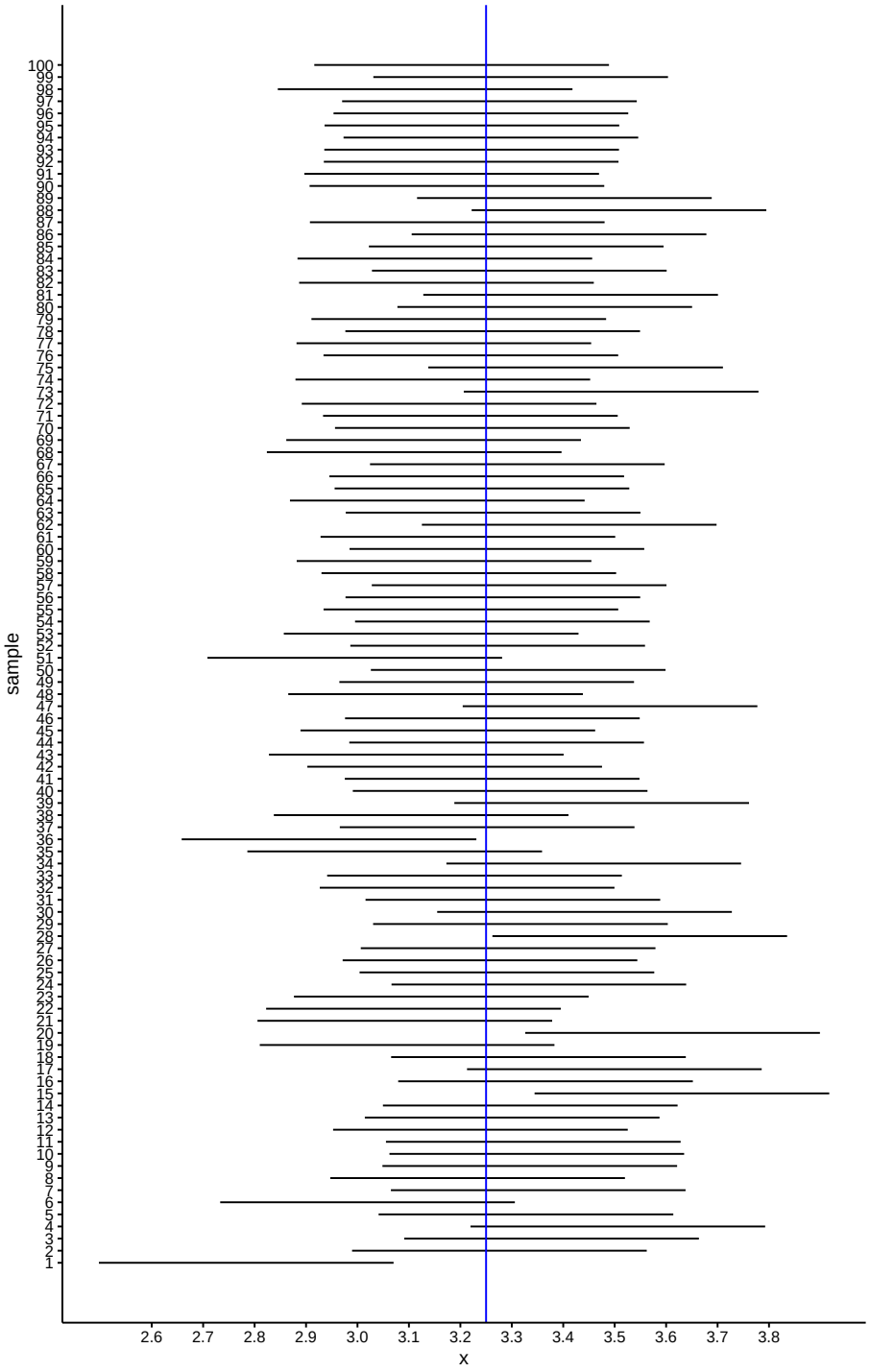


Figure 2.10: Confidence intervals

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.

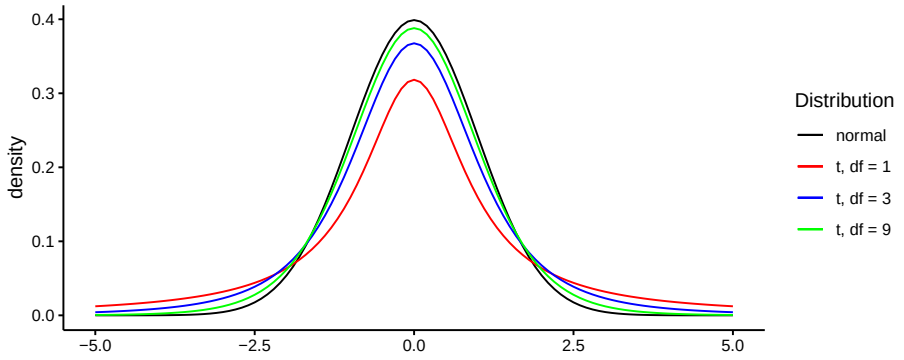


Figure 2.11: Difference in the shapes of the standard normal distribution and t-distributions with 1, 3 and 9 degrees of freedom.

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 check whether 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

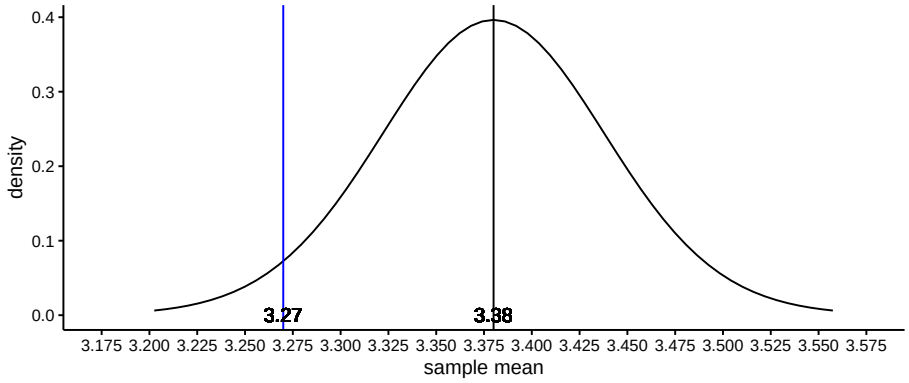


Figure 2.12: The sampling distribution under the null-hypothesis that the South-African population mean is 3.38. The blue line represents the sample mean for our observed sample mean of 3.27.

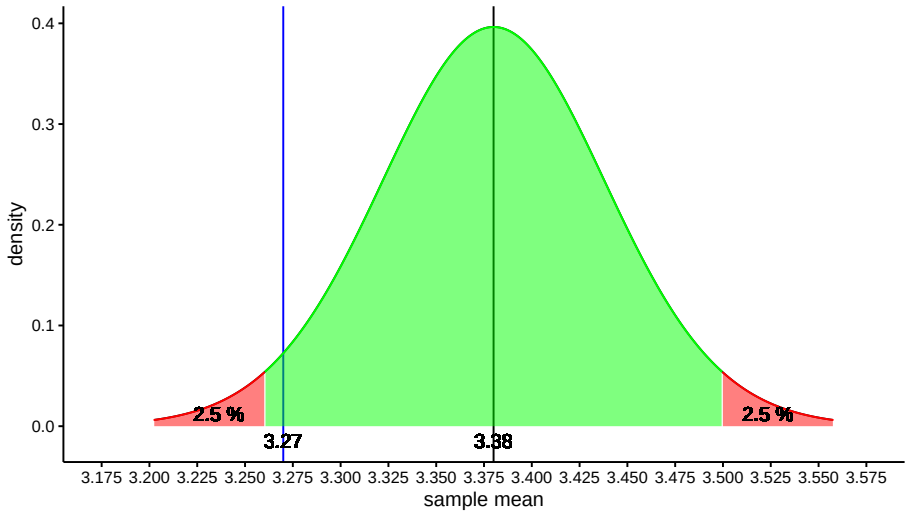


Figure 2.13: The sampling distribution under the null-hypothesis that the South-African population mean is 3.38. The red area represents the range of values for which the null-hypothesis is rejected (rejection region), the green area represents the range of values for which the null-hypothesis is not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

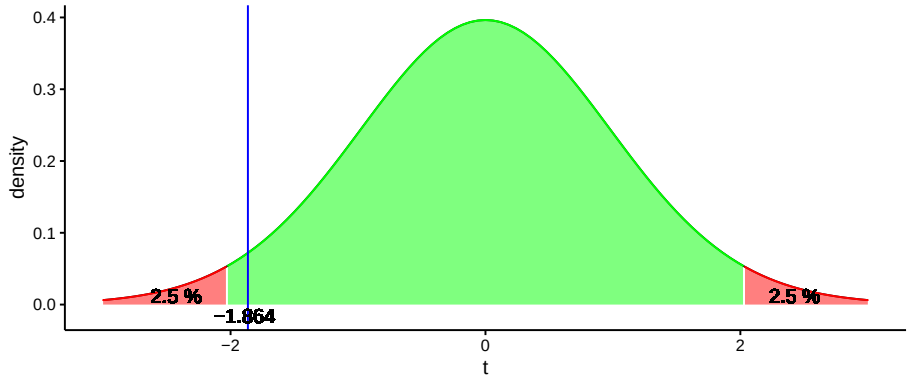


Figure 2.14: A t-distribution with 39 degrees of freedom to test the null-hypothesis that the South-African population mean is 3.38. The blue line represents the T-score for our observed sample mean of 3.27.

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 5% or smaller 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.

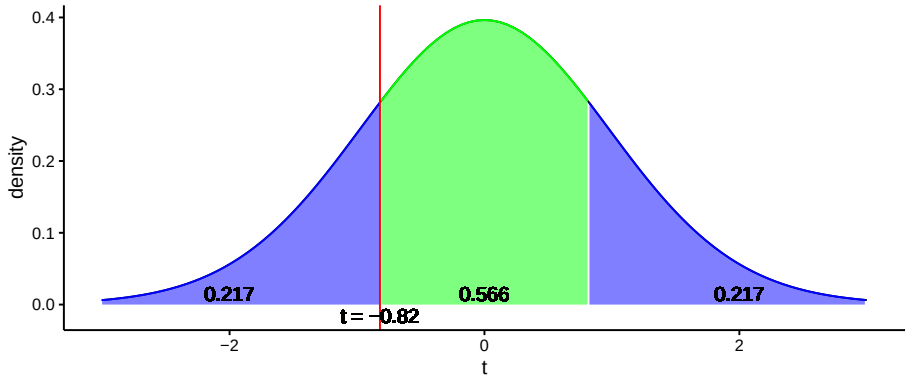


Figure 2.15: Illustration of what a p-value is. The total blue area represents the probability that under the null-hypothesis, you find a more extreme value than the t-score or its opposite. The blue area covers a proportion of .217 + .217 = 0.434 of the t-distribution. This amounts to a p-value of .434.

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.

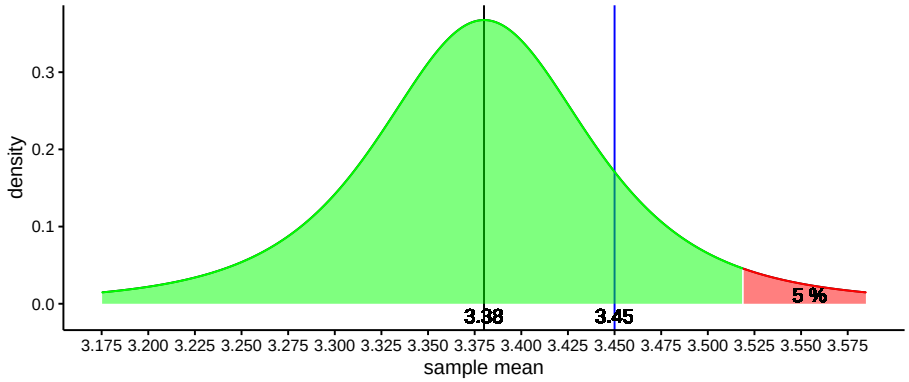


Figure 2.16: The sampling distribution under the null-hypothesis that the South-African population mean is 3.38. In one-tailed testing, the rejection area is located in only one of the tails. The red area represents the range of values for which the null-hypothesis is rejected (rejection region), the green area represents 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

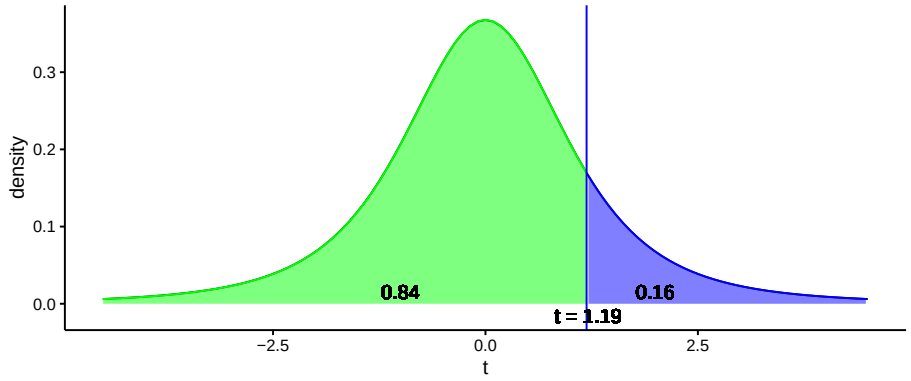


Figure 2.17: The sampling distribution under the null-hypothesis that the South-African population mean is 3.38. For one-tailed testing, the rejection area is located in only one of the tails. The green area represents the probability of seeing a t-value smaller than 1.19, the blue are represents the probability of seeing a t-value larger than 1.19. The latter probability is the p-value.

2.17. The one-tailed p-value for a t-statistic of 1.19 and 3 degrees of freedom 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

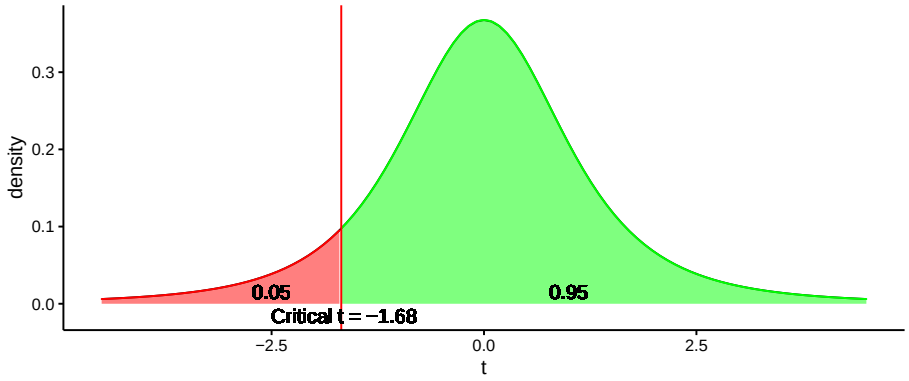


Figure 2.18: One-tailed decision process for deciding whether the average LH level in a woman is too low.

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).

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

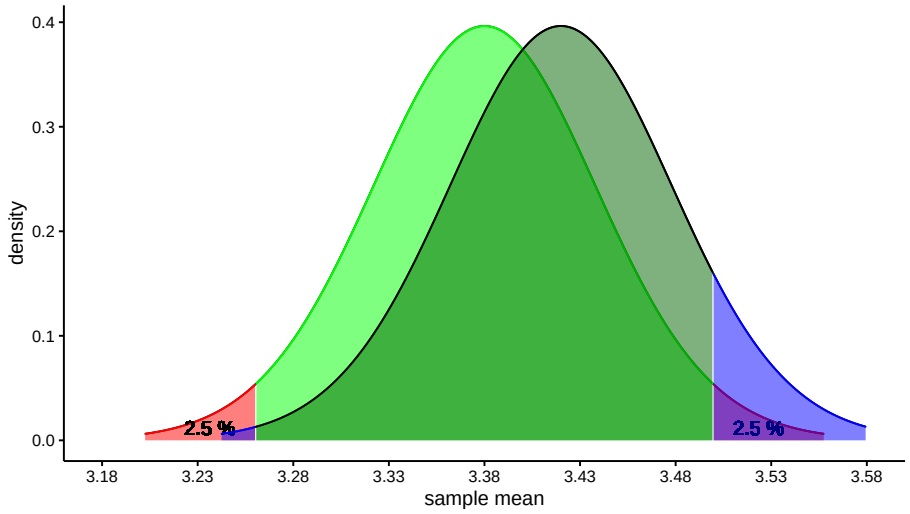


Figure 2.19: Two sampling distributions, one for a population mean of 3.38 (null-hypothesis) and one for a population mean of 3.42 (alternative hypothesis). The red areas represent the probability of a type I error, the dark green area the probability of a type II error. The blue areas represent the 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.

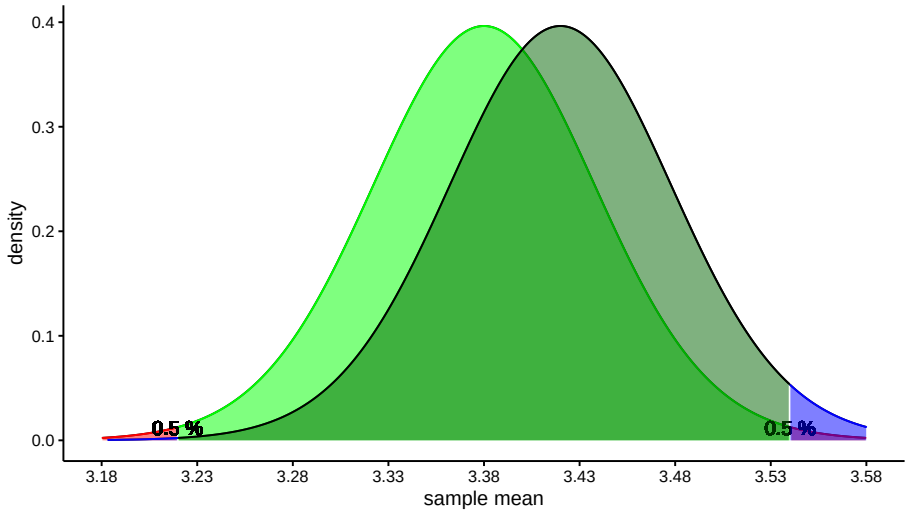


Figure 2.20: Two sampling distributions, one for a population mean of 3.38 (null-hypothesis) and one for a population mean of 3.42 (alternative hypothesis). The red areas represent the probability of a type I error, the dark green area the probability of a type II error. The blue area represents the 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.

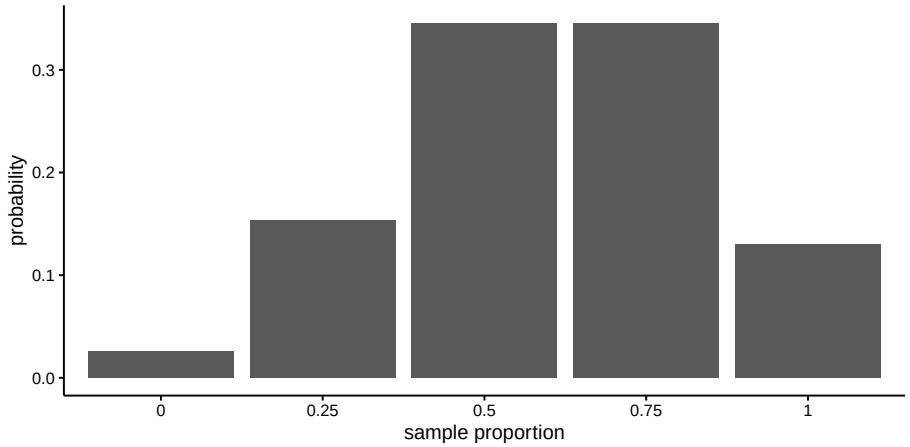


Figure 3.1: Sampling distribution of the sample proportion, when the population proportion is 0.60

Actually, the sampling distribution that we see in Figure 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 to denote a tall elephant and B to denote a short elephant, the 6 possible combinations of having two As and two Bs are in fact: AABB, BBAA, ABAB, BABA, ABBA, and BAAB.

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

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

This number $\binom{4}{2}$ ('four choose two') 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 combinations, 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 $\binom{n}{r} = \frac{n!}{r!(n-r)!}$.

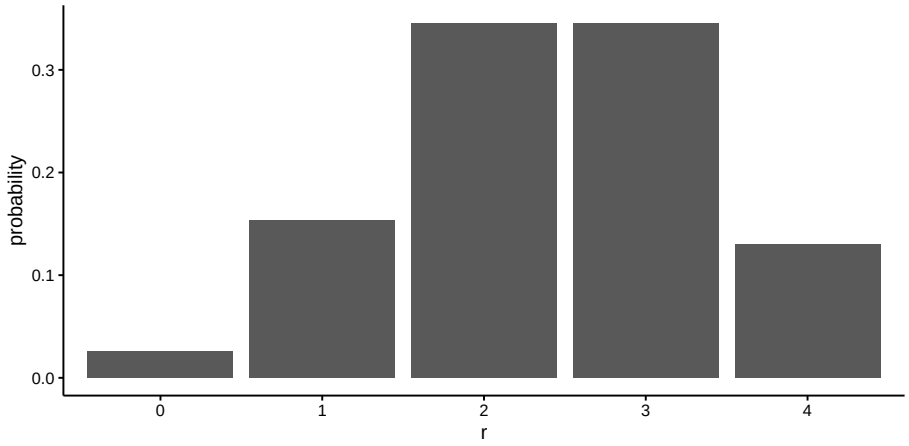


Figure 3.2: Binomial distribution with $N = 4$ and $p = 0.60$.

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

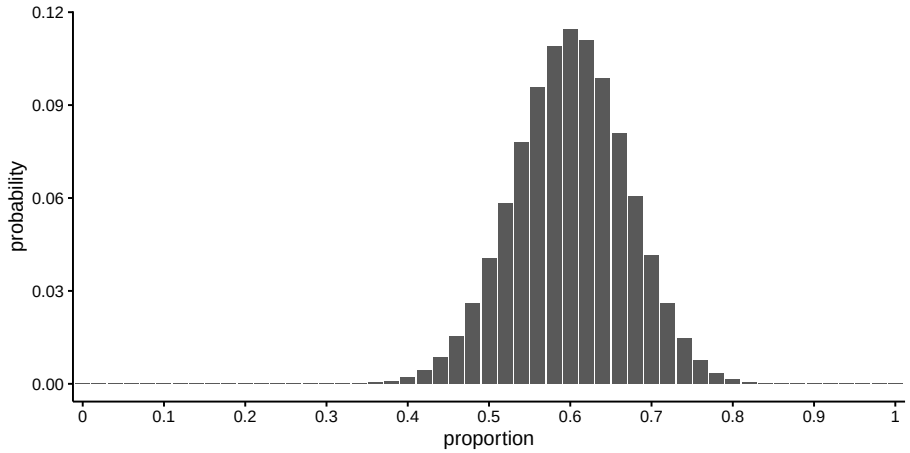


Figure 3.3: Sampling distribution with $N = 50$ and $p = 0.60$.

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

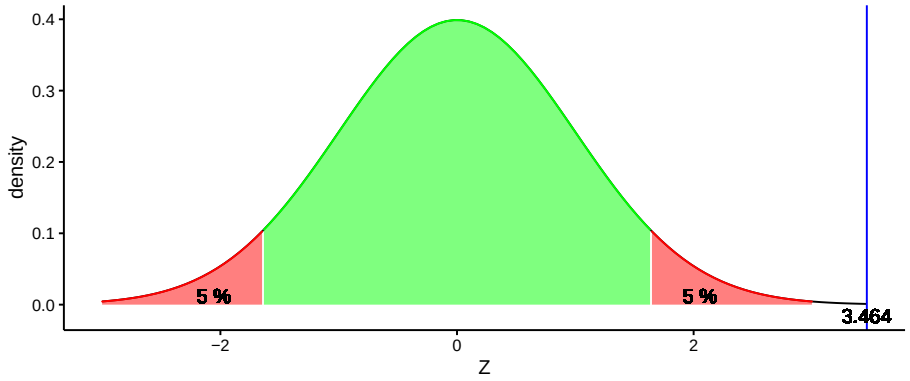


Figure 3.4: A normal distribution to test the null-hypothesis that the population proportion is 0.6. The blue line represents the z-score for our observed sample proportion of 0.84.

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

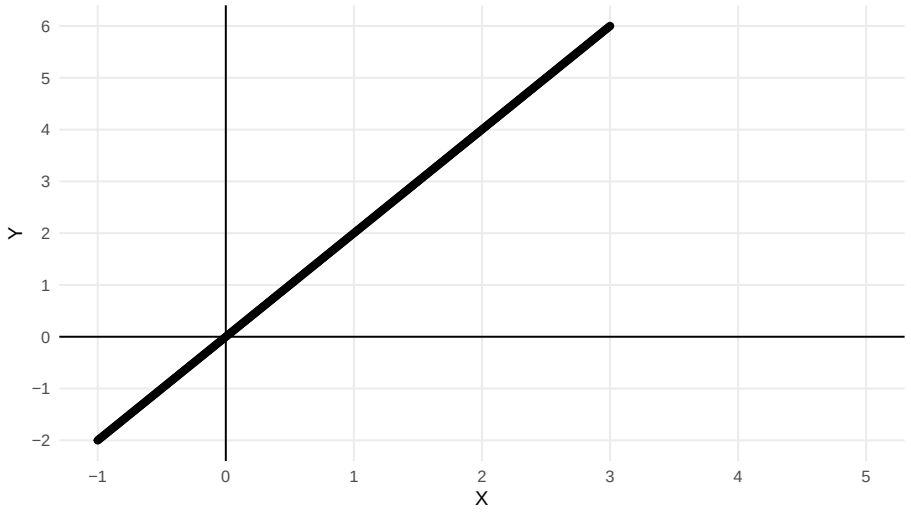


Figure 4.1: Straight line with intercept 0 and slope 2.

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.

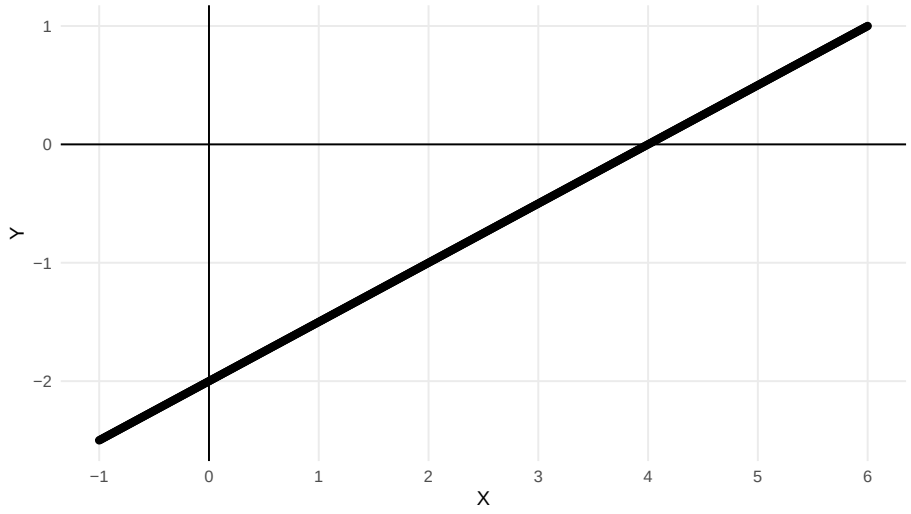


Figure 4.2: Straight line with intercept -2 and slope 0.5.

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:

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

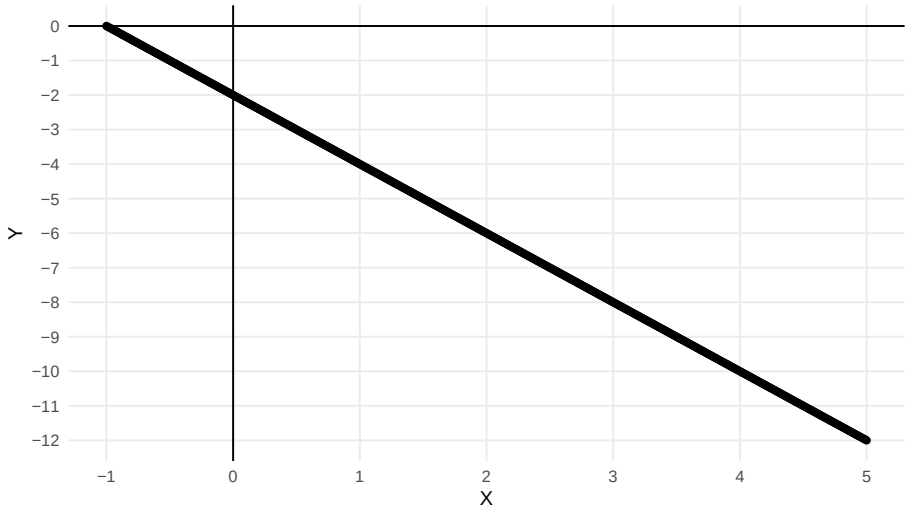


Figure 4.3: Straight line with intercept -2 and slope -2.

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

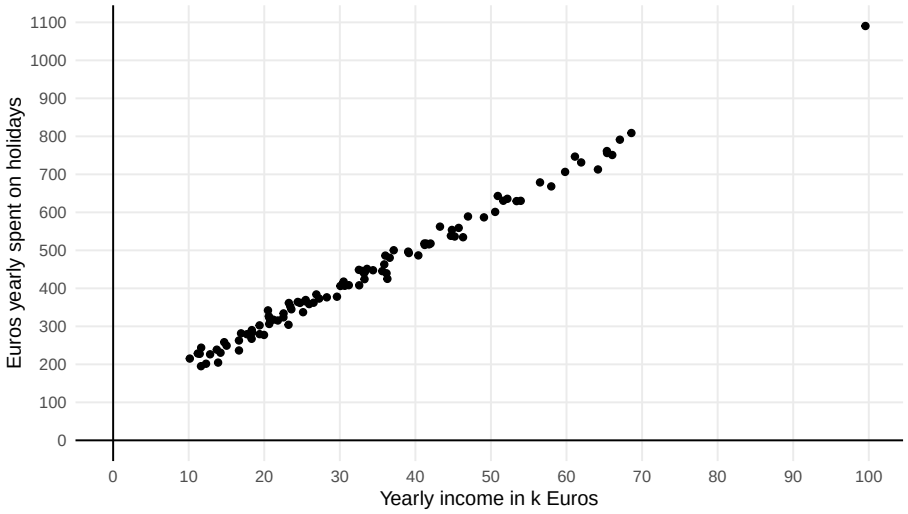


Figure 4.4: Data on holiday spending.

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

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

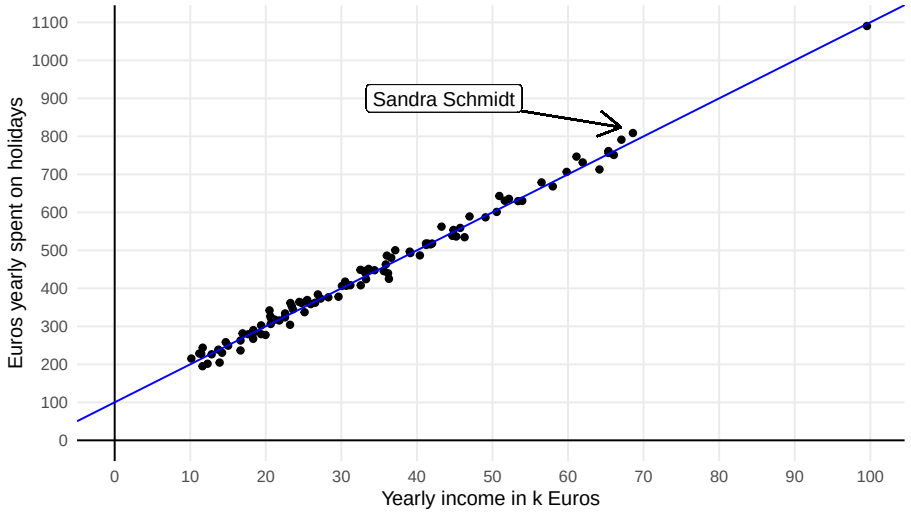


Figure 4.5: Data on holiday spending with an added straight line.

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.

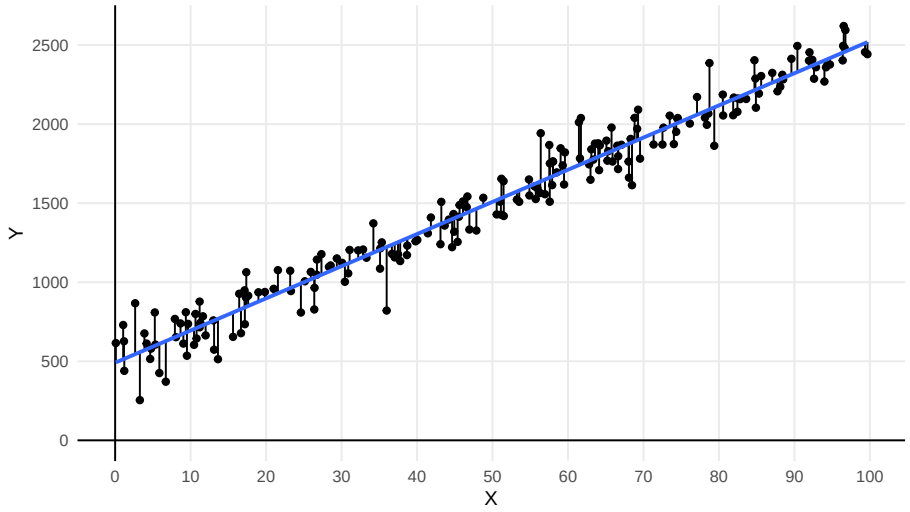


Figure 4.6: Data on variables X and Y with an added straight line.

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).

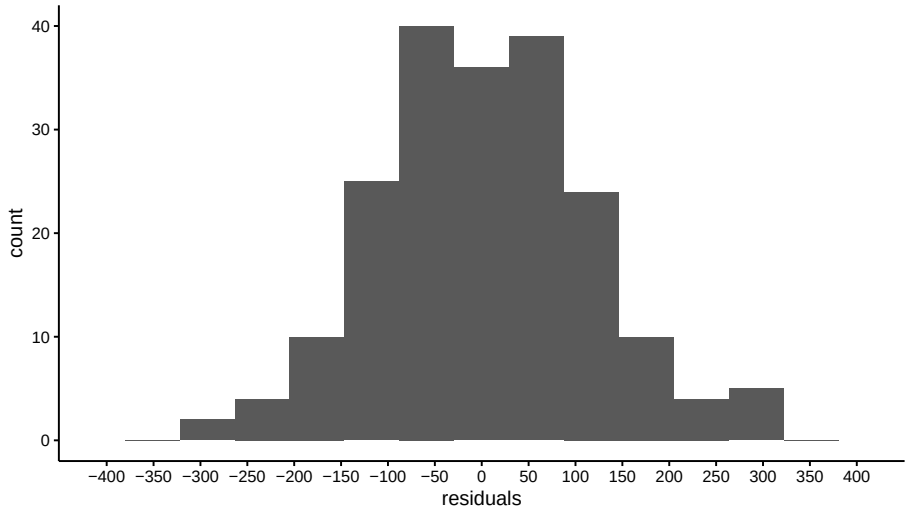


Figure 4.7: Histogram of the residuals (errors).

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.

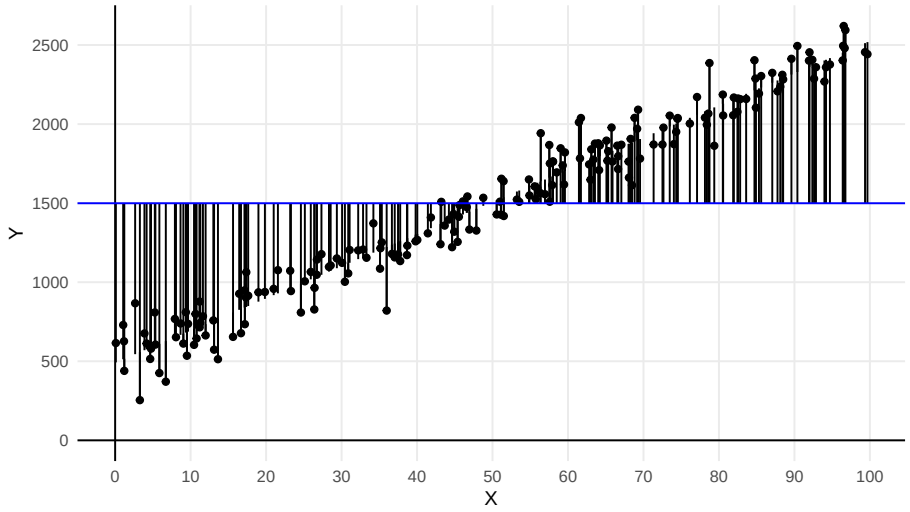


Figure 4.8: Data on variables X and Y with an added straight line. The sum of the residuals equals 0.

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.

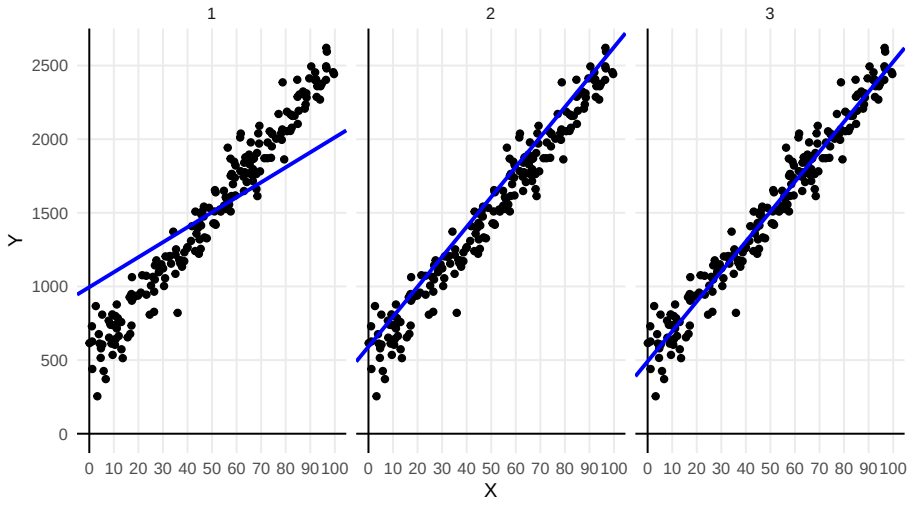


Figure 4.9: Three times the same data set, but with different regression lines.

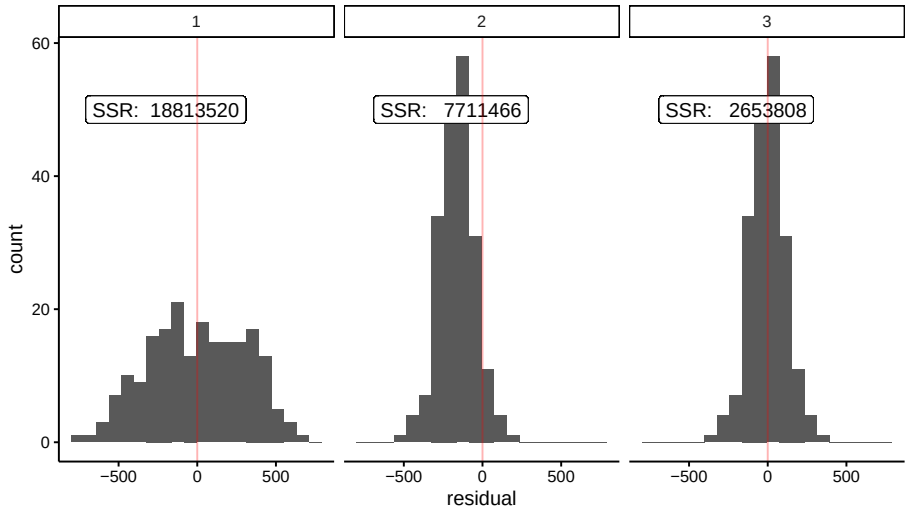


Figure 4.10: Histogram of the residuals (errors) for three different regression lines, and the respective sums of squared residuals (SSR).

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.

$$\begin{aligned} Y_i &= b_0 + b_1 X_i + e_i && (4.12) \\ e_i &\sim N(0, \sigma_e^2) && (4.13) \end{aligned}$$

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 5.

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

In the syntax we first indicate that we start from the `mtcars` data set. Next, we use the `lm()` function to indicate that we want to apply the linear model to these data. 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, because we

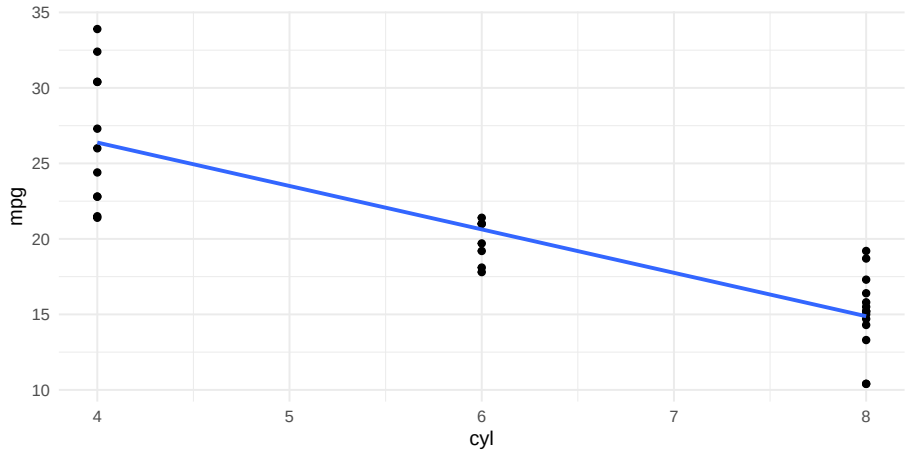


Figure 4.11: Data set on number of cylinders (`cyl`) and miles per gallon (`mpg`) in 32 cars.

already indicated we use the `mtcars` data set, the `data` argument for the `lm()` function should be left empty. Finally, we store the results in the object `model`.

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

```
model

##
## Call:
## lm(formula = mpg ~ cyl, data = .)
##
## 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 Equations 4.12 and 4.13 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)
model <- mtcars %>%
  lm(mpg ~ cyl, data = .)
model %>%
  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 Chapter 5.

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

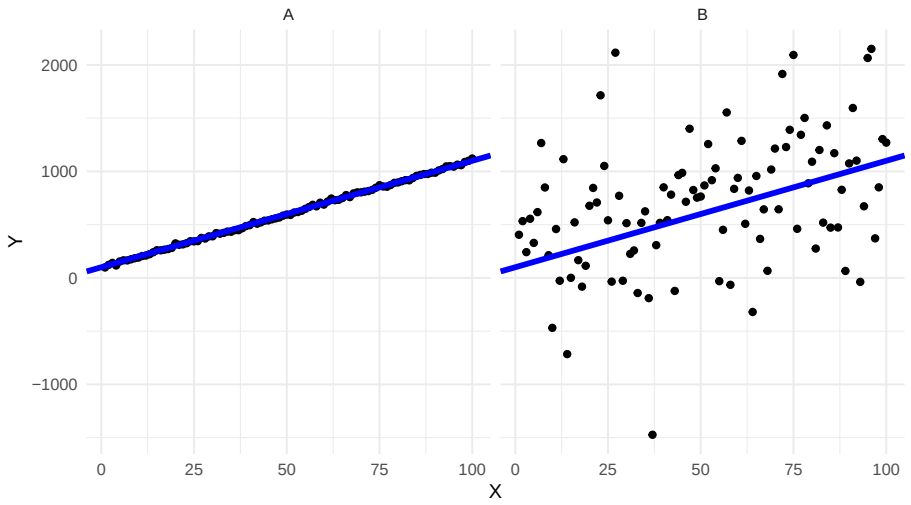


Figure 4.12: Two data sets with the same regression line.

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 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

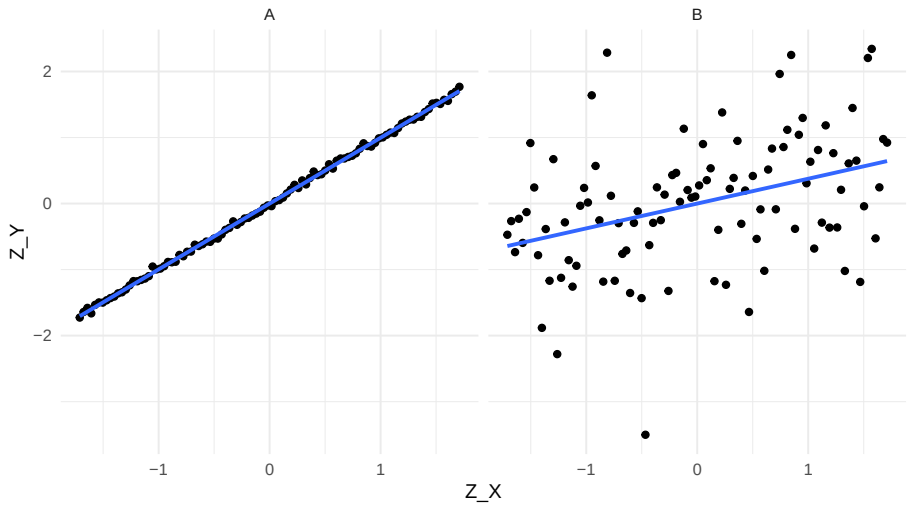


Figure 4.13: Two data sets, with different regression lines after standardisation.

the slopes are different: in data set A, the slope is 0.997 and in data set B, the slope is 0.376.

$$Z_Y = 0 + 0.997 \times Z_X = 0.997 \times Z_X \tag{4.17}$$
$$Z_Y = 0 + 0.376 \times Z_X = 0.376 \times Z_X \tag{4.18}$$

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* or *standardised regression 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 4.14 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.

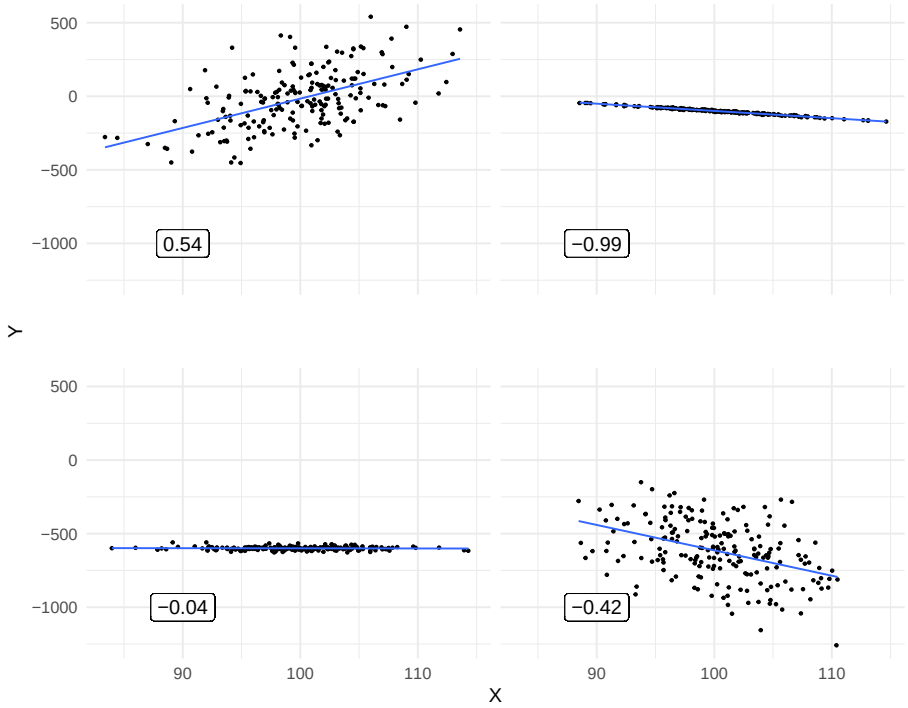


Figure 4.14: Various plots showing different correlations between variables X and Y.

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. It is standardised in such a way that it can never be outside the (-1, 1) interval. This standardisation happened through the division of X and Y-values by their respective standard deviation. 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 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

[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 $n - 1$

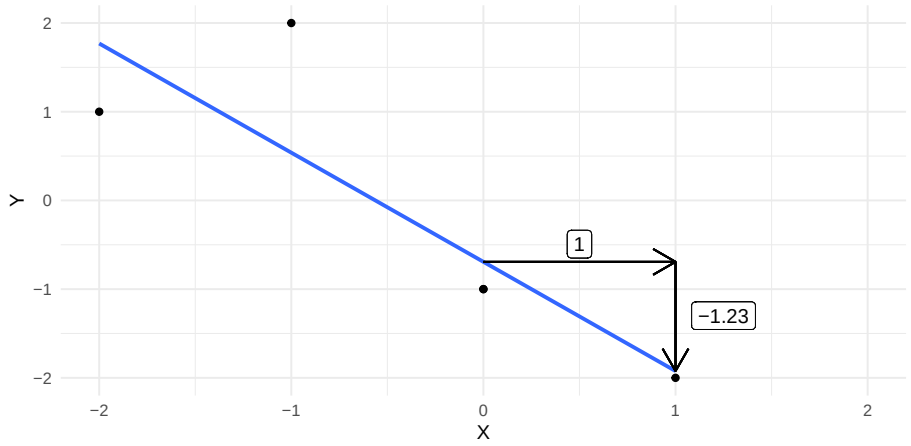


Figure 4.15: Data example and the regression line.

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: 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 4.15 together with the regression line. The standardised data and the corresponding regression line are plotted in Figure 4.16. 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()`:

```
mtcars %>%
  select(cyl, mpg) %>%
  cor()

##           cyl       mpg
```

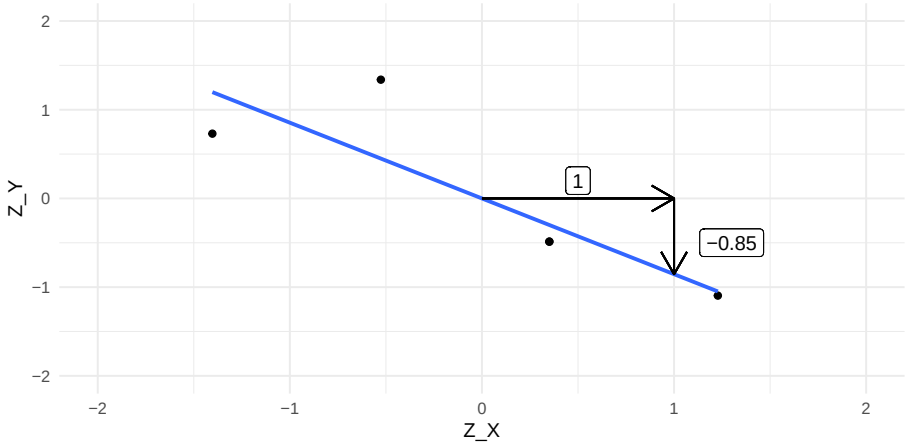


Figure 4.16: Data example (standardised values) and the regression line.

```
## cyl  1.000000 -0.852162
## mpg -0.852162  1.000000
```

In the output we see a correlation matrix. On the diagonal are the correlations of `cyl` and `mpg` with themselves, which are perfect (a correlation of 1). On the off-diagonal, we see that the correlation between `cyl` and `mpg` equals -0.852162. This is a strong negative correlation, which means that generally, the more cylinders a car has, the lower the mileage. We can also compute the covariance, with the function `cov`:

```
mtcars %>%
  select(cyl, mpg) %>%
  cov()

##           cyl       mpg
## cyl  3.189516 -9.172379
## mpg -9.172379 36.324103
```

On the off-diagonal we see that the covariance between `cyl` and `mpg` equals -9.172379. On the diagonal we see the variances of `cyl` and `mpg`. Note that R uses the formula with $n - 1$ in the denominator. If we want R to compute the (co-)variance 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 5.1):

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

## [1] -2.87579
```

Note that both `cov()` and `var()` use $n - 1$. Since this cancels out if we do the division, it doesn't matter whether we use n or $n - 1$.

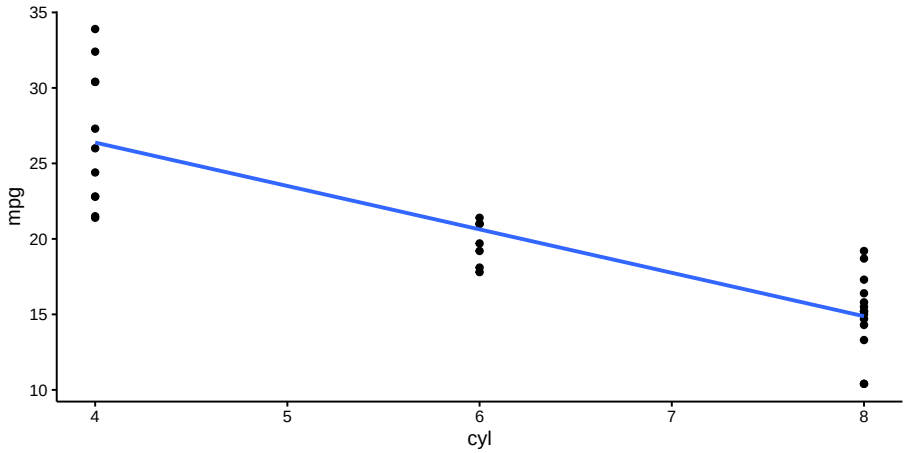
If we first standardise the data with the function `scale()` and then compute the least squares slope, we get

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

cov(z_mpg, z_cyl) / var(z_cyl)

##           [,1]
## [1,] -0.852162

cor(z_mpg, z_cyl)

##           [,1]
## [1,] -0.852162

cov(z_mpg, z_cyl)

##           [,1]
## [1,] -0.852162
```

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.

The data and the least squares regression line can be plotted using `geom_smooth()`:

```
mtcars %>%
  ggplot(aes(x = cyl, y = mpg)) +
  geom_point() +
  geom_smooth(method = "lm", se = F)
```



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. These data are available in R, and we will show how to perform the followint analyses in a later section. 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:

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

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, $\text{Var}(e)$, 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_Y^2 and σ_e^2. We should not use the variance of Y observed in the sample, but the unbiased estimator of the variance of Y in the population

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

where n is sample size (see Section 2.3).

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

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

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

$$\begin{aligned} \widehat{R^2} &= \frac{\widehat{\sigma_Y^2} - \widehat{\sigma_e^2}}{\widehat{\sigma_Y^2}} = \frac{\frac{\Sigma(Y_i - \overline{Y})^2}{n-1} - \frac{\Sigma e_i^2}{n-1}}{\frac{\Sigma(Y_i - \overline{Y})^2}{n-1}} \\ &= \frac{\Sigma(Y_i - \overline{Y})^2 - \Sigma e_i^2}{\Sigma(Y_i - \overline{Y})^2} = \frac{\Sigma(Y_i - \overline{Y})^2}{\Sigma(Y_i - \overline{Y})^2} - \frac{\Sigma e_i^2}{\Sigma(Y_i - \overline{Y})^2} \\ &= 1 - \frac{SSR}{SST} \end{aligned} \tag{4.37}$$

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 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.

[2]In the literature and online, sometimes you see SSR and sometimes you see SSE, both referring to the sum of the squared residuals

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)
library(broom)
model <- allbacks %>%
lm(weight ~ volume + area, data = .)
model %>%
tidy()
```

Below we see the output:

```
library(DAAG)
library(broom)
model <- allbacks %>%
  lm(weight ~ volume + area, data = .)
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 22.4. 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:

$$\texttt{weight} = 22.4 + 0.708 \times \texttt{volume} + 0.468 \times \texttt{area} + e \tag{4.38}$$

Thus, the predicted weight of a book that has a volume of 10 and an area of 5, the expected weight is equal to $22.4 + 0.708 \times 10 + 0.468 \times 5 = 31.82$.

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

```
library(DAAG)
data(cfseal) # available in package DAAG
out1 <- cfseal %>%
  lm(age ~ weight , data = .)
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 $(1090.8551724 - 200)/1090.8551724 = 0.8166576$.

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

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

```
out2 <- cfseal %>%
  lm(age ~ heart , data = .)
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

sum2 <- out2 %>%
  summary()
sum2$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 <- cfseal %>%
  lm(age ~ heart + weight , data = .)
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

sum3 <- out3 %>% summary()
sum3$sigma^2  # variance of residuals

## [1] 203.55
```

with model equation:

$$\texttt{age} = 10.3 - 0.03 \times \texttt{heart} + 0.99 \times \texttt{weight} + e \quad (4.46)$$

$$e \sim N(0, 204) \quad (4.47)$$

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.

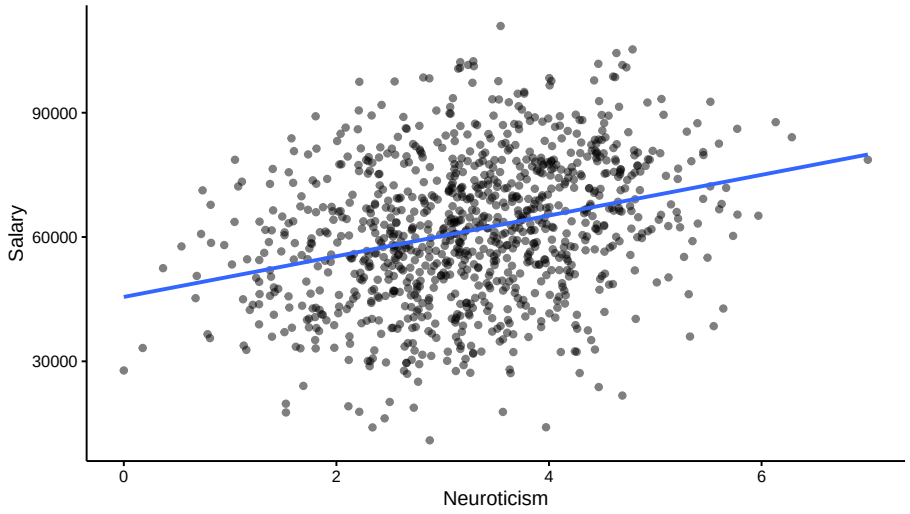


Figure 4.17: Simulated HR data set.

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 4.17 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 4.18: 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

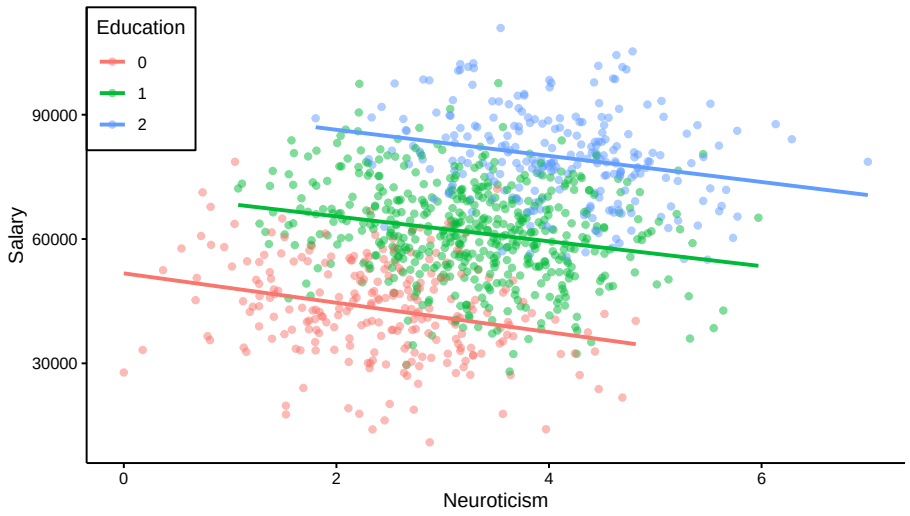


Figure 4.18: Same HR data, now with markers for different education levels.

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 4.19. If you look at all the data points at once, you see a steep slope in the least squares regression line for 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

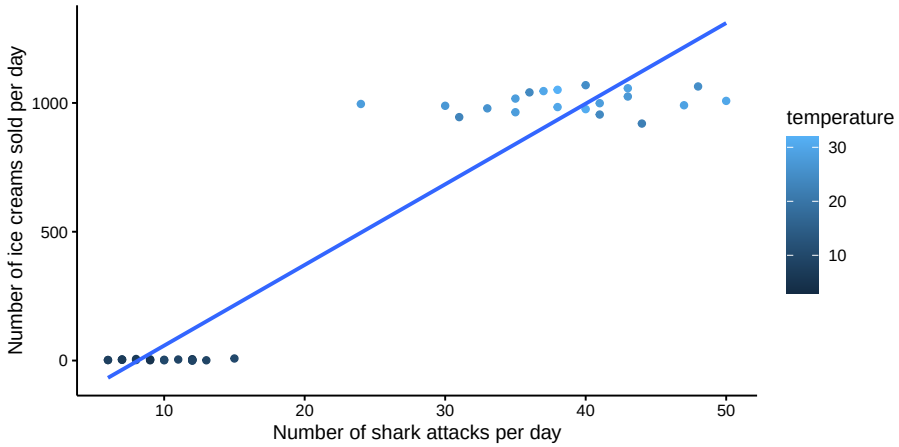


Figure 4.19: A spurious correlation between the number of shark attacks and ice cream sales.

the dark blue data points (low temperatures).

Chapter 5

Inference for linear models

In Chapter 4 on regression we saw how a linear equation can describe a data set: the linear equation describes the behaviour of one variable, the dependent variable, on the basis of one or more other variables, the independent variable(s). Sometimes we are indeed interested in the relationship between two variables in one given data set. For instance, a primary school teacher wants to know how well the exam grades in her class of last year predict how well the same students do on another exam a year later.

But very often, researchers are not interested in the relationships between variables in one data set on one specific group of people, but interested in the relationship between variables in general, not limited to only the observed data. For example, a researcher would like to know what the relationship is between the temperature in a brewery and the amount of beer that goes into the beer bottles. In order to study the effect of temperature on volume, the researcher measures the volume of beer in a limited collection of 200 bottles under standard conditions of 20 degrees Celsius and determines from log files the temperature in the factory during production for each measured bottle. The linear equation might be $\texttt{volume} = 32.35 - 0.1207 \times \texttt{temp} + e$, see Figure 5.1. Thus, for every unit increase in degrees Celsius, say from 20 to 21 degrees, the volume of beer that is measured increases by -0.1207 centilitres, or put differently, the volume of beer decreases by 0.1207.

But the researcher is not at all interested in these 200 bottles specifically: the question is what would the linear equation be if the researcher had used information about *all* bottles produced in the same factory? In other words, we may know about the linear relationship between temperature and volume in a *sample* of bottles, but we might really be interested to know what the relationship would look like *had we been able to measure the volume in the population of all bottles.*

In this chapter we will see how to do inference in the case of a linear model. Many important concepts that we already saw in earlier chapters will be mentioned again. Some repetition of those rather difficult concepts will be helpful, especially when now discussed within the context of linear models.

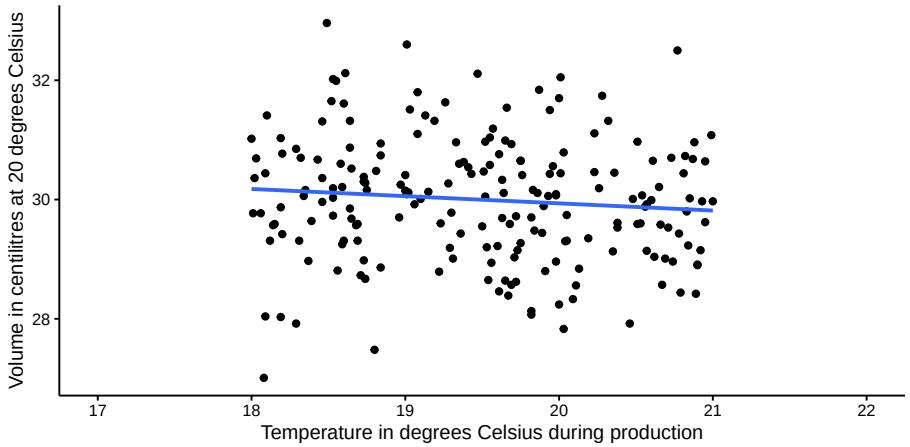


Figure 5.1: The relationship between temperature and volume in a sample of 200 bottles.

5.1 Population data and sample data

In the beer bottle example above, the volume of beer was measured in a total of 200 bottles. Let's do a thought experiment, similar to the one in Chapter 2. Suppose we could have access to volume data about all bottles of beer on all days on which the factory was operating, including information about the temperature for each day of production. Suppose that the total number of bottles produced is 80,000 bottles. When we plot the volume of each bottle against the temperature of the factory we get the scatter plot in Figure 5.2.

In our thought experiment, we could determine the regression equation using all bottles that were produced: all 80,000 of them. We then find the blue regression line displayed in Figure 5.2. Its equation is $\texttt{volume} = 29.98 + 0.001 \times \texttt{temp} + e$. Thus, for every unit increase in temperature, the volume increases by 0.001 centilitres. Thus, the slope is slightly positive in the population, but negative in the sample of 200 bottles.

In the data example above, data were only collected on 200 bottles. These bottles were randomly selected[1]: there were many more bottles but we could measure only a limited number of them. This explains why the regression equation based on the sample differed from the regression equation based on all bottles: we only see part of the data.

Here we see a discrepancy between the regression equation based on the sample, and the regression equation based on the population. We have a slope of 0.001 in the population, and we have a slope of -0.1207 in the sample. Also the intercepts differ. To distinguish between the coefficients of the population and coefficients of the sample, a population coefficient is often denoted by the

[1]Random selection means that each of the 80,000 bottles had an equal probability to end up in this sample of 200 bottles.

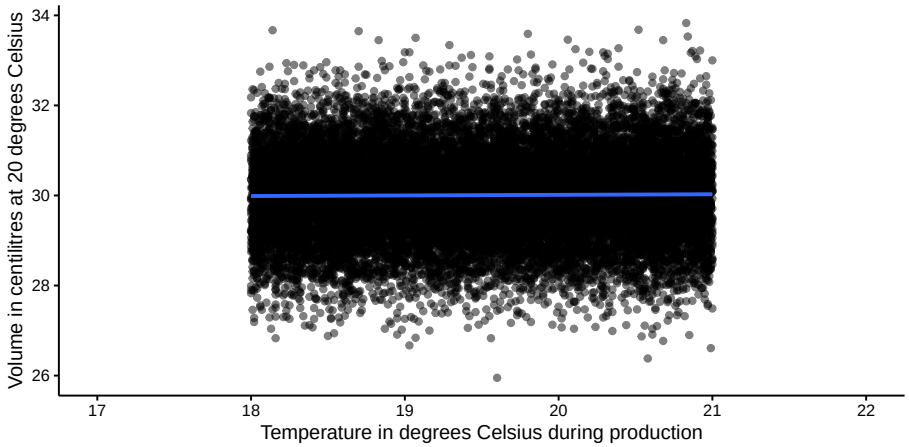


Figure 5.2: The relationship between temperature and volume in all 80,000 bottles.

Greek letter β and a sample coefficient by the Roman letter b.

$$\begin{aligned} \textit{Population} : \texttt{volume} &= \beta_0 + \beta_1 \times \texttt{temp} = 29.98 + 0.001 \times \texttt{temp} \\ \textit{Sample} : \texttt{volume} &= b_0 + b_1 \times \texttt{temp} = 32.35 - 0.1207 \times \texttt{temp} \end{aligned}$$

The discrepancy between the two equations is simply the result of chance: had we selected another sample of 200 bottles, we probably would have found a different sample equation with a different slope and a different intercept. The intercept and slope based on sample data are the result of chance and therefore different from sample to sample. The population intercept and slope (the true ones) are fixed, but unknown. If we want to know something about the population intercept and slope, we only have the sample equation to go on. Our best guess for the population equation is the sample equation; the unbiased estimator for a regression coefficient in the population is the sample coefficient. But how certain can we be about how close the sample intercept and slope are to the population intercept and slope?

5.2 Random sampling and the standard error

In order to know how close the intercept and slope in a sample are to their values in the population, we do another thought experiment. Let's see what happens if we take more than one random sample of 200 bottles.

We put the 200 bottles that we selected earlier back into the population and we again blindly pick a new collection of 200 bottles. We then measure for each bottle the volume of beer it contains and we determine the temperature in the factory on the day of its production. We then apply a regression analysis and

determine the intercept and the slope. Next, we put these bottles back into the population, draw a second random sample of 200 bottles and calculate the intercept and slope again.

You can probably imagine that if we repeat this procedure of randomly picking 200 bottles from a large population of 80,000, each time we find a different intercept and a different slope. Let's carry out this procedure 100 times by a computer. Table 5.1 shows the first 10 regression equations, each based on a random sample of 200 bottles. If we then plot the histograms of all 100 sample intercepts and sample slopes we get Figure 5.3. Remember from Chapters 2 and 3 that these are called *sampling distributions*. Here we look at the sampling distributions of the intercept and the slope.

The sampling distributions in Figure 5.3 show a large variation in the intercepts, and a smaller variation in the slopes (i.e., all values very close to another).

Table 5.1: Ten different sample equations based on ten different random samples from the population of bottles.

sample	equation
1	volume = 28.87 + 0.06 x temperature + e
2	volume = 30.84 – 0.05 x temperature + e
3	volume = 31.05 – 0.06 x temperature + e
4	volume = 31.67 – 0.09 x temperature + e
5	volume = 30.59 – 0.03 x temperature + e
6	volume = 29.53 + 0.02 x temperature + e
7	volume = 28.36 + 0.08 x temperature + e
8	volume = 27.78 + 0.11 x temperature + e
9	volume = 28.29 + 0.09 x temperature + e
10	volume = 30.75 – 0.03 x temperature + e

For now, let's focus on the slope. We do that because we are mostly interested in the linear relationship between volume and temperature. However, everything that follows also applies to the intercept. In Figure 5.4 we see the histogram of the slopes if we carry out the random sampling 1000 times. We see that on average, the sample slope is around 0.001, which is the population slope (the slope if we analyse all bottles). But there is variation around that mean of 0.001: the standard deviation of all 1000 sample slopes turns out to be 0.08.

Remember from Chapter 2 that the standard deviation of the sample distribution is called the *standard error*. The standard error for the sampling distribution of the sample slope represents the uncertainty about the population slope. If the standard error is large, it means that if we would draw many different random samples from the same population data, we would get very different sample slopes. If the standard error is small, it means that if we would draw many different random samples from the same population data, we would get sample slopes that are very close to one another, and very close to the

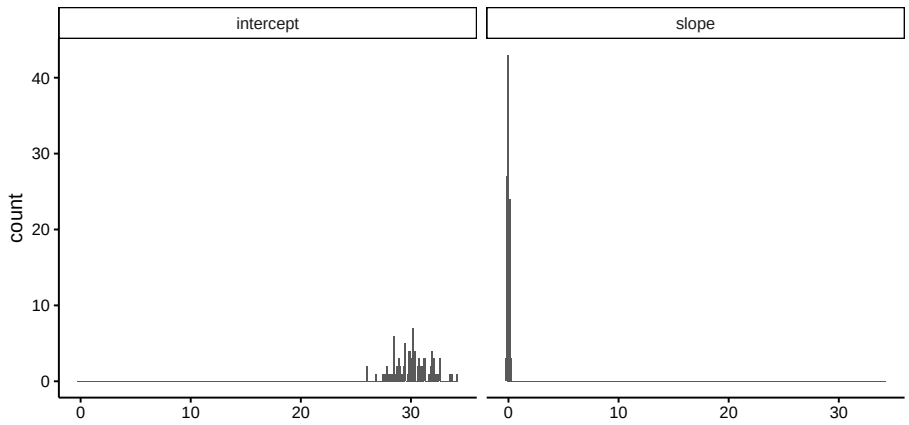


Figure 5.3: Distribution of the 100 sample intercepts and 100 sample slope.

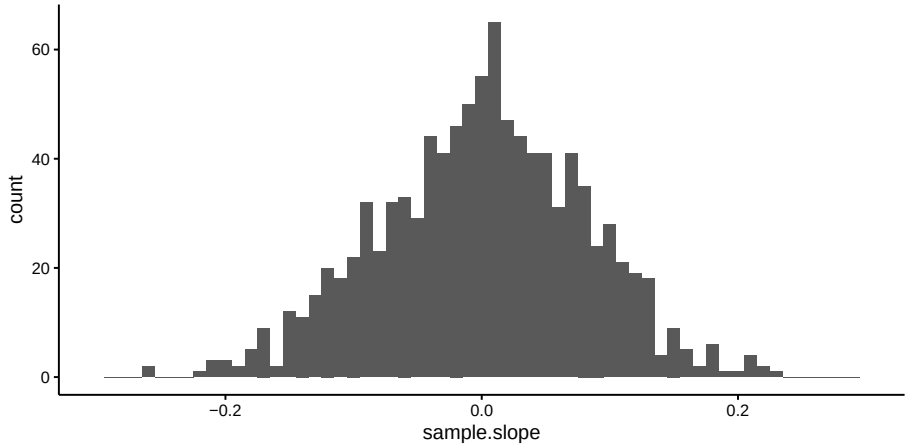


Figure 5.4: Distribution of 1000 sample slopes.

population slope.[2]

5.2.1 Standard error and sample size

Similar to the sample mean, the standard error for a sample slope depends on the *sample size*: how many bottles there are in each random sample. The larger the sample size, the smaller the standard error, the more certain we are about the population slope. In the above example, the sample size is 200 bottles.

The left panel of Figure 5.6 shows the distribution of the sample slope where the sample size is 2. You see that for quite a number of samples, the slope is larger than 10, even if the population slope is 0.001. But when you increase the number of bottles per sample to 20 (in the right panel), you are less dependent on chance observations. With large sample sizes, your results from a regression analysis become less dependent on chance, become more stable, and therefore more reliable.

In Figure 5.6 we see the sampling distributions of the sample slope where the sample size is either 2 (left panel) or 20 (right panel). We see quite a lot of variation in sample slopes with sample size equal to 2, and considerably less variation in sample slopes if sample size is 20. This shows that the larger the sample size, the smaller the standard error, the larger the certainty about the population slope. The dependence of a sample slope on chance and sample size is also illustrated in Figure 5.5.

5.2.2 From sample slope to population slope

In the previous section we saw that if we have a small standard error, we can be relatively certain that our sample slope is close to the population slope. We did a thought experiment where we knew everything about the population intercept and slope, and we drew many samples from this population. In reality, we don't know anything about the population: we only have one sample of data. So suppose we draw a sample of 200 bottles from an unknown population of bottles, and we find a slope of 1, we have to look at the standard error to know how close that sample slope is to the population slope.

For example, suppose we find a sample slope of 1 and the standard error is equal to 0.1. Then we know that the population slope is more likely to be in the neighbourhood of values like 0.9, 1.0, or 1.1 than in the neighbourhood of 10 or -10 (we know that when using the empirical rule, see Chap. 1).

Now suppose we find a sample slope of 1 and the standard error is equal to 10. Then we know that the sample slope is more likely to be somewhere in the neighbourhood of values like -9, 1 or 11, than around values in the neighbourhood of -100 or +100. However, values like -9, 1 and 11 are quite far apart, so actually we have no idea what the population slope is; we don't even know whether the population slope is positive or negative! The standard error is simply too large.

[2]Because sample slopes cluster around the population slope, the sample slope is very close to the population slope when the standard error is small.

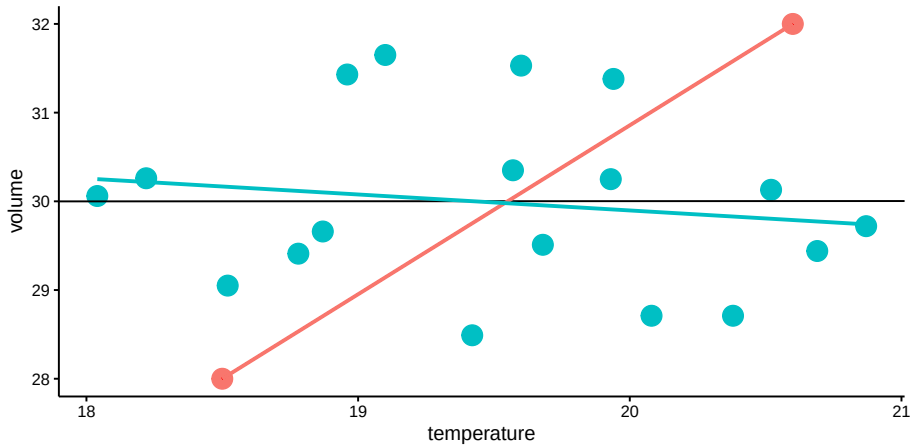


Figure 5.5: The averaging effect of increasing sample size. The scatter plot shows the relationship between temperature and volume for a random sample of 20 bottles (the dots); the first two bottles in the sample are marked in red. The red line would be the sample slope based on these first two bottles, the blue line is the sample slope based on all 20 bottles, and the black line represents the population slope, based on all 80,000 bottles. This illustrates that the larger the sample size, the closer the sample regression line is expected to be to the population regression line.

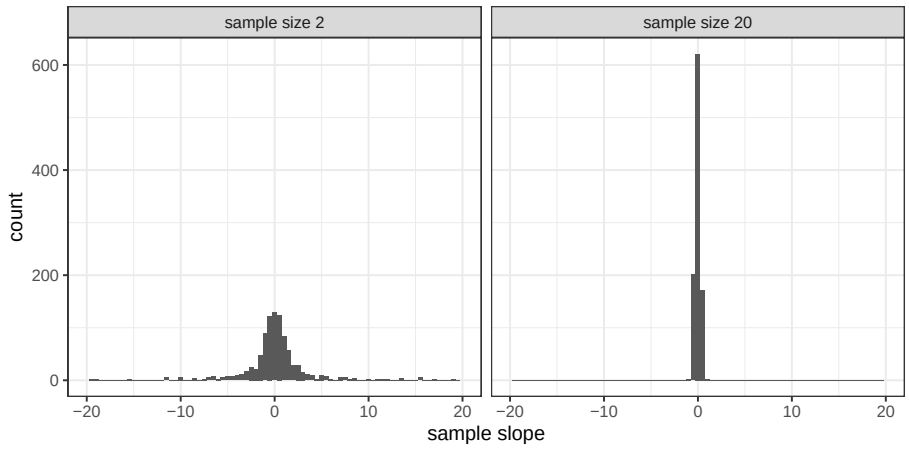


Figure 5.6: Distribution of the sample slope when sample size is 2 (left panel) and when sample size is 20 (right panel).

As we have seen, the standard error depends very much on sample size. Apart from sample size, the standard error for a slope also depends on the variance of the independent variable, the variance of the dependent variable, and the correlations between the independent variable and other independent variables in the equation. We will not bore you with the complicated formula for the standard error for regression coefficients in the case of multiple regression [3]. But here is the formula for the standard error for the slope coefficient if you have only one predictor variable X:

$$\begin{aligned} \sigma_{\widehat{b_1}} &= \frac{\sqrt{\frac{SSR}{n-2}}}{\sqrt{SS_X}} \\ &= \frac{\sqrt{\frac{\Sigma_i(Y_i-\widehat{Y_i})^2}{n-2}}}{\sqrt{\Sigma_i(X_i-\overline{X})^2}} = \sqrt{\frac{\Sigma_i(Y_i-\widehat{Y_i})^2}{(n-2)\sqrt{\Sigma_i(X_i-\overline{X})^2}}} \end{aligned} \tag{5.1}$$

where b_1 is the slope coefficient in the sample, n is sample size, SSR is the sum of the squared residuals, and SS_X the sum of squares for independent variable X. From the formula, you can see that the standard error $\sigma_{\widehat{b_1}}$ becomes smaller when sample size n becomes larger.

It's not very useful to memorise this formula; you'd better let R do the calculations for you. But an interesting part of the formula is the nominator: $\frac{SSR}{n-2}$. This is the sum of the squared residuals, divided by $n-2$. Remember from Chapter 1 that the definition of the variance is the sum of squares divided by the number of values. Thus it looks like we are looking at the variance of the residuals. Remember from Chapter 2 that when we want to estimate a population variance, a biased estimator is the variance in the sample. In order to get an unbiased estimate of the variance, we have to divide by $n-1$ instead of n. This was because when computing the sum of squares, we assume we know the mean. Here we are computing the variance of the residuals, but it's actually an unbiased estimator of the variance in the population, because we divide by $n-2$: when we compute the residuals, we assume we know the intercept and the slope. We assume two parameters, so we divide by $n-2$. Thus, when we have a linear model with 2 parameters (intercept and slope), we have to divide the sum of squared residuals by $n-2$ in order to obtain an unbiased estimator of the variance of the residuals in the population.

From the equation, we see that the standard error becomes larger when there is a large variation in the residuals, it becomes smaller when there is a large variation in predictor variable X, and it becomes smaller with large sample size n.

[3] See https://www3.nd.edu/ rwilliam/stats1/x91.pdf for the formula. In this pdf, 'IV' means independent variable

5.3 t-distribution for the model coefficients

When we look at the sample distribution of the sample slope, for instance in Figure 5.4, we notice that the distribution looks very much like a normal distribution. From the Central Limit Theorem, we know that the sampling distribution will become very close to normal for large sample sizes. Using this sampling distribution for the slope we could compute confidence intervals and do null-hypothesis testing, similar to what we did in Chapters 2 and 3.

For large sample sizes, we could assume the normal distribution, and when we standardise the slope coefficient, we can look up in tables such as in Appendix A the critical value for a particular confidence interval. For instance, 200 bottles is a large sample size. When we standardise the sample slope – let's assume we find a slope of 0.05 –, we need to use the values -1.96 and +1.96 to obtain a 95% confidence interval around 0.05. The margin of error (MoE) is then 1.96 times the standard error. Suppose that the standard error is 0.10. The MoE is then equal to $1.96 \times 0.10 = 0.196$. The 95% interval then runs from $0.05 - 0.196 = -0.146$ to $0.05 + 0.196 = 0.246$.

However, this approach does not work for small sample sizes. Again this can be seen when we standardise the sampling distribution. When we standardise the slope for each sample, we subtract the sample slope from the population slope β_1, and have to divide each time by the standard error (the standard deviation). But when we do that

$$t = \frac{b_1 - \beta_1}{\widehat{\sigma_{b_1}}} = \frac{b_1 - \beta_1}{\frac{\sqrt{\frac{SSR}{n-2}}}{\sqrt{SS_X}}} \tag{5.2}$$

we immediately see the problem that when we only have sample data, we have to estimate the standard error. In each sample, we get a slightly different estimated standard error, because each time, the variation in the residuals (SSR) is a little bit different, and also the variation in the predictor variable (SS_X). If sample size is large, this is not so bad: we then can get very good estimates of the standard error so there is little variation across samples. But when sample size is small, both SSR and SS_X are different from sample to sample (due to chance), and the estimate of the standard error will therefore also vary a lot. The result is that the distribution of the standardised t-value from Equation 5.2 will only be close to normal for large sample size, but will have a t-distribution in general.

Because the standard error is based on the variance of the residuals, and because the variance of the residuals can only be computed if you assume a certain intercept and a certain slope, the degrees of freedom will be $n - 2$.

Let's go back to the example of the beer bottles. In our first random sample of 200 bottles, we found a sample slope of -0.121. We also happened to know the population slope, which was 0.001. From our computer experiment, we saw that the standard deviation of the sample slopes with sample size 200 was equal to 0.08. Thus, if we fill in the formula for the standardised slope t, we get for

this particular sample

$$t = \frac{-0.1207 - 0.001}{0.08} = -1.52 \tag{5.3}$$

In this section, when discussing t-statistics, we assumed we knew the population slope β, that is, the slope of the linear equation based on all 80,000 bottles. In reality, we never know the population slope: the whole reason to look at the sample slope is to have an idea about the population slope. Let's look at the confidence interval for slopes.

5.4 Confidence intervals for the slope

Since we don't know the actual value of the population slope β_1, we could ask the personnel in the beer factory what they think is a likely value for the slope. Suppose Mark says he believes that a slope of 0.1 could be true. Well, let's find out whether that is a reasonable guess, given that the sample slope is -0.121. Now we *assume* that the population slope β_1 is 0.1, and we compute the t-statistic for our sample slope -0.121:

$$t = \frac{-0.121 - 0.1}{0.08} = -2.7 \tag{5.4}$$

Thus, we compute how many standard errors the sample value is away from the hypothesised population value 0.1. If the population value is indeed 0.1, how likely is it that we find a sample slope of -0.121?

From the t-distribution, we know that such a t-value is very unlikely: the probability of finding a sample slope -2.7 standard deviations or more away from a population slope of 0.1 is less than 0.0075341. How do we know that? Well, the t-statistic is -2.7 and the degrees of freedom is $200 - 2 = 198$. The cumulative proportion of a t-value can be looked up in R:

```
pt(-2.7, df = 198)
```

```
## [1] 0.003767051
```

That means that a proportion of 0.0037671 of all values in the t-distribution with 198 degrees of freedom are lower than -2.7. Because the t-distribution is symmetric, we then also know that 0.0037671 of all values are larger than 2.7. If we add up these two numbers, we know that 0.0075341 of all values in a t-distribution are less than -2.7 or more than 2.7. That means that if the population slope is 0.1, we only find a sample slope of ± -0.121 or more extreme with a probability of 0.0075341. That's very unlikely.

Because we know that such a t-value of ± -2.7 or more extreme is unlikely, we know that a sample slope of -0.1206874 is unlikely *if the population slope is equal to 0.1*. Therefore, we feel 0.1 is not a realistic value for the population slope.

Now let's ask Martha. She thinks a reasonable value for the population slope is 0, as she doesn't believe there is a linear relationship between temperature and volume. She suspects that the fact that we found a sample slope that was not 0 was a pure coincidence. Based on that hypothesis, we compute t again and find:

$$t = \frac{-0.121 - 0}{0.08} = -1.5 \tag{5.5}$$

In other words, if we believe Martha, our sample slope is only about 1 standard deviation away from her hypothesised value. That's not a very bad idea, since from the t-distribution we know that the probability of finding a value more than 1.5 standard deviations away from the mean (above or below) is 13.35%. You can see that by asking R:

```
pt(-1.5, df = 198) * 2
```

```
## [1] 0.1352072
```

Thirteen percent, that's about 1 in 7 or 8 times. That's not so improbable. In other words, if the population slope is truly 0, then our sample slope of -0.121 is quite a reasonable finding. If we reverse this line of reasoning: if our sample slope is -0.121, with a standard error of 0.08, then a population slope of 0 is quite a reasonable guess! It is reasonable, since the difference between the sample slope and the hypothesised value is only 1.5 standard errors.

So when do we no longer feel that a person's guess of the population slope is reasonable? Perhaps if the probability of finding a sample slope of at least a certain size given a hypothesised population slope is so small that we no longer believe that the hypothesised value is reasonable. We might for example choose a small probability like 1%. We know from the t-distribution with 198 degrees of freedom that 1% of the values lie at least 2.6 standard deviations above and below the mean.

```
qt(0.005, df = 198)
```

```
## [1] -2.600887
```

This is shown in Figure 5.7. So if our sample slope is more than 2.6 standard errors away from the hypothesised population slope, then that population slope is *not* a reasonable guess. In other words, if the *distance* between the sample slope and the hypothesised population slope is more than 2.6 standard errors, then the hypothesised population slope is no longer reasonable.

This implies that *any* value closer than 2.6 standard errors from the sample slope is a collection of reasonable values for the population slope.

Thus, in our example of the 200 bottles with a sample slope of -0.121 and a standard error of 0.08, the interval from $-0.121 - 2.6 \times 0.08$ to $-0.121 + 2.6 \times 0.08$ contains reasonable values for the population slope. If we do the calculations,

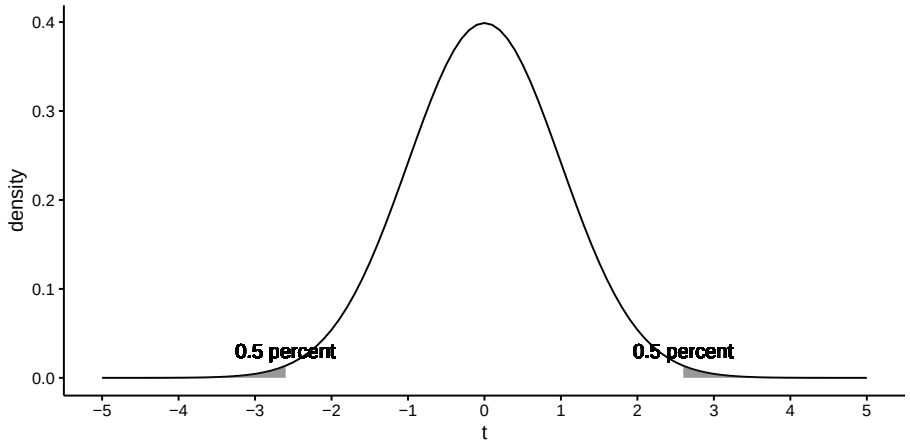


Figure 5.7: The t-distribution with 198 degrees of freedom.

we get the interval from -0.33 to 0.09. If we would have to guess the value for the population slope, our guess would be that it would lie somewhere between between -0.33 and 0.09, *if we feel that 1% is a small enough probability.*

In data analysis, such an interval that contains reasonable values for the population value, if we only know the sample value, is called a *confidence interval*, as we know from Chapter 2. Here we've chosen to use 2.6 standard errors as our cut-off point, because we felt that 1% would be a small enough probability to dismiss the real population value as a reasonable candidate (type I error rate). Such a confidence interval based on this 1% cut-off point is called a 99% confidence interval.

Particularly in social and behavioural sciences, one also sees 95% confidence intervals. The critical t-value for a type I error rate of 0.05 and 198 degrees of freedom is 1.97.

```
qt(0.975, df = 198)
```

```
## [1] 1.972017
```

Thus, 5% of the observations lie more than 1.97 standard deviations away from the mean, so that the 95% confidence interval is constructed by subtracting/adding 1.97 standard errors from/to the sample slope. Thus, in the case of our bottle sample, the 95% confidence interval for the population slope is from $-0.121 - 1.97 \times 0.08$ to $-0.121 + 1.97 \times 0.08$, so reasonable values for the population slope are those values between -0.28 and 0.04. Luckily, this corresponds to the truth, because we happen to know that the population slope is equal to 0.001. In real life, we don't know the population slope and of course it might happen that the true population value is not in the 95% confidence interval. If you want to make the likelihood of this being the case smaller, then you can use

a 99%, a 99.9% or an even larger confidence interval.

5.5 Residual degrees of freedom in linear models

What does the term, "degrees of freedom" mean? In Chapter 2 we discussed degrees of freedom in the context of doing inference about a population mean. We saw that degrees of freedom referred to the number of values in the final calculation of a statistic that are free to vary. More specifically, 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. There, we computed a t-statistic for the sample mean. Because in the computation of the t-statistic for a sample mean, we divide by the standard error for the mean, and that this in turn requires assuming a certain value for the mean, we had $n - 1$ degrees of freedom.

Here, we are talking about t-statistics for linear models. In the case of a simple regression model, we only have one intercept and one slope. In order to compute a t-value, for the slope for example, we have to estimate the standard error as well. In Equation 5.1 we see that in order to estimate the standard error, we need to compute the residuals $e_i = Y_i - \hat{Y}_i$. But you can only compute residuals if you have predictions for the dependent variable, $\hat{Y}_i$, and for that you need an intercept and a slope coefficient. Thus, we need to assume we know two parameters, in order to calculate a t-value. With sample means we only assumed we knew the mean, and therefore had $n - 1$ degrees of freedom. In case of a linear model where we assume one intercept and one slope, we have $n - 2$ degrees of freedom. For the same reason, if we have a linear model with one intercept and two slopes (multiple regression with two predictors), we have $n - 3$ degrees of freedom. In general then, if we have a linear model with K independent variables, we have $n - K - 1$ degrees of freedom associated with our t-statistic.

To convince you of this, we illustrate the idea of degrees of freedom in a numerical example. Suppose that we have a sample with four Y values: 2, 6, 5, 2. There are four separate pieces of information here. There is no particular connection between these values. They are free to take any values, in principle. We could say that there are four degrees of freedom associated with this sample of data.

Now, suppose that I tell you that three of the values in the sample are 2, 6, and 2; and I also tell you that the sample mean is 3.75. You can immediately deduce that the remaining value has to be 5. Were it any other value, the mean would not be 3.75.

$$\overline{Y} = \frac{\Sigma Y_i}{n} = \frac{2 + 6 + Y_3 + 2}{4} = \frac{10 + Y_3}{4} = 3.75$$
$$10 + Y_3 = 4 \times 3.75 = 15$$
$$Y_3 = 15 - 10 = 5$$

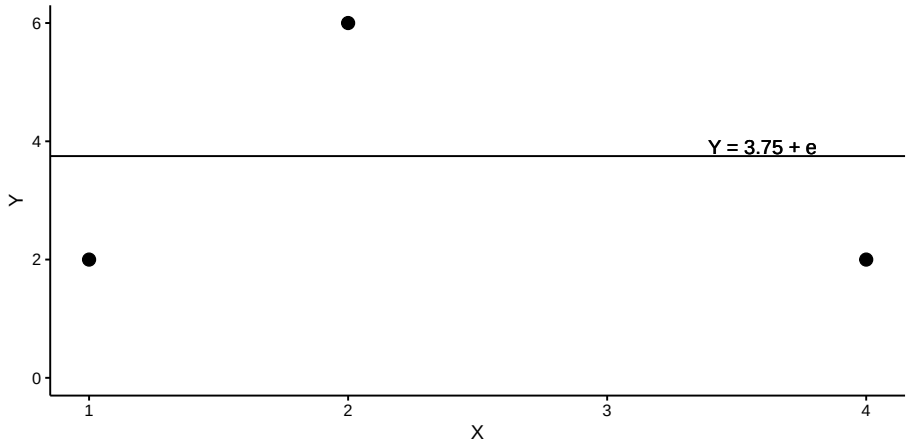


Figure 5.8: Illustration of residual degrees of freedom in a linear model, in case there is no slope and the intercept equals 3.75.

Once I tell you that the sample mean is 3.75, I am effectively introducing a *constraint*. The value of the unknown sample value is implicitly being determined from the other three values plus the constraint. That is, once the constraint is introduced, there are only three logically independent pieces of information in the sample. That is to say, there are only three "degrees of freedom", once the sample mean is revealed.

Let's carry this example to regression analysis. Suppose I have four observations of variables X and Y, where the values for X are 1, 2, 3 and 4. Each value of $Y = y$ is one piece of information. These Y-values could be anything, so we say that we have 4 degrees of freedom. Now suppose I use a linear model for these data points, and suppose I only use an intercept. Let the intercept be 3.75 so that we have $Y = 3.75 + e$. Now the first bit of information for $X = 1$, Y could be anything, say 2. The second and third bits of information for $X = 2$ and $X = 4$ could also be anything, say 6 and 2. Figure 5.8 shows these bits of information as dots in a scatter plot. Since we know that the intercept is equal to 3.75, with no slope (slope=0), we can also draw the regression line.

Before we continue, you must know that if we talk about degrees of freedom in regression analysis, we generally talk about *residual degrees of freedom*. We therefore look at residuals. If we compute the residuals, we have residuals -1.75, 2.25 and -1.75 for these data points. When we sum them, we get -1.25. Since we know that all residuals should sum to 0 in a regression analysis (see Chapter 4), we can derive the fourth residual to be +1.25, since only then the residuals sum to 0. Therefore, the Y-value for the fourth data point (for $X = 3$) has to be 5, since then the residual is equal to $5 - 3.75 = 1.25$.

In short, when we use a linear model with only an intercept, the degrees of freedom is equal to the number of data points (combinations of X and Y) minus

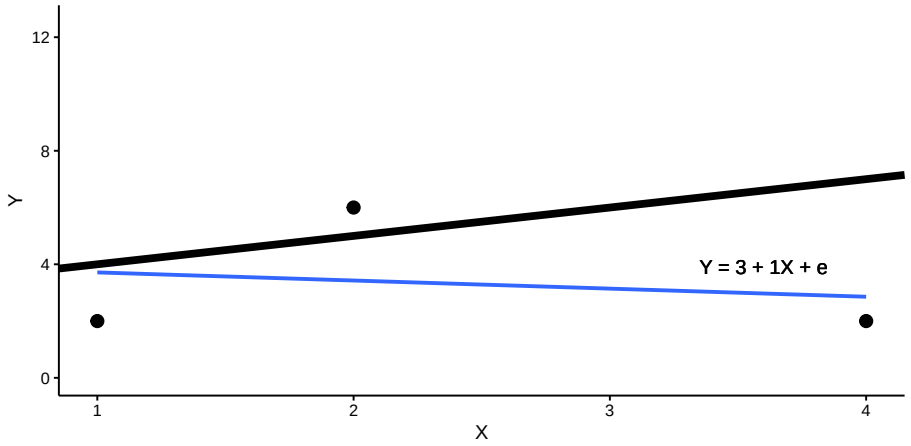


Figure 5.9: Illustration of residual degrees of freedom, in case of a linear model with both intercept and slope for four data points (black line). The blue line is the regression line only using the three known data points.

1, or in short notation: $n - 1$, where n stands for sample size.

Now let's look at the situation where we use a linear model with both an intercept and a slope: suppose the intercept is equal to 3 and the slope is equal to 1: $Y = 3 + 1X + e$. Then suppose we have the same X-values as the example above: 1, 2 and 4. When we give these X-values corresponding Y-values, 2, 6, and 2, we get the plot in Figure 5.9.

The black line is the regression line that we get when we analyse the complete data set of four points, $Y = 3 + 1X$. The blue line is the regression line based on only the three visible data points. Now the question is, is it possible for a fourth data point with $X = 3$, to think of a Y-value such that the regression line based on these four data points is equal to $Y = 3 + 1X$? In other words, can we choose a Y-value such that the blue line exactly overlaps with the black line?

Figure 5.10 shows a number of possibilities for the value of Y if $X = 3$. It can be seen, that it is impossible to pick a value for Y_3 such that we get a regression equation $Y = 3+1X$. The blue line and green line intersect the black line at $X = 1$, but they have slopes that are less steep than the black line. If you use lower values for Y such as 9 (red line) or higher values like 15 (purple line), the regression lines still do not overlap. It turns out to be impossible to choose a value for Y_3 in such a way that the regression line matches $Y = 3+1X$.

So, with sample size $n = 4$, we can never freely choose 3 residuals in order to satisfy the constraint that a particular regression equation holds for all 4 data points. We have less than 3 degrees of freedom because it is impossible to think of a fitting fourth value. It turns out, that in this case we can only choose 2 residuals freely, and the remaining residuals are then already determined. To

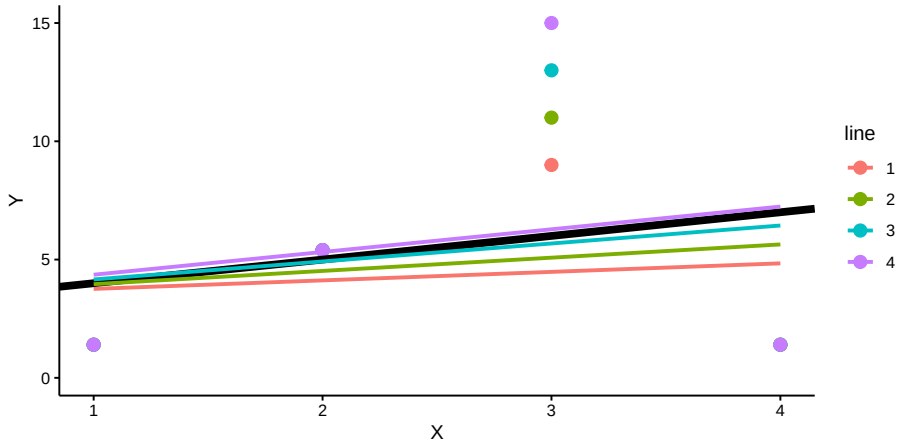


Figure 5.10: Different regression lines for different values of Y if $X = 3$.

prove this requires matrix algebra, but you can see it when you try it yourself.

The gist of it is that if you have a regression equation with both an intercept and a slope, the degrees of freedom is equal to the number of data points (sample size) minus 2: $n - 2$. Generalising this to linear models with K predictors: $n - K - 1$.

Generally, these degrees of freedom based on the number of residuals that could be freely chosen, given the constraints of the model, are termed *residual degrees of freedom*. When using regression models, one usually only reports these residual degrees of freedom. Later on in this book, we will see instances where one also should use *model degrees of freedom*. For now, it suffices to know what is meant by residual degrees of freedom.

5.6 Null-hypothesis testing with linear models

Often, data analysis is about finding an answer to the question whether there is a relationship between two variables. In most cases, the question pertains to the population: is there a relationship between variable Y and variable X in the population? In many cases, one looks for a linear relationship between two variables.

One common method to answer this question is to analyse a sample of data, apply a linear model, and look at the slope. However, one then knows the slope in the sample, but not the slope in the population. We have seen that the slope in the sample can be very different from the slope in the population. Suppose we find a slope of 1: does that mean there is a slope in the population or that there is no slope in the population?

In inferential data analysis, one often works with two hypotheses: the *null-hypothesis* and the *alternative hypothesis*. The null-hypothesis states that the

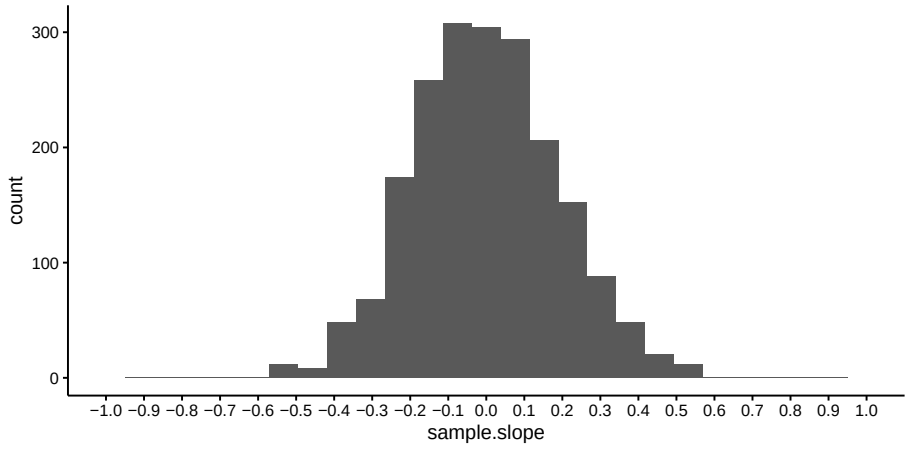


Figure 5.11: Distribution of the sample slope when the population slope is 0 and sample size equals 40.

population slope is equal to 0 and the alternative hypothesis states that there is a slope that is different from 0. Remember that if the population slope is equal to 0, that is saying that there is no linear relationship between X and Y (that is, you cannot predict one variable on the basis of the other variable). Therefore, the null-hypothesis states there is no linear relationship between X and Y in the population. If there is a slope, whether positive or negative, is the same as saying there is a linear relationship, so the alternative hypothesis states that there is a linear relationship between X and Y in the population.

In formula form, we have

$$H_0 : \beta_{slope} = 0 \tag{5.6}$$
$$H_A : \beta_{slope} \neq 0 \tag{5.7}$$

The population slope, β_{slope}, is either 0 or it is not. Our data analysis is then aimed at determining which of these two hypotheses is true. Key is that we do a thought experiment on the null-hypothesis: we wonder what would happen if the population slope would be really 0. In our imagination we draw many samples of a certain size, say 40 data points, and then determine the slope for each sample. Earlier we learned that the many sample slopes would form a histogram in the shape of a t-distribution with $n - 2 = 38$ degrees of freedom. For example, suppose we would draw 1000 samples of size 40, then the histogram of the 1000 slopes would look like depicted Figure 5.11.

From this histogram we see that all observed sample slopes are well between -0.8 and 0.8. This gives us the information we need. Of course, we have only one sample of data, and we don't know anything about the population data. But we *do* know that *if the population slope is equal to 0*, then it is very unlikely

to find a sample slope of say 1 or -1. Thus, with our sample slope of 1, we know that this finding is very unlikely *if we hold the null-hypothesis to be true*. In other words, if the population slope is equal to 0, it would be quite improbable to find a sample slope of 1 or larger. Therefore, we regard the null-hypothesis to be false, since it does not provide a good explanation of why we found a sample slope of 1. In that case, we say that *we reject the null-hypothesis*.

5.7 p-values

A p-value is a probability. It represents the probability of observing certain events, given that the null-hypothesis is true.

In the previous section we saw that if the population slope is 0, and we drew 1000 samples of size 40, we did not observe a sample slope of 1 or larger. In other words, the frequency of observing a slope of 1 or larger was 0. If we would draw more samples, we theoretically could observe a sample slope of 1, but the probability that that happens for any new sample we can estimate at less than 1 in a 1000, so less than 0.001: $p < 0.001$.

This estimate of the p-value was based on 1000 randomly drawn samples of size 40 and then looking at the frequency of certain values in that data set. But there is a short-cut, for we know that the distribution of sample slopes has a t-distribution if we standardise the sample slopes. Therefore we do not have to take 1000 samples and estimate probabilities, but we can look at the t-distribution directly, using tables online or in statistical packages.

Figure 5.12 shows the t-distribution that is the theoretical distribution corresponding to the histogram in Figure 5.11. If the standard error is equal to 0.19, and the hypothetical population slope is 0, then the t-statistic associated with a slope of 1 is equal to $t = \frac{1-0}{0.19} = 5.26$. With this value, we can look up in the tables, how often such a value of 5.26 *or larger* occurs in a t-distribution with 38 degrees of freedom. In the tables or using R, we find that the probability that this occurs is 0.00000294.

```
1 - pt(5.26, df = 38)
```

```
## [1] 0.000002939069
```

So, the fact that the t-statistic has a t-distribution gives us the opportunity to exactly determine certain probabilities, including the p-value.

Now let's suppose we have only one sample of 40 bottles, and we find a slope of 0.1 with a standard error of 0.19. Then this value of 0.1 is $(0.1-0)/0.19 = 0.53$ standard errors away from 0. Thus, the t-statistic is 0.53. We then look at the t-distribution with 38 degrees of freedom, and see that such a t-value of 0.53 is not very strange: it lies well within the middle 95% of the t-distribution (see Figure 5.12).

Let's determine the p-value again for this slope of 0.1: we determine the probability that we obtain such a t-value of 0.53 or larger. Figure 5.13 shows

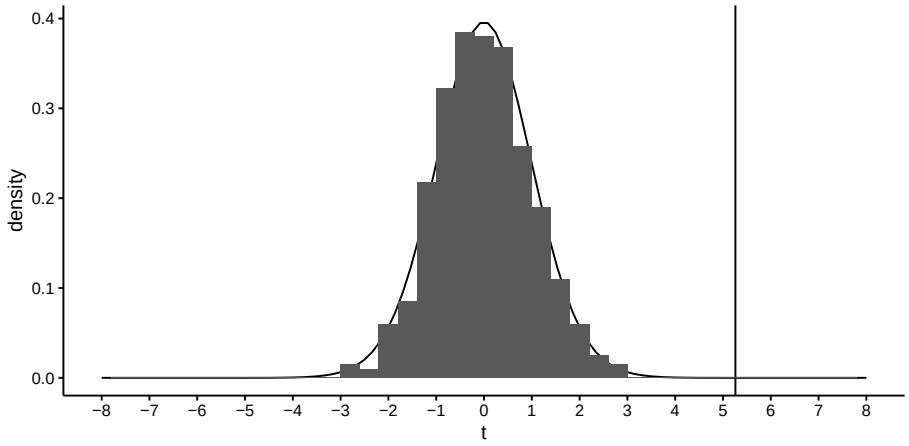


Figure 5.12: The histogram of 1000 sample slopes and its corresponding theoretical t-distribution with 38 degrees of freedom. The vertical line represents the t-value of 5.26.

the area under the curve for values of t that are larger than 0.53. This area under the curve can be seen as a probability. The total area under the curve of the t-distribution amounts to 1. If we know the area of the shaded part of the total area, we can compute the probability of finding t-values larger than 0.53.

In tables online, in Appendix B, or available in statistical packages, we can look up how large this area is. It turns out to be 0.3.

```
1 - pt(0.53, df = 38)
```

```
## [1] 0.2995977
```

So, if the population slope is equal to 0 and we draw an infinite number of samples of size 40 and compute the sample slopes, then 30% of them will be larger than our sample slope of 0.1. The proportion of the shaded area is what we call a *one-sided* p-value. We call it one-sided, because we only look at one side of the t-distribution: we only look at values that are larger than our t-value of 0.53.

We conclude that a slope value of 0.1 is not that strange to find if the population slope is 0. By the same token, it would also have been probable to find a slope of -0.1, corresponding to a t-value of -0.53. Since the t-distribution is symmetrical, the probability of finding a t-value of less than -0.53 is depicted in Figure 5.14, and of course this probability is also 0.3.

```
pt(-0.53, df = 38)
```

```
## [1] 0.2995977
```

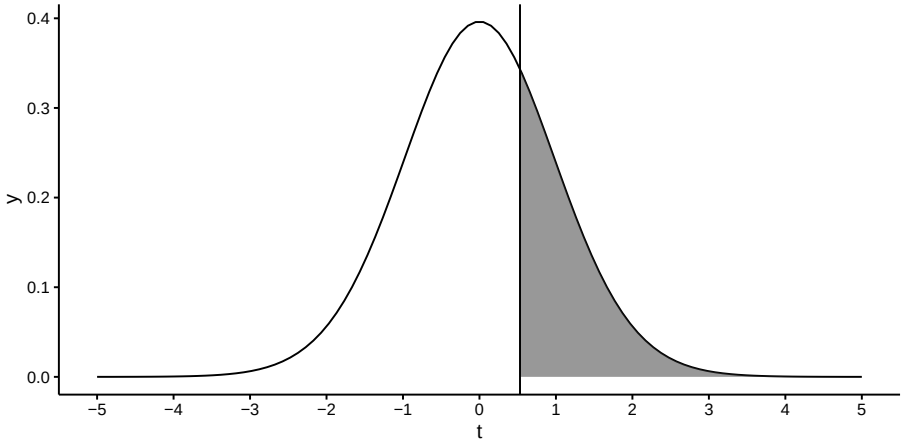


Figure 5.13: Probability of a t-value larger than 0.53.

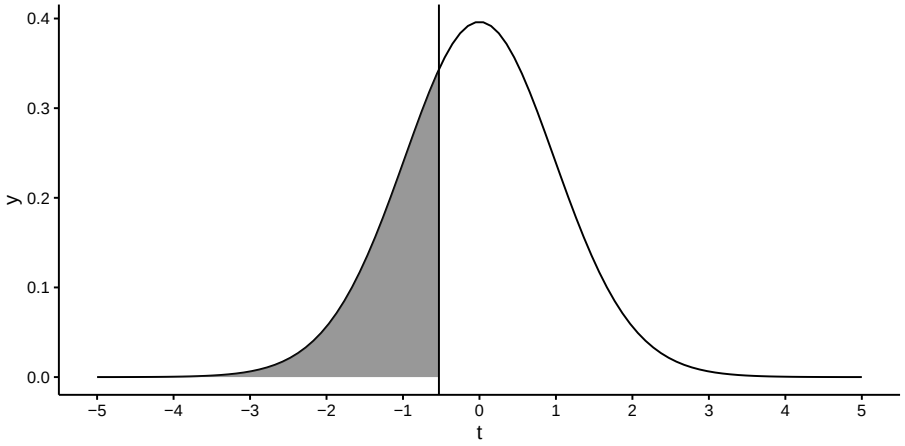


Figure 5.14: Probability of finding a t-value smaller than -0.53.

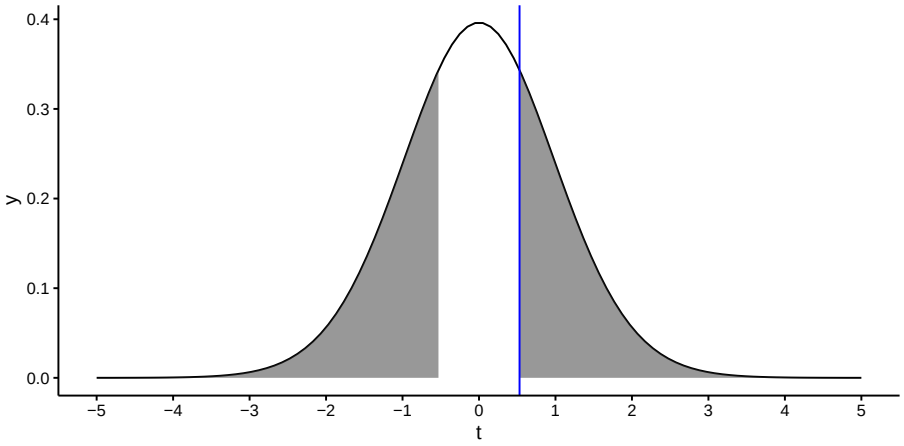


Figure 5.15: The blue vertical line represents a t-value of 0.53. The shaded area represents the two-sided p-value: the probability of obtaining a t-value smaller than -0.53 or larger than 0.53.

Remember that the null-hypothesis is that the population slope is 0, and the alternative hypothesis is that the population slope is *not* 0. We should therefore conclude that if we find a very large positive *or* negative slope, large in the sense of the number of standard errors away from 0, that the null-hypothesis is unlikely to be true. Therefore, if we find a slope of 0.1 or -0.1, then we should determine the probability of finding a t-value that is larger than 0.53 or smaller than -0.53. This probability is depicted in Figure 5.15 and is equal to twice the one-side p-value, $2 \times 0.2995977 = 0.5991953$.

This probability is called the *two-sided* p-value. This is the one that should be used, since the alternative hypothesis is also two-sided: the population slope can be positive or negative. The question now is: is a sample slope of 0.1 enough evidence to reject the null-hypothesis? To determine that, we determine how many standard errors away from 0 the sample slope is and we look up in tables how often that happens. Thus in our case, we found a slope that is 0.53 standard errors away from 0 and the tables told us that the probability of finding a slope that is at least 0.53 standard errors away from 0 (positive or negative) is equal to 0.5991953. We find this probability rather large, so we decide that we *do not reject the null-hypothesis*.

5.8 Hypothesis testing

In the previous section, we found a one-sided p-value of 0.00000294 for a sample slope of 1 and more or less concluded that this probability was rather small. The two-sided p-value would be twice this value, so 0.00000588, which is still very small. Next, we determined the p-value associated with a slope of 0.1 and

found a p-value of 0.60. This probability was rather large, and we decided to *not* reject the null-hypothesis. In other words, the probability was so large that we thought that the hypothesis that the population slope is 0 should not be rejected based on our findings.

When should we think the p-value is small enough to conclude that the null-hypothesis can be rejected? When can we conclude that the hypothesis that the population slope is 0 is not supported by our sample data? This was a question posed to the founding father of statistical hypothesis testing, Sir Ronald Fischer. In his book *Statistical Methods for Research Workers* (1925), Fisher proposed a probability of 5%. He advocated 5% as a standard level for concluding that there is evidence against the null-hypothesis. However, he did not see it as an absolute rule: "If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05...". So Fisher saw the p-value as an informal index to be used as a measure of discrepancy between the data and the null-hypothesis: The null-hypothesis is never proved, but is possibly disproved.

Later, Jerzy Neyman and Egon Pearson saw the p-value as an instrument in decision making: is the null-hypothesis true, or is the alternative hypothesis true? You either reject the null-hypothesis or you don't, there is nothing in between. A slightly milder view is that you either decide that there is enough empirical evidence to reject the null-hypothesis, or there is not enough empirical evidence to reject the null-hypothesis (not necessarily accepting H_0 as true!). This view to data-analysis is rather popular in the social and behavioural sciences, but also in particle physics. In order to make such black-and-white decisions, you decide before-hand, that is, before collecting data, what *level of significance* you choose for your p-value to decide whether to reject the null-hypothesis. For example, as your significance level, you might want to choose 1%. Let's call this chosen significance level α. Then you collect your data, you apply your linear model to the data, and find that the p-value associated with the slope equals p. If this p is smaller than or equal to α, you *reject the null-hypothesis*, and if p is larger than α then you *do not reject the null-hypothesis*. A slope with a $p \leq \alpha$ is said to be *significant*, and a slope with a $p > \alpha$ is said to be *non-significant*. If the sample slope is significant, then one should reject the null-hypothesis and say there is a slope in the population different from zero. If the sample slope is not significant, then one should not reject the null-hypothesis and say there is no slope in the population (i.e., the slope is 0). Alternatively, one could say there is no empirical evidence for the existence of a slope (this leaves the possibility that there is a slope in the population but that our method of research failed to find evidence for it).

5.9 Inference for linear models in R

So far, we have focused on standard errors and confidence intervals for the slope parameter in simple regression, that is, a linear model where there is only one independent variable. However, the same logic can be applied to the intercept parameter, and to other slope variables in case you have multiple independent variables in your model (multiple regression).

For instance, suppose we are interested in the knowledge university students have of mathematics. We start measuring their knowledge at time 0, when the students start doing a bachelor programme in mathematics. At time 1 (after 1 year) and at time 2 (after two years), we also perform measures. Our dependent variable is mathematical knowledge, a measure with possible values between 200 and 700. The independent variables are `time` (the time of measurement) and `distance`: the distance in kilometers between university and their home. There are two research questions. First question is about the level of knowledge when students enter the bachelor programme, and the second question is how much knowledge is acquired in one year of study. The linear model is as follows:

$$\begin{aligned} \texttt{knowledge} &= b_0 + b_1\texttt{time} + b_2\texttt{distance} + e \\ e &\sim N(0, \sigma^2) \end{aligned} \tag{5.8}$$

Table 5.2: Regression table as obtained from R, with knowledge predicted by time and distance.

term	estimate	std.error	statistic	p.value
(Intercept)	299.35	7.60	39.40	0.00
time	18.13	2.60	6.96	0.00
distance	-0.76	0.67	-1.15	0.25

The first question could be answered by estimating the intercept b_0: that is the level of knowledge we expect for a student at time 0 and with a home 0 kilometres from the university. The second question could be answered by estimating the slope coefficient for `time`: the expected increase in knowledge per year. In Chapter 4 we saw how to estimate the regression parameters in R. We saw that we then get a *regression table*. For our mathematical knowledge example, we could obtain the regression table, displayed in Table 5.2. We discussed the first column with the regression parameters in Chapter 4. We see that the intercept is estimated at 299.35, and the slopes for time and distance are 18.13 and -0.76, respectively. So we can fill in the equation:

$$\texttt{knowledge} = 299.35 + 18.13\texttt{time} - 0.76\texttt{distance} + e \tag{5.9}$$

Let's look at the other columns in the regression table. In the second column we see the standard errors for each parameter. The third column gives statistics;

these are the t-statistics for the null-hypotheses that the respective parameters in the population are 0. For instance, the first statistic has the value 39.40. It belongs to the intercept. If the null-hypothesis is that the population intercept is 0 ($\beta_0 = 0$), then the t-statistic is computed as

$$t = \frac{b_0 - \beta_0}{\sigma_{\hat{\beta}}} = \frac{299.35 - 0}{7.60} = \frac{299.35}{7.60} = 39.40 \tag{5.10}$$

You see that the t-statistic in the regression table is simply the regression parameter divided by its standard error. This is also true for the slope parameters. For instance, the t-statistic of 6.96 for `time` is simply the regression coefficient 18.13 divided by the standard error 2.60:

$$t = \frac{b_1 - \beta_0}{\sigma_{\hat{\beta}}} = \frac{18.13 - 0}{7.60} = \frac{18.13}{2.60} = 6.96 \tag{5.11}$$

The last column gives the two-sided p-values for the respective null-hypotheses. For instance, the p-value of 0.00 for the intercept says that the probability of finding an intercept of 299.35 or larger (plus or minus), under the assumption that the population intercept is 0, is very small (less than 0.01).

If you want to have confidence intervals for the intercept and the slope for time, you can use the information in the table to construct them yourself. For instance, according to the table, the standard error for the intercept equals 7.60. Suppose the sample size equals 90 students, then you know that you have $n - K - 1 = 90 - 2 - 1 = 87$ degrees of freedom. The critical value for a t-statistic with 84 degrees of freedom for a 95% confidence interval can be looked up in Appendix B. It must be somewhere between 1.98 and 2.00, so let's use 1.99. The 95% interval for the intercept then runs between $299.35 - 1.99 \times 7.60$ and $299.35 - 1.99 \times 7.60$, so the expected level of knowledge at the start of the bachelor programme for students living close to or on campus is somewhere between from 284.23 to 314.47.

To show you how this can all be done using R, we have a look at the R dataset called "freeny" on quarterly revenues. We would like to predict the variable `market.potential` by the predictors `price.index` and `income.level`. Apart from the `tidyverse` package, we also need the `broom` package for the `tidy()` function. When we run the following code, we obtain a regression table.

```
library(broom)
data("freeny")
out <- freeny %>%
  lm(market.potential ~ price.index + income.level, data = .)
out %>%
  tidy()

## # A tibble: 3 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
```

```
## 1 (Intercept)    13.3      0.291      45.6  1.86e-33
## 2 price.index    -0.309    0.0263    -11.8  6.92e-14
## 3 income.level    0.196    0.0291      6.74 7.20e- 8
```

We can have R compute the respective confidence intervals by indicating that we want intervals of a certain confidence level, say 99%:

```
out <- freeny %>%
  lm(market.potential ~ price.index + income.level, data = .)
out %>%
  tidy(conf.int = 0.99)

## # A tibble: 3 x 7
##   term         estimate std.error statistic  p.value conf.low conf.high
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)    13.3      0.291      45.6  1.86e-33   12.7      13.9
## 2 price.index    -0.309    0.0263    -11.8  6.92e-14   -0.363    -0.256
## 3 income.level    0.196    0.0291      6.74 7.20e- 8    0.137     0.255
```

In the last two columns we see for example that the 99% confidence interval for the `price.index` slope runs from -0.363 to -0.256.

5.10 Type I and Type II errors in decision making

Since data analysis is about probabilities, there is always a chance that you make the wrong decision: you can wrongfully reject the null-hypothesis, or you can wrongfully fail to reject the null-hypothesis. Pearson and Neyman distinguished between two kinds of error: one could reject the null-hypothesis while it is actually true (error of the first kind, or type I error) and one could accept the null-hypothesis while it is not true (error of the second kind, or type II error). We already discussed these types of error in Chapter 2. Table 5.3 gives an overview.

Table 5.3: Four different scenarios for hypothesis tests.

		Test conclusion	
		do not reject H_0	reject H_0
Truth	H_0 true	OK	Type I Error
	H_A true	Type II Error	OK

To illustrate the difference between type I and type II errors, let's recall the famous fable by Aesop about the boy who cried wolf. The tale concerns a shepherd boy who repeatedly tricks other people into thinking a wolf is attacking his flock of sheep. The first time he cries "There is a wolf!", the men working in

an adjoining field come to help him. But when they repeatedly find there is no wolf to be seen, they realise they are being fooled by the boy. One day, when a wolf *does* appear and the boy again calls for help, the men believe that it is another false alarm and the sheep are eaten by the wolf.

In this fable, we can think of the null-hypothesis as the hypothesis that there is no wolf. The alternative hypothesis is that there is a wolf. Now, when the boy cries wolf the first time, there is in fact no wolf. The men from the adjoining field make a type I error: they think there is a wolf while there isn't. Later, when they are fed up with the annoying shepherd boy, they don't react when the boy cries "There is a wolf!". Now they make a type II error: they think there is no wolf, while there actually is a wolf. See Table 5.4 for the overview.

		Men in the field	
		Think there is no wolf	Think there is a wolf
Truth	There is no wolf	OK	waste of time and energy
	There is a wolf	devoured sheep	OK

Table 5.4: Four different scenarios for wolves and men working in the field.

Let's now discuss these errors in the context of linear models. Suppose you want to determine the slope for the effect of age on height in children. Let the slope now stand for the wolf: either there is no slope (no wolf, H_0) or there is a slope (wolf, H_A). The null-hypothesis is that the slope is 0 in the population of all children (a slope of 0 means there is no slope) and the alternative hypothesis that the slope is not 0, so there is a slope. You might study a sample of children and you might find a certain slope. You might decide that if the p-value is below a critical value you conclude that the null-hypothesis is not true. Suppose you think a probability of 10% is small enough to reject the null-hypothesis as true. In other words, if $p \leq 0.10$ then we no longer think 0 is a reasonable value for the population slope. In this case, we have fixed our α or type I error rate to be $\alpha = 0.10$. This means that if we study a random sample of children, we look at the slope and find a p-value of 0.11, then we do not reject the null-hypothesis. If we find a p-value of 0.10 or less, then we reject the null-hypothesis.

Note that the probability of a type I error is the same as our α for the significance level. Suppose we set our $\alpha = 0.05$. Then for any p-value equal or smaller than 0.05, we reject the null-hypothesis. Suppose the null-hypothesis is true, how often do we then find a p-value smaller than 0.05? We find a p-value smaller than 0.05 if we find a t-value that is above a certain threshold. For instance, for the t-distribution with 198 degrees of freedom, the critical value is ± 1.97, *because only in 5% of the cases we find a t-value of* ± 1.97 *or more if the null-hypothesis is true*! Thus, if the null-hypothesis is true, we see a t-value of at least ± 1.97 in 5% of the cases. Therefore, we see a significant p-value in 5% of the cases if the null-hypothesis is true. This is exactly the definition of a Type I error: the probability that we reject the null-hypothesis (finding a significant p-value), given that the null-hypothesis is true. So we call our α-value the type I error rate.

Suppose 100 researchers are studying a particular slope. Unbeknownst to them, the population slope is exactly 0. They each draw a random sample from the population and test whether their sample slope is significantly different from 0. Suppose they all use different sample sizes, but they all use the same α of 0.05. Then we can expect that about 5 researchers will reject the null-hypothesis (finding a p-value less than or smaller than 0.05) and about 95 will not reject the null-hypothesis (finding a p-value of more than 0.05).

Fixing the type I error rate should always be done *before* data collection. How willing are you to take a risk of a type I error? You are free to make a choice about α, as long as you do it before looking at the data, and report what value you used.

If α represents the probability of making a type I error, then we can use β to represent the opposite: the probability of not rejecting the null-hypothesis while it is not true (type II error, thinking there is no wolf while there is). However, setting the β-value prior to data collection is a bit trickier than choosing your α. It is not possible to compute the probability that we find a non-significant effect ($p > \alpha$), given that the alternative hypothesis is true, because the alternative hypothesis is only saying that the slope is not equal to 0. In order to compute β, we need to think first of a reasonable size of the slope that we expect. For example, suppose we believe that a slope of 1 is quite reasonable, given what we know about growth in children. Let that be our alternative hypothesis:

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 = 1$$

Next, we determine the distribution of sample slopes under the assumption that the population slope is 1. We know that this distribution has a mean of 1 and a standard deviation equal to the standard error. We also know it has the shape of a t-distribution. Let sample size be equal to 102 and the standard error 2. If we standardise the slopes by dividing by the standard error, we get the two t-distributions in Figure 5.16: one distribution of t-values if the population slope is 0 (centred around $t = 0$), and one distribution of t-values if the population slope is 1 (centred around $t = 1/2 = 0.5$).

Let's fix α to 10%. The shaded areas represent the area where $p \leq \alpha$: for all values of t smaller than -1.6859545 and larger than 1.6859545, we reject the null-hypothesis. The probability that this happens, *if the null-hypothesis is true*, is equal to α, which is 0.10 in this example. The probability that this happens *if the alternative hypothesis is true* (i.e., population slope is 1), is depicted in Figure 5.17.

The shaded area in Figure 5.17 turns out to be 0.1415543. This represents the probability that we find a significant effect, *if the population slope is 1*. This is actually 1 minus the probability of finding a *non*-significant effect, *if the population slope is 1*, which is defined as β. Therefore, the shaded area in Figure 5.17 represents $1 - \beta$: the probability of finding a significant p-value, if the population slope is 1. In this example, $1 - \beta$ is equal to 0.1415543, so β is

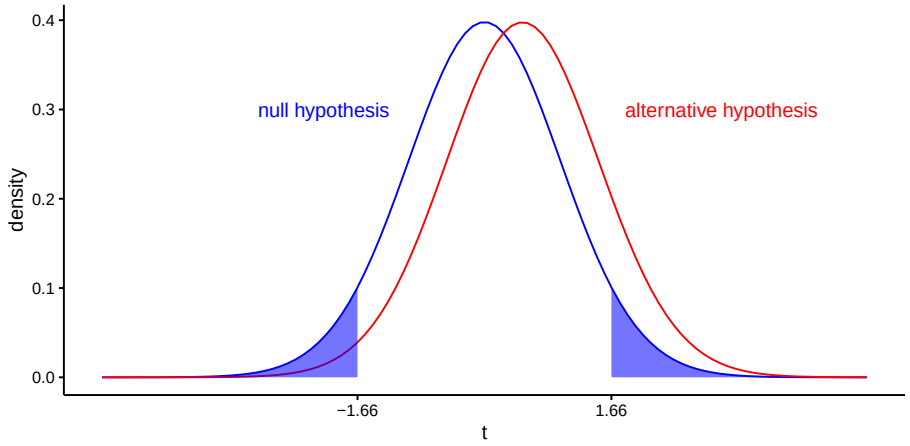


Figure 5.16: Different t-distributions of the sample slope if the population slope equals 0 (left curve in blue), and if the population slope equals 1 (right curve in red). Blue area depicts the probability that we find a p-value value smaller than 0.10 if the population slope is 0 (α).

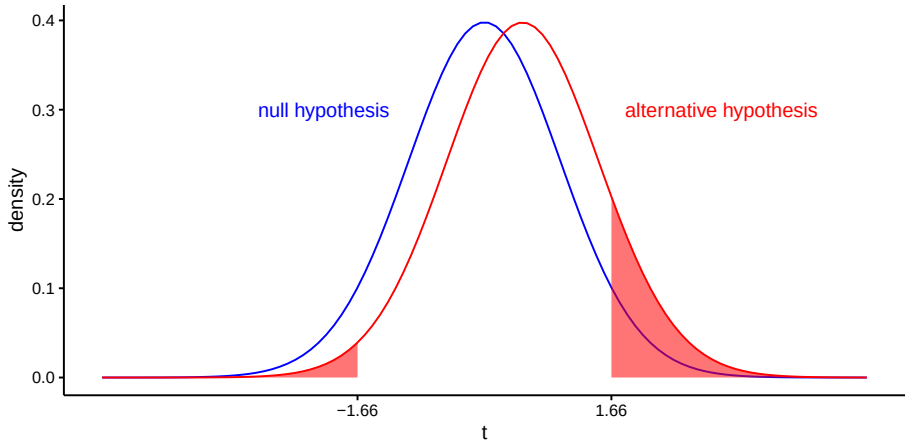


Figure 5.17: Different t-distributions of the sample slope if the population slope equals 0 (left curve in blue), and if the population slope equals 1 (right curve in red). Shaded area depicts the probability that we find a p-value value smaller than 0.10 if the population slope is 1 ($1 - \beta$).

equal to $1 - 0.1415543 = 0.8584457$.

In sum, in this example with an α of 0.10 and assuming a population slope of 1, we find that the probability of a type II error is 0.86: if there is a slope of 1, then we have an 86% chance of wrongly concluding that the slope is 0.

Type I and II error rates α and β are closely related. If we feel that a significance level of $\alpha = 0.10$ is too high, we could choose a level of 0.01. This ensures that we are less likely to reject the null-hypothesis when it is true. The critical value for our t-statistic is then equal to ± 2.6258905, see Figure 5.18. In Figure 5.19 we see that if we change α, we also get a different value for $1 - \beta$, in this case 0.0196567.

Table 5.5 gives an overview of how α and β are related to type I and type II error rates. If a p-value for a statistical test is equal to or smaller than a pre-chosen significance level α, the probability of a type I error equals α. The probability of a type II error rate is equal to β.

		Statistical outcome	
		$p > \alpha$	$p \leq \alpha$
Truth	H_0	$1 - \alpha$	α
	H_A	β	$1 - \beta$

Table 5.5: The probabilities of a statistical outcome under the null-hypothesis and the alternative hypothesis.

Thus, if we use smaller values for α, we get smaller values for $1-\beta$, so we get larger values for β. This means that if we lower the probability of rejecting the null-hypothesis given that it is true (type I error) by choosing a lower value for α, we inadvertently increase the probability of failing to reject the null-hypothesis given that it is not true (type II error).

Think again about the problem of the sheep and the wolf. Instead of the boy, the men could choose to put a very nervous person on watch, someone very scared of wolves. With the faintest hint of a wolf's presence, the man will call out "Wolf!". However, this will lead to many false alarms (type I errors), but the men will be very sure that when there actually is a wolf, they will be warned. Alternatively, they could choose to put a man on watch that is very laid back, very relaxed, but perhaps prone to nod off. This will lower the risk of false alarms immensely (no more type I errors) but it will dramatically increase the risk of a type II error!

One should therefore always strike a balance between the two types of errors. One should consider how bad it is to think that the slope is not 0 while it is, and how bad it is to think that the slope is 0, while it is not. If you feel that the first mistake is worse than the second one, then make sure α is really small, and if you feel that the second mistake is worse, then make α not too small. Another option, and a better one, to avoid type II errors, is to increase sample size, as we will see in the next section.

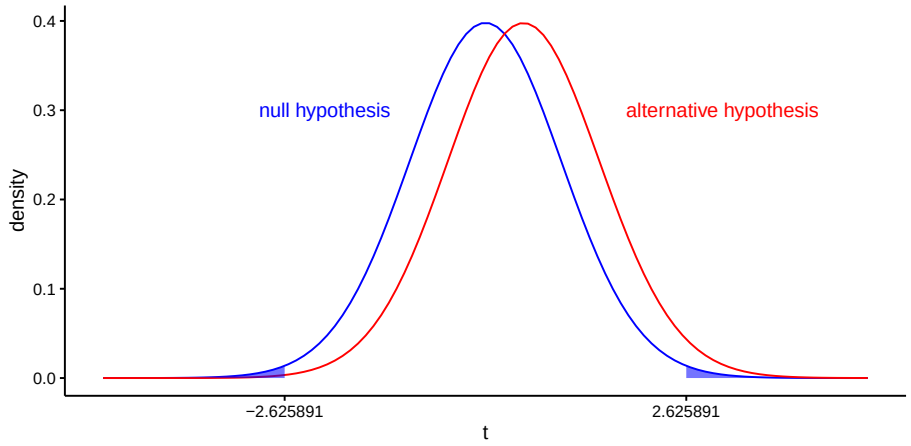


Figure 5.18: Different t-distributions of the sample slope if the population slope equals 0 (left curve), and if the population slope equals 1 (right curve). Blue area depicts the probability that we find a p-value value smaller than 0.01 if the population slope is 0.

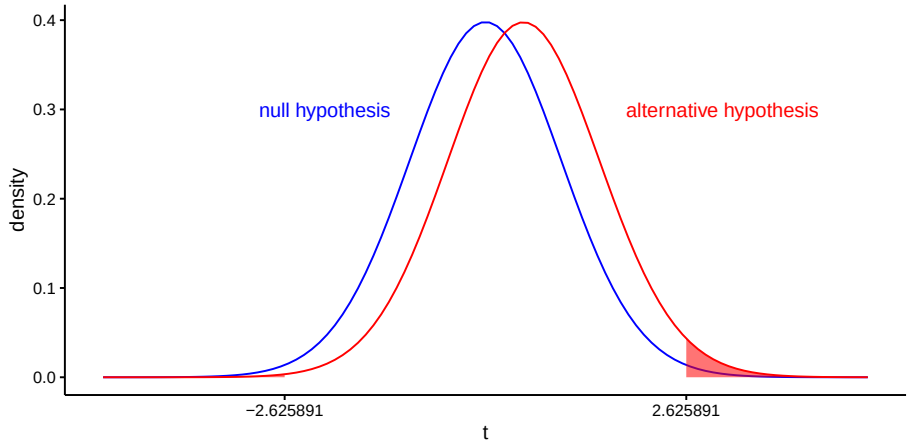


Figure 5.19: Different t-distributions of the sample slope if the population slope equals 0 (left curve in blue), and if the population slope equals 1 (right curve in red). Red area depicts the probability that we find a p-value value smaller than 0.01 if the population slope is 1: $1 - \beta$.

5.11 Statistical power

Null-hypothesis testing only involves the null-hypothesis: we look at the sample slope, compute the t-statistic and then see how often such a t-value and larger values occur given that the population slope is 0. Then we look at the p-value and if that p-value is smaller than or equal to α, we reject the null-hypothesis. Therefore, null-hypothesis testing does not involve testing the alternative hypothesis. We can decide what value we choose for our α, but not our β. The β is dependent on what the actual population slope is, and we simply don't know that.

As stated in the previous section, we can compute β only if we have a more specific idea of an alternative value for the population slope. We saw that we needed to think of a reasonable value for the population slope that we might be interested in. Suppose we have the intuition that a slope of 1 could well be the case. Then, we would like to find a p-value of less than α if indeed the slope were 1. We hope that the probability that this happens is very high: the conditional probability that we find a t-value large enough to reject the null-hypothesis, given that the population slope is 1. This probability is actually the *complement* of β, $1-\beta$: the probability that we reject the null-hypothesis, given that the alternative hypothesis is true. This $1-\beta$ is often called the *statistical power* of a null-hypothesis test. When we think again about the boy who cried wolf: the power is the probability that the men think there is a wolf if there is indeed a wolf. The power of a test should always be high: if there is a population slope that is not 0, then of course you would like to detect it by finding a significant t-value!

In order to get a large value for $1-\beta$, we should have large t-values in our data-analysis. There are two ways in which we can increase the value of the t-statistic. Since with null-hypothesis testing $t = \frac{b-0}{\sigma_{\hat{b}}} = \frac{b}{\sigma_{\hat{b}}}$, we can get large values for t if 1) we have a small standard error, $\sigma_{\hat{b}}$, or 2) if we have a large value for b.

Let's first look at the first option: a small standard error. We get a small standard error if we have a large sample size, see Section 5.2.1. If we go back to the example of the previous section where we had a sample size of 102 children and our alternative hypothesis was that the population slope was 1, we found that the t-distribution for the alternative hypothesis was centred around 0.5, because the standard error was 2. Suppose that we would increase sample size to 1200 children, then our standard error might be 0.2. Then our t-distribution for the alternative hypothesis is centred at 5. This is shown in Figure 5.20.

We see from the shaded area that if the population slope is really 1, there is a very high chance that the t-value for the sample slope will be larger than 2.58, the cut-off point for an α of 0.01 and 1198 degrees of freedom. The probability of rejecting the null-hypothesis while it is not true, is therefore very large. This is our $1-\beta$ and we call this the power of the null-hypothesis test. We see that with increasing sample size, the power to find a significant t-value increases too.

Now let us look at the second option, a large value of b. Sample slope b_1

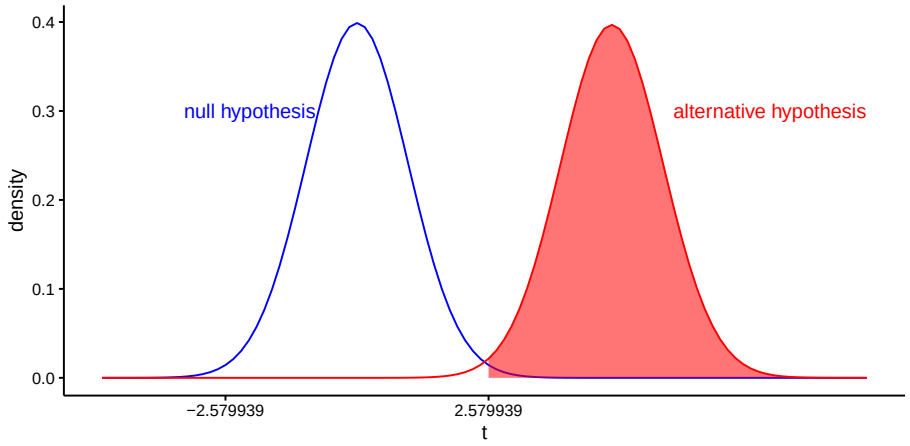


Figure 5.20: Different t-distributions of the sample slope if the population slope equals 0 (left curve in blue), and if the population slope equals 1 (right curve in red). Now for a larger sample size. Shaded area depicts the probability that we find a p-value value smaller than 0.01 if the population slope is 1.

depends of course on the population slope β_1. The power becomes larger when the population slope is further away from zero. If the population slope were 10, and we only had a sample of 102 children (resulting in a standard error of 2), the t-distribution for the alternative hypothesis that the population slope is centred around $\frac{b}{\sigma_b} = 10/2 = 5$, resulting in the same plot as in Figure 5.20, with a large value for $1 - \beta$. Unfortunately, the population slope is beyond our control: the population slope is a given fact that we cannot change. The only thing we can change most of the times is sample size.

In sum: the statistical power of a test is the probability that the null-hypothesis is rejected, given that it is not true. This probability is equal to $1 - \beta$. The statistical power of a test increases with sample size, and depends on the actual population slope. The further away the population slope is from 0 (positive or negative), the larger the statistical power. Earlier we also saw that $1 - \beta$ increases with increasing α: the larger α, the higher the power.

5.12 Power analysis

Because of these relationships between statistical power, α, sample size n, and the actual population slope β_1, we can compute the statistical power for any combination thereof.

If you really care about the quality of your research, you carry out a *power analysis* prior to collecting data. With such an analysis you can find out how large your sample size should be. You can find many tools online that can help you with that.

Suppose you want to minimise the probability of a type I error, so you choose an $\alpha = 0.01$. Next, you think of what kind of population slope you would like to find, if it indeed has that value. You could perhaps base this expectation on earlier research. Suppose that you feel that if the population slope is 0.15, you would really like to find a significant t-value so that you can reject the null-hypothesis. Next, you have to specify how badly you want to reject the null-hypothesis if indeed the population slope is 0.15. If the population slope is really 0.15, then you would like to have a high probability to find a t-value large enough to reject the null-hypothesis. This is of course the power of the test, $1 - \beta$. Let's say you want to have a power of 0.90. Now you have enough information to calculate how large your sample size should be.

Let's look at G*power[4], an application that can be downloaded from the web. If we start the app, we can ask for the sample size required for a slope of 0.15, an α of 0.01, and a power $(1 - \beta)$ of 0.90. Let the standard deviation of our dependent variable (Y) be 3 and the standard deviation of our independent variable (X) be 2. These numbers you can guess, preferably based on some other data that were collected earlier. Then we get the input as displayed in Figure 5.21. Note that you should use two-sided p-values, so `tails = two`. From the output we see that the required sample size is 1477 children.

5.13 Criticism on null-hypothesis testing and p-values

The practice of null-hypothesis significance testing (NHST) is widespread. However, from the beginning it has received much criticism. One of the first to criticise the approach was the inventor of the p-value, Sir Ronald Fisher himself. Fisher explicitly contrasted the use of the p-value for statistical inference in science with the Pearson-Neyman approach, which he termed "Acceptance Procedures". Whereas in the Pearson-Neyman approach the only relevance of the p-value is whether it is smaller or larger than the fixed significance level α, Fisher emphasised that the exact p-value should be reported to indicate the strength of evidence against the null-hypothesis. He emphasised that no single p-value can refute a hypothesis, since chance always allows for type I and type II errors. Conclusions can and will be revised with further experimentation; science requires more than one study to reach solid conclusions. Decision procedures with clear-cut decisions based on one study alone hamper science and lead to tunnel-vision.

Apart from these science-theoretical considerations of the NHST, there are also practical reasons why pure NHST should be avoided. In at least a number of research fields, the p-value has become more than just the criterion for finding an effect or not: it has become the criterion of whether the research is publishable or not. Editors and reviewers of scientific journals have increasingly interpreted a study with a significant effect to be more interesting than a study

[4] http://www.gpower.hhu.de/

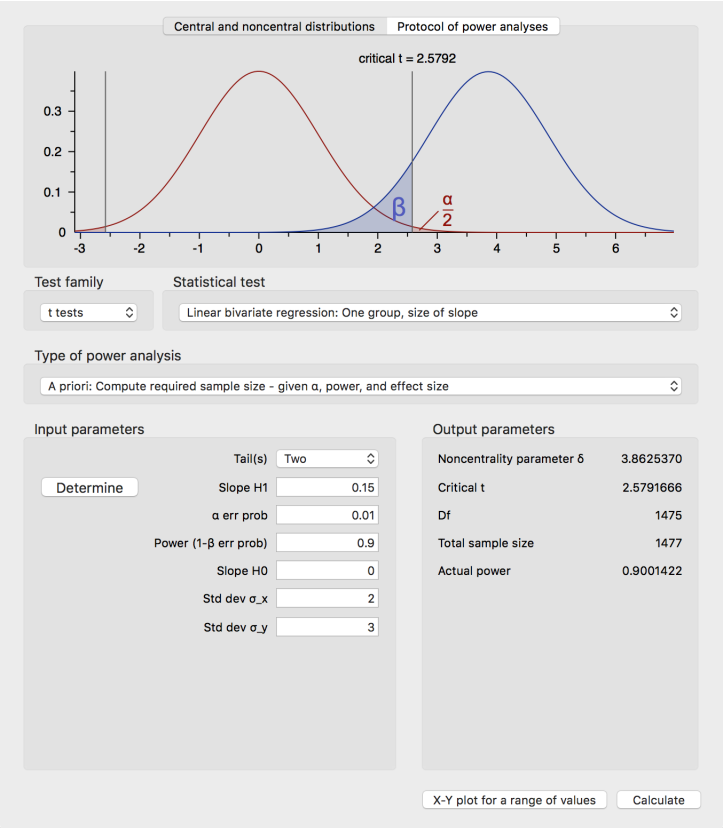


Figure 5.21: G*power output for a simple regression analysis.

with a non-significant effect. For that reason, in scientific journals you will find mostly studies reported with a significant effect. This has led to *the file-drawer problem*: the literature reports significant effects for a particular phenomenon, but there can be many unpublished studies with non-significant effects for the same phenomenon. These unpublished studies remain unseen in file-drawers (or these days on hard-drives). So based on the literature there might seem to exist a particular phenomenon, but if you would put all the results together, including the unpublished studies, the effect might disappear completely.

Remember that if the null-hypothesis is true and everyone uses an α of 0.05, then out of 100 studies of the same phenomenon, only 5 studies will be significant and are likely to be published. The remaining 95 studies with insignificant effects are more likely to remain invisible.

As a result of this bias in publication, scientists who want to publish their results are tempted to fiddle around a bit more with their data in order to get a significant result. Or, if they obtain a p-value of 0.07, they decide to increase

their sample size, and perhaps stop as soon as the p-value is 0.05 or less. This horrible malpractice is called *p-hacking* and is extremely harmful to science. As we saw earlier, if you want to find an effect and not miss it, you should carry out a power analysis *before* you collect the data and make sure that your sample size is large enough to obtain the power you want to have. Increasing sample size *after* you have found a non-significant effect increases your type I error rate dramatically: if you stop collecting data *until* you find a significant p-value, the type I error rate is equal to 1!

There have been wide discussions the last few years about the use and interpretation of p-values. In a formal statement, the American Statistical Association published six principles that should be well understood by anyone, including you, who uses them.

The six principles are:

1. p-values can indicate how incompatible the data are with a specified statistical model (usually the null-hypothesis).

2. p-values *do not* measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone. Instead, they measure how likely it is to find a sample slope of at least the size that you found, given that the population slope is 0.

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold. For instance, also look at the size of the effect: is the slope large enough to make policy changes worth the effort? Have other studies found effects of similar sizes?

4. Proper inference requires full reporting and transparency. Always report your sample slope, the standard error, the t-statistic, the degrees of freedom, and the p-value. Only report about null-hypotheses that your study was designed to test.

5. A p-value or statistical significance *does not* measure the size of an effect or the importance of a result. (See principle 1)

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis. At least as important is the design of the study.

These six principles are further explained in the statement online[5]. The bottom line is, p-values have worth but only when used and interpreted in a proper way. Some disagree. The philosopher of science William Rozeboom once called NHST "surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students." The scientific journal *Basic and Applied Social Psychology* even banned NHST altogether: t-values and p-values are not allowed if you want to publish your research in that journal.

[5] https://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108

Most researchers now realise that reporting confidence intervals is often a lot more meaningful than reporting whether a p-value is significant or not. A p-value only says something about evidence against the hypothesis that the slope is 0. In contrast, a confidence interval gives a whole range of reasonable values for the population slope. If 0 lies within the confidence interval, then 0 is a reasonable value; if it is not, then 0 is not a reasonable value so that we can reject the null-hypothesis.

Using confidence intervals also counters one fundamental problem of null-hypotheses: nobody believes in them! Remember that the null-hypothesis states that a particular effect (a slope) is exactly 0: not 0.0000001, not -0.000201, but exactly 0.00000000000000000000.

Sometimes a null-hypothesis doesn't make sense at all. Suppose we are interested to know what the relationship is between age and height in children. Nobody believes that the population slope coefficient for the regression of height on age is 0. Why then test this hypothesis? More interesting would be to know *how large* the population slope is. A confidence interval would then be much more informative than a simple rejection of the null-hypothesis.

In some cases, a null-hypothesis can be slightly more meaningful: suppose you are interested in the effect of cognitive behavioural therapy on depression. You hope that the number of therapy sessions has a negative effect on the severity of the depression, but it is entirely possible that the effect is very close to non-existing. Of course you can only look at a sample of patients and determine the sample slope. But think now about the population slope: think about all patients in the world with depression that theoretically could partake in the research. Some of them have 0 sessions, some have 1 session, and so on. Now imagine that there are 1 million of such people. How likely is it that in the population, the slope for the regression is exactly 0? Not 0.00000001, not -0.0000000002, but exactly 0.0000000000. Of course, this is extremely unlikely. The really interesting question in such research is whether there is a *meaningful* effect of therapy. For instance, an effect of at least half a point decrease on the Hamilton depression scale for 5 sessions. That would translate to a slope of $\frac{-0.5}{5} = -0.1$. Also in this case, a confidence interval for the effect of therapy on depression would be more helpful than a simple p-value. A confidence interval of -2.30 to -0.01 says that a small population effect of -0.01 might be there, but that an effect of -0.0001 or 0.0000 is rather unlikely. It also states that a meaningful effect of at least -0.1 is likely. You can then conclude that the therapy is helpful. The p-value less than α only tells you that a value of exactly 0.0000 is not realistic, but who cares.

So, instead of asking research questions like "Is there a linear relationship between x and y?" you might ask: "How large is the linear effect of x on y?" Instead of a question like "Is there an effect of the intervention?" it might be more interesting to ask: "How large is the effect of the intervention?"

Summarising, remember the following principles when doing your own research or evaluating the research done by others:

- Inference about a population slope or intercept can be made on the basis

of sample data, but only in probabilistic terms. This means that a simple statement like "the value of the population slope is definitely not zero" cannot be made. Only statements like "A population slope of 0 is not very likely given the sample data" can be made.

- Science is cumulative. No study is definitive. Effects should be replicated by independent researchers.

- Always report your regression slope or intercept, with the standard error and the sample size. Based on these, the t-statistics can be computed with the degrees of freedom. Then if several other researchers have done the same type of research, the results can be combined in a so-called meta-analysis, so that a stronger statement about the population can be made, based on a larger total sample size. The standard error and sample size moreover allow for the construction of confidence intervals. But better is to report confidence intervals yourself.

5.14 Relationship between p-values and confidence intervals

In previous sections we stated that if the value 0 lies within a confidence interval, it is a reasonable value for the population slope. If 0 is not within the interval, 0 is not a reasonable value for the population slope, so we have to reject the null-hypothesis. Here we will elaborate a little on this theme.

Both the confidence interval and the p-value are based on the same t-distribution. Suppose we set our α to 0.05, and our sample size is 102. This means that if we find a p-value $p \leq 0.05$ we reject the null-hypothesis that the slope is 0. The p-value depends on how many standard deviations our sample slope deviates from 0. We calculate this by computing a standardised slope. For example, for a sample slope of 1 and a standard error of 0.5, our standardised slope is $t = (1 - 0)/0.5 = 2$. In other words, our sample slope of 1 is 2 standard errors away from 0. From t-tables, we know that with 100 degrees of freedom, the 2.5th and 97.5th percentiles are -1.98 and 1.98, respectively (see Appendix B). Therefore, the p-value depends on the size of the t-statistic. If it is equal to -1.98 or 1.98, the p-value is exactly 0.05. If the t-statistic is smaller than -1.98 or larger than 1.98, the p-value is smaller than 0.05.

The values -1.98 and 1.98 are also used for the construction of the 95% confidence interval. The lower bound lies at 1.98 times the standard error below the sample slope, and the upper bound lies at 1.98 times above the sample slope. Therefore, if 0 lies more than 1.98 standard errors away from the mean, it lies outside the confidence interval. But if 0 lies more than 1.98 standard errors away from the mean, this implies that the sample slope lies more than -1.98 standard errors away from 0, which corresponds to a t-statistic of more than ± 1.98. Thus, if 0 is not within the 95% confidence interval, we know that the p-value is smaller than 0.05.

Using the same reasoning as above, we also know that if 0 is not within the 99% confidence interval, we know that the p-value is smaller than 0.01, and if 0 is not within the 99.9% confidence interval, we know that the p-value is smaller than 0.001, etcetera.

A 95% confidence interval can therefore also be seen as the range of possible values for the null-hypothesis that cannot be rejected with an α of 5%. By the same token, a 99% confidence interval can be seen as the range of possible values for the null-hypothesis that cannot be rejected with an α of 1%, etcetera.

5.15 The intercept only model

So far in this chapter, we have only discussed inference regarding the linear model with both an intercept and one or more slope parameters.

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + e \tag{5.12}$$

Remember that in Chapter 2 we discussed inference regarding only a mean. Here we show that inference regarding the mean can also be done within the linear model framework. In Chapter 2 we wanted to get a confidence interval for the mean luteinising hormone (LH) for a woman. We had 48 measures ($n = 48$) and the sample mean was 2.4. We computed the standard error as $\sqrt{\frac{s^2}{n}} = 0.0796$, so that we could construct a confidence interval using a t-distribution of $48 - 1 = 47$ degrees of freedom. In Chapter 2 we saw that we can compute a 95% confidence interval for a population mean as

```
t.test(lh, conf.level = 0.95)$conf.int
```

```
## [1] 2.239834 2.560166
## attr(,"conf.level")
## [1] 0.95
```

Here we show that the same inference can be done with a very simple version of the linear model: an intercept-only model. An intercept-only model has only an intercept and no slopes.

$$\begin{aligned} Y &= b_0 + e \\ e &\sim N(0, \sigma^2) \end{aligned} \tag{5.13}$$

We briefly discussed such a model when we discussed degrees of freedom. This model says that the predicted/expected Y-value for any observation, is equal to b_0, with residuals e that are normally distributed. On average they are 0, and that implies that their sum is equal to 0.

We know that when we have a bunch of Y-values, and we compute the mean, the deviations between the Y-values and the mean also sum to 0. As a very simple example, if we observe the Y-values 4, 5 and 6, the mean is 5. When we

take the deviations between this mean of 5 and the Y-values, we get -1, 0 and 1. And these sum to 0. This is true for any set of Y-values. Thus, we could use the mean of Y as our estimate for b_0, since then the deviations with the mean (i.e., the residuals) sum to 0.

Earlier we said that the unbiased estimator of the population mean is the sample mean. Therefore, our b_0 parameter represents the unbiased estimator of the population mean of Y. Let's see if this works by fitting this model in R. In R, an intercept is indicated by a `1`:

```
library(broom)
data(lh)
out <- lh %>%
  lm(lh ~ 1, data = .)
out %>%
  tidy(conf.int = 0.95)

## # A tibble: 1 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)      2.4    0.0796      30.1 2.14e-32     2.24      2.56
```

In the output we only see an intercept. It is equal to 2.4, which is also the mean of LH as we saw earlier. The standard error is also exactly the same as we computed by hand (0.0796), as is the 95% confidence interval. We get the same results, because in both cases, we use exactly the same standard error and the same t-distribution with $n - K - 1 = 48 - 0 - 1 = 47$ degrees of freedom (K equals 0, the number of independent variables).

In summary, inference about the mean of a set of values can be done using an intercept-only linear model.

Chapter 6

Categorical predictor variables

6.1 Dummy coding

As we have seen in Chapter 1, there are largely two different types of variables: numeric variables and categorical variables. Numeric variables say something about *how much* of an attribute is in an object: for instance height (measured in inches) or heat (measured in degrees Celsius). Categorical variables say something about the quality of an attribute: for instance colour (red, green, yellow) or type of seating (aisle seat, window seat). We have also seen a third type of variable: ordinal variables. Ordinal variables are somewhat in the middle between numeric variables and categorical variables: they are about quantitative differences between objects (e.g., size) but the values are sharp disjoint categories (small, medium, large), and the values are not expressed using units of measurements.

In the chapters on simple and multiple regression we saw that both the dependent and the independent variables were all numeric. The linear model used in regression analysis always involves a numeric dependent variable. However, in such analyses it is possible to use categorical independent variables. In this chapter we explain how to do that and how to interpret the results.

The basic trick that we need is *dummy coding*. Dummy coding involves making one or more new variables, that reflects the categorisation seen with a categorical variable. First we focus on categorical variables with only two categories (dichotomous variables). Later in this chapter, we will explain what to do with categorical variables with more than two categories (nominal variables).

Imagine we study bus companies and there are two different types of seating in buses: aisle seats and window seats. Suppose we ask 5 people, who have travelled from Amsterdam to Paris by bus during the last 12 months, whether they had an aisle seat or a window seat during their last trip, and how much they paid for the trip. Suppose we have the variables `person`, `seat` and `price`.

Table 6.1 shows the anonymised data. There we see the dichotomous variable `seat` with values 'aisle' and 'window'.

Table 6.1: Bus trips to Paris.

person	seat	price
001	aisle	57.00
002	aisle	59.00
003	window	68.00
004	window	60.00
005	aisle	61.00

With dummy coding, we make a new variable that only has values 0 and 1, that conveys the same information as the `seat` variable. The resulting variable is called a *dummy variable*. Let's call this dummy variable `window` and give it the value 1 for all persons that travelled in a window seat. We give the value 0 for all persons that travelled in an aisle seat. We can also call the new variable `window` a *boolean variable* with TRUE and FALSE, since in computer science, TRUE is coded by a 1 and FALSE by a 0. Another name that is sometimes used is an *indicator variable*. Whatever you want to call it, the data matrix including the new variable is displayed in Table 6.2.

Table 6.2: Bus trips to Paris.

person	seat	window	price
001	aisle	0	57.00
002	aisle	0	59.00
003	window	1	68.00
004	window	1	60.00
005	aisle	0	61.00

What we have done now is coding the old categorical variable `seat` into a variable `window` with values 0 and 1 that looks numeric. Let's see what happens if we use a linear model for the variables `price` (dependent variable) and `window` (independent variable). The linear model is:

$$\texttt{price} = b_0 + b_1\texttt{window} + e \tag{6.1}$$

$$e \sim N(0, \sigma_e^2) \tag{6.2}$$

Let's use the bus trip data and determine the least squares regression line. We find the following linear equation:

$$\widehat{\texttt{price}} = 59 + 5 \times \texttt{window} \tag{6.3}$$

If the variable `window` has the value 1, then the expected or predicted price of the bus ticket is, according to this equation, $59 + 5 \times 1 = 64$. What does

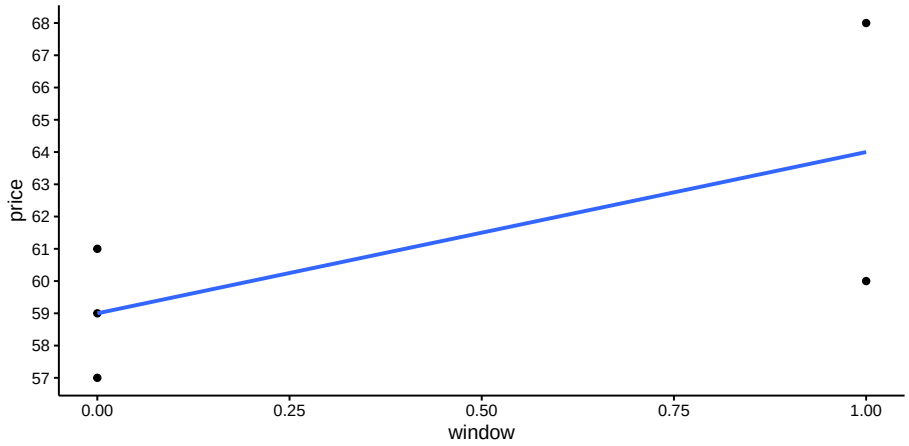


Figure 6.1: Relationship between dummy variable window and price.

this mean? Well, all persons who had a window seat also had a value of 1 for the `window` variable. Therefore the expected price of a window seat equals 64. By the same token, the expected price of an aisle seat (`window = 0`) is $59 + 5 \times 0 = 59$, since all those with an aisle seat scored 0 on the `window` variable.

You see that by coding a categorical variable into a numeric dummy variable, we can describe the 'linear' relationship between the type of seat and the price of the ticket. Figure 6.1 shows the relationship between the numeric variable `window` and the numeric variable `price`.

Note that the blue regression line goes straight through the mean of the prices for window seats (`window = 1`) and the mean of the prices for aisle seats (`window = 0`). In other words, the linear model with the dummy variable actually models the *group means* of people with window seats and people with aisle seats.

Figure 6.2 shows the same regression line but now for the original variable `seat`. Although the analysis was based on the dummy variable `window`, it is more readable for others to show the original categorical variable `seat`.

6.2 Using regression to describe group means

In the previous section we saw that if we replace a categorical variable with a numeric dummy variable with values 0 and 1, we can use a linear model to describe the relationship between a categorical independent variable and a numeric dependent variable. We also saw that if we take the least squares regression line, this line goes straight through the averages, the group means. The line goes straight through the group means because then the sum of the squared residuals is at its smallest value (the least squares principle). Have a look at the bus trip data again in Figure 6.1 and see if you can derive the

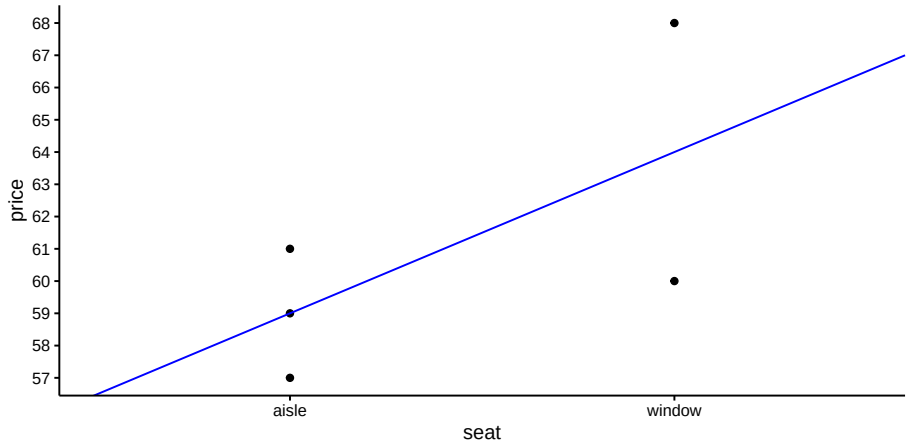


Figure 6.2: Relationship between type of seat and price.

residuals and the squared residuals. These are displayed in Table 6.3.

Table 6.3: Bus trip to Paris data, together with residuals and squared residuals from the least squares regression line.

person	seat	window	price	e	e_squared
001	aisle	0.00	57.00	-2.00	4.00
002	aisle	0.00	59.00	0.00	0.00
003	window	1.00	68.00	4.00	16.00
004	window	1.00	60.00	-4.00	16.00
005	aisle	0.00	61.00	2.00	4.00

If we take the sum of the squared residuals we obtain 40. Now if we use a slightly different slope, so that we no longer go straight through the average prices for aisle and window seats (see Figure 6.3) and we compute the predicted values, the residuals and the squared residuals (see Table 6.4), we obtain a higher sum: 40.05.

Only the least squares regression line goes through the observed average prices of aisle seats and window seats. Thus, we can use the least squares regression equation to describe observed group means for categorical variables.

Conversely, when you know the group means, it is very easy to draw the regression line: the intercept is then the mean for the category coded as 0, and the slope is equal to the mean of the category coded as 1 minus the mean of the category coded as 0 (i.e., the intercept). Check Figure 6.1 to verify this yourself. But we can also show this for a new data set.

We look at results from an experiment to compare yields (as measured by dried weight of plants) obtained under a control and two different treatment conditions. Let's plot the data first, where we only compare the two experimental

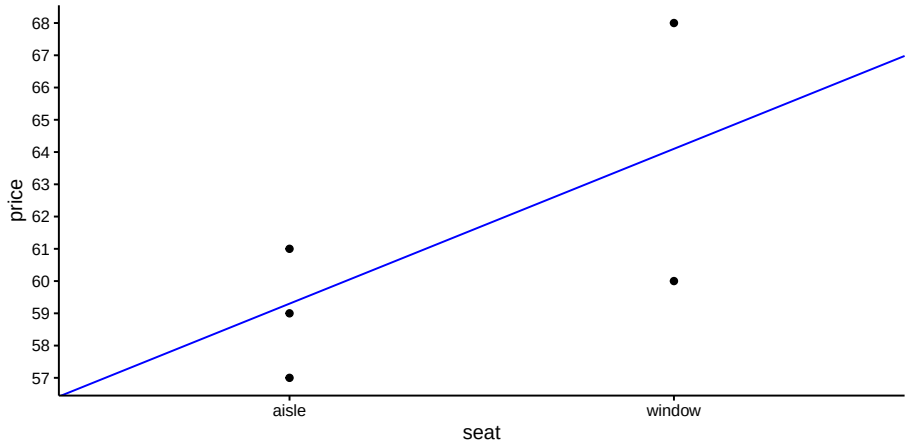


Figure 6.3: Relation between type of seat and price, with the regression line being not quite the least squares line.

Table 6.4: Bus trips to Paris, together with residuals and squared residuals from a suboptimal regression line.

person	seat	window	price	wrongpredict	e	e_squared
001	aisle	0.00	57.00	59.10	-2.10	4.41
002	aisle	0.00	59.00	59.10	-0.10	0.01
003	window	1.00	68.00	63.90	4.10	16.81
004	window	1.00	60.00	63.90	-3.90	15.21
005	aisle	0.00	61.00	59.10	1.90	3.61

conditions (see Figure 6.4).

With treatment 1, the average yield turns out to be 4.661, and with treatment 2, the average yield is 5.526. Suppose we make a new dummy variable `treatment2` that is 0 for treatment 1, and 1 for treatment 2. Then we have the linear equation:

$$\widehat{\texttt{weight}} = b_0 + b_1 \times \texttt{treatment2} \tag{6.4}$$

If we fill in the dummy variable and the expected weights (the means!), then we have the linear equations:

$$4.661 = b_0 + b_1 \times 0 = b_0 \tag{6.5}$$

$$5.526 = b_0 + b_1 \times 1 = b_0 + b_1 \tag{6.6}$$

So from this, we know that intercept $b_0 = 4.661$, and if we fill that in for the

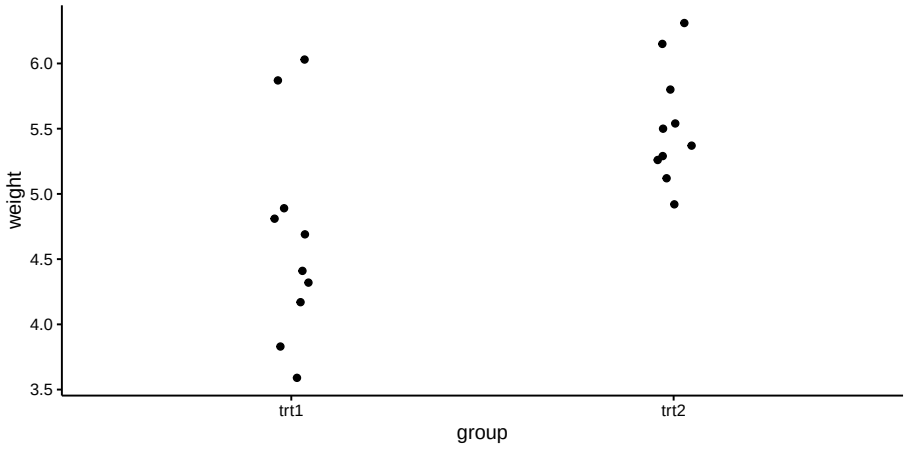


Figure 6.4: Data on yield under two experimental conditions: treatment 1 and treatment 2.

second equation above, we get the slope:

$$b_1 = 5.526 - b_0 = 5.526 - 4.661 = 0.865. \tag{6.7}$$

Thus, we get the linear equation

$$\widehat{\texttt{weight}} = 4.661 + 0.865 \times \texttt{treatment} \tag{6.8}$$

Since this regression line goes straight through the average yield for each treatment, we know that this is the least squares regression equation. We could have obtained the exact same result with a regression analysis using statistical software. But this was not necessary: because we knew the group means, we could find the intercept and the slope ourselves by doing the math.

The interesting thing about a dummy variable is that the slope of the regression line is exactly equal to the differences between the two averages. If we look at Equation 6.8, we see that the slope coefficient is 0.865 and this is exactly equal to the difference in mean weight for treatment 1 and treatment 2. Thus, the slope coefficient for a dummy variable indicates how much the average of the treatment that is coded as 1 differs from the treatment that is coded as 0. Here the slope is positive so that we know that the treatment coded as 1 (trt2), leads to a higher average yield than the treatment coded as 0 (trt1). This makes it possible to draw inferences about differences in group means.

6.3 Making inferences about differences in group means

In the previous section we saw that the slope in a dummy regression is equal to the difference in group means. Suppose researchers are interested in the effects of different treatments on yield. They'd like to know what the difference is in yield between treatments 1 and 2, using a limited sample of 20 data points. Based on this sample, they'd like to generalise to the population of all yields based on treatments 1 and 2. They adopt a type I error rate of $\alpha = 0.05$.

Table 6.5: Yield by treatment.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.66	0.20	22.94	0.00	4.23	5.09
grouptrt2	0.87	0.29	3.01	0.01	0.26	1.47

The researchers analyse the data and they find the regression table as displayed in Table 6.5. The 95% confidence interval for the slope is from 0.26 to 1.47. This means that reasonable values for the *population* difference between the two treatments on yield lie within this interval. All these values are positive, so we reasonably believe that treatment 2 leads to a higher yield than treatment 1. We know that it is treatment 2 that leads to a higher yield, because the slope in the regression equation refers to a variable `grouptrt2` (see Table 6.5). Thus, a dummy variable has been created, `grouptrt2`, where trt2 has been coded as 1 (and trt1 consequently coded as 0). In the next section, we will see how to do this yourself.

If the researchers had been interested in testing a null-hypothesis about the differences in mean yield between treatment 1 and 2, they could also use the 95% confidence interval for the slope. As it does not contain 0, we can reject the null-hypothesis that there is no difference in group means at an α of 5%. The exact p-value can be read from Table 6.5 and is equal to 0.01.

Thus, based on this regression analysis the researchers can write in a report that there is a significant difference between the yield after treatment 1 and the yield after treatment 2, $t(18) = 3.01, p = 0.01$. Treatment 2 leads to a yield of about 0.87 (SE = 0.29) more than treatment 1 (95% CI: 0.26 – 1.47).

6.4 Regression analysis using a dummy variable in R

When your independent variable is a categorical variable, the code that you use in R is the same as with a numeric independent variable. For instance, if you want to predict yield from the treatment group, you could run the following R code:

```
data("Plantgrowth")
```

```
PlantGrowth %>%
filter(group != "ctrl") %>%
lm(weight ~ group, .) %>%
tidy()
```

In this code, we take the `PlantGrowth` data frame that is available in R, we omit the data points from the control group (because we are only interested in the two treatment groups), and we model `weight` as a function of `group`. What then happens depends on the data type of `group`. Let's take a quick look at the variables:

```
PlantGrowth %>%
  select(weight, group) %>%
  str()

## 'data.frame': 30 obs. of  2 variables:
##  $ weight: num  4.17 5.58 5.18 6.11 4.5 4.61 5.17 4.53 5.33 5.14 ...
##  $ group : Factor w/ 3 levels "ctrl","trt1",..: 1 1 1 1 1 1 1 1 1 1 ...
```

We see that the dependent variable `weight` is of type numeric (`num`), and that the independent variable `group` is of type factor. If the independent variable is of type factor, R will automatically make a dummy variable for the factor variable. This will not happen if the independent variable is of type numeric.

So here `group` is a factor variable. Below we see the regression table that results from the linear model analysis.

```
data("PlantGrowth")
out <- PlantGrowth %>%
  filter(group != "ctrl") %>%
  lm(weight ~ group, .)
out %>%
  tidy()

## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    4.66      0.203     22.9  8.93e-15
## 2 grouptrt2      0.865     0.287      3.01 7.52e- 3
```

We no longer see the `group` variable, but we see a new variable called `grouptrt2`. Apparently, this new variable was created by R to deal with the `group` variable being a factor variable. The slope value of 0.865 now refers to the effect of treatment 2, that is, treatment 1 is the reference category and the value 0.865 is the added effect of treatment 2 on the yield. We should therefore interpret these results as that in the sample data, the mean of the treatment 2 data points was 0.865 higher than the mean of the treatment 1 data points.

Here, R automatically picked the treatment 1 group as the reference group. In case you want to have treatment 2 as the reference group, you could make your own dummy variable. For instance, make your own dummy variable `grouptrt1` in the following way and check whether it is indeed stored as numeric in R:

```
PlantGrowth <- PlantGrowth %>%
 mutate(grouptrt1 = ifelse(group == "trt1", 1, 0))

PlantGrowth %>%
  select(weight, group, grouptrt1) %>%
  str()

## 'data.frame': 30 obs. of  3 variables:
##  $ weight   : num  4.17 5.58 5.18 6.11 4.5 4.61 5.17 4.53 5.33 5.14 ...
##  $ group    : Factor w/ 3 levels "ctrl","trt1",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ grouptrt1: num  0 0 0 0 0 0 0 0 0 0 ...
```

Next, you can run a linear model with the `grouptrt1` dummy variable that you created yourself:

```
out <- PlantGrowth %>%
  filter(group != "ctrl") %>%
  lm(weight ~ grouptrt1, .)
out %>%
  tidy()

## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    5.53      0.203     27.2  4.52e-16
## 2 grouptrt1     -0.865     0.287     -3.01 7.52e- 3
```

The results now show the effect of treatment 1, with treatment 2 being the reference category. Of course the effect of -0.865 is now the opposite of the effect that we saw earlier (+0.865), when the reference category was treatment 2. The intercept has also changed, as the intercept is now the expected weight for the other treatment group. In other words, the reference group has now changed: the intercept is equal to the expected weight of the treatment 1 group.

In general, store variables that are essentially categorical as factor variables in R. For instance, you could have a variable `group` that has two values 1 and 2 and that is stored as numeric. It would make more sense to first turn this variable into a factor variable, before using this variable as a predictor in a linear model. You could turn the variable into a factor only for the analysis and leaving the data frame unchanged, like this:

```
model <- dataset %>%
lm(y ~ factor(group), data = .)
```

or change the data frame before the analysis

```
dataset <- dataset %>%
mutate(group = factor(group))

model <- dataset %>%
lm(y ~ group, data = .)
```

R will always choose the category with the lowest internal integer value as the reference category. The internal values are chosen alphabetically. If that however makes the output hard to interpret, think of the best way to code your own dummy variable. For experimental designs, it makes sense to code control conditions as 0, and experimental conditions as 1 (you're often interested in the effect of the experimental condition *compared* to the control condition, so the control condition is the reference group). For social surveys, if you want to compare how a social minority group scores relative to a social majority group, it makes sense to code the minority group as 1 and the social majority as 0. In educational studies, it makes sense to code an old teaching method as 0 and a new method as 1. In all of these cases, the slope can then be interpreted as the difference of the experimental procedure/new method/minority compared to the reference group, and the intercept can be interpreted as the mean of the reference group.

6.5 Two independent variables: one dummy and one numeric variable

In Chapter 4 we saw that we can have more than one predictor variable in a linear model. If we have two or more numeric variables, one usually talks about multiple regression models. When we have a categorical variable that we treat as a numeric dummy variable, then we can therefore also have linear models with both a categorical variable and a numeric variable.

Let's return to the bus trip to Paris data. Suppose that one can choose the amount of leg room. There are seats with 60, 70 or 80 centimetres of leg room. You might expect that seats with more leg room are more expensive. To find out whether this is the case, you analyse the sample data from the 5 travellers. Leg room varies between 60 and 80 centimetres. Since you already know that the price of seats also depends on the type of seat (aisle or window), you analyse the data with the following linear model:

$$\begin{aligned} \texttt{price} &= b_0 + b_1\texttt{window} + b_2\texttt{legroom} + e \\ e &\sim N(0, \sigma^2) \end{aligned} \tag{6.9}$$

That is, your independent variables are the dummy variable `window` (1 coding for window seat, 0 coding for aisle seat) and the numeric variable `legroom`.

Table 6.6: Regression table for the regression of price on the dummy variable window and the numeric variable legroom.

term	estimate	std.error	statistic	p.value
(Intercept)	41.50	9.71	4.27	0.05
window	5.00	2.50	2.00	0.18
leg_room	0.25	0.14	1.83	0.21

When we look at the output, we see the regression table in Table 6.6. When we fill in the coefficients, we obtain the following linear equation:

$$\widehat{\texttt{price}} = 41.5 + 5 \times \texttt{window} + 0.25 \times \texttt{legroom} \tag{6.10}$$

From Chapter 4 we know how to interpret the coefficients. The slope coefficient for `window` should be interpreted as "the increase in price if we change seat from aisle to window, given a certain amount of legroom". That is, holding `legroom` constant, for instance at 60 centimetres, the difference between an aisle and a window seat equals 5 Euros. And of course, this difference is also 5 Euros when legroom equals 70 centimetres, and also when legroom equals 80 centimetres. Along the same vein, the slope coefficient for `legroom` should be interpreted as "the increase in price for every unit increase in `legroom`, given a certain type of seat". Thus, for an aisle seat, you pay 0.25 Euros more for every extra centimetre. And this is also true for window seats: every extra centimetre for your legs costs you 0.25 Euros.

The intercept of 41.5 means that the model predicts that you pay 41.5 if you happen to have 0 centrimeters of leg room and an aisle seat (the reference category). Of course, a seat with 0 leg room does not exist, but it is simply what the model predicts, based on the data that are observed. The data and these predictions are visualised in Figure 6.5.

In order to visualise the relationship between the three variables `price`, `legroom` and `window`, we plotted `legroom` on the horizontal axis, and used different colours for the variable `window`. There are a couple of things you should notice in this figure.

The first you should notice is that there are now two regression lines, one for window seats and one for aisle seats. This is so because the model makes different predictions for window and aisle seats. If we take Equation 6.10 and make predictions for aisle seats, we fill in `window = 0` and we get the following linear equation:

$$\widehat{\texttt{price}} = 41.5 + 5 \times 0 + 0.25 \times \texttt{legroom} = 41.5 + 0.25 \times \texttt{legroom} \tag{6.11}$$

Thus, the regression line for aisle seats has an intercept of 41.5 and a slope of 0.25. Now let's fill in the equation for window seats, that is, `window = 1`. Then we obtain

$$\widehat{\texttt{price}} = 41.5 + 5 \times 1 + 0.25 \times \texttt{legroom} = 46.5 + 0.25 \times \texttt{legroom} \tag{6.12}$$

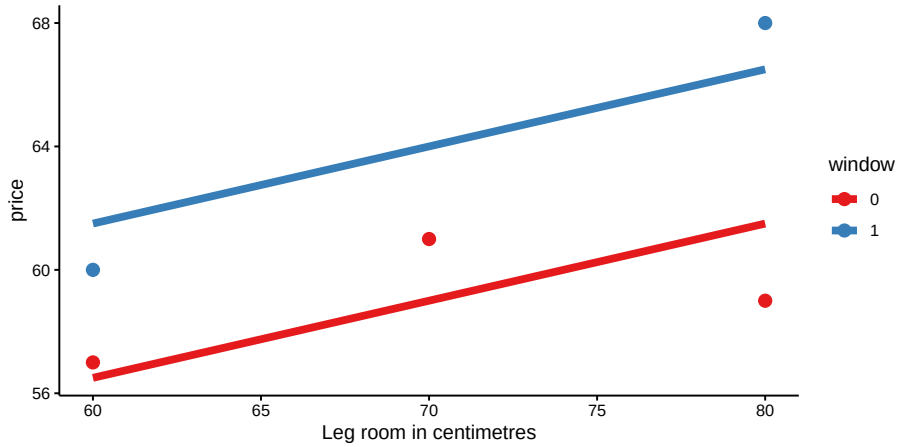


Figure 6.5: The bus trip to Paris data, with the predictions from a linear model with legroom and window as independent variables.

That is, the regression line for window seats has an intercept that is different: it is equal to the original intercept plus the slope of the `window` variable, $41.5 + 5 = 46.5$. On the other hand, the slope for `legroom` is unchanged. With the same slope for `legroom`, the two regression lines are therefore parallel.

The second you should notice from Figure 6.5 is that these two regression lines are not the least squares regression lines for window and aisle seats respectively. For instance, the regression line for window seats (the top one) should be more positive in order to minimise the difference between the data points and the regression line (the residuals). On the other hand, the regression line for aisle seats (the bottom one) should be less steep in order to have smaller residuals. Why is this so? Shouldn't the regression lines minimise the residuals?

Yes they should! But there is a problem, because the model also implies, as we saw above, that the lines are parallel. Whatever we choose for values for the multiple regression equation, the regresion lines for aisle and window seats will always be parallel. And under that constraint, the current parameter values give the lowest possible value for the sum of the squared residuals, that is, the the sum of the squared residuals for both regression lines taken together. The aisle seat regression line should be less steep, and the window seat regression line should be steeper to have a better fit with the data, but taken together, the estimated slope of 0.25 gives the lowest overall sum of squared residuals.

It is possible though to have linear models where the lines are not parallel. This will be discussed in Chapter 9.

Table 6.7: Height across three different countries.

ID	Country	height
001	A	120
002	A	160
003	B	121
004	B	125
005	C	140
...	...	...

Table 6.8: Height across three different countries with dummy variables.

ID	Country	height	countryA	countryB
001	A	120	1	0
002	A	160	1	0
003	B	121	0	1
004	B	125	0	1
005	C	140	0	0
...	...	...	...	...

6.6 Dummy coding for more than two groups

In the previous sections we saw how to code a categorical variable with 2 categories (a dichotomous variable) into 1 dummy variable. In this section, we learn how to code a categorical variable with 3 categories into 2 dummy variables, and to code a categorical variable with 4 categories into 3 dummy variables, etcetera. That is, how to code nominal variables into sets of dummy variables.

Take for instance a variable `Country`, where in your data set, there are three different values for this variable, for instance, Norway, Sweden and Finland, or Zimbabwe, Congo and South-Africa. Let's call these countries A, B and C. Table 6.7 shows a data example where we see height measurements on people from three different countries.

We can code this `Country` variable with three categories into two dummy variables in the following way. First, we create a variable `countryA`. This is a dummy variable, or indicator variable, that indicates whether a person comes from country A or not. Those persons that do are coded 1, and those that do not are coded 0. Next, we create a dummy variable `countryB` that indicates whether or not a person comes from country B. Again, persons that do are coded 1, and those that do not are coded 0. The resulting variables are displayed in Table 6.8

Note that we have now for every value of `Country` (A, B, or C) a unique combination of the variables `countryA` and `countryB`. All persons from country A have a 1 for `countryA` and a 0 for `countryB`; all those from country B have a 0 for `countryA` and a 1 for `countryB`, and all those from country C have a 0 for `countryA` and a 0 for `countryB`. Therefore a third dummy variable `countryC`

is not necessary (i.e., is redundant): the two dummy variables give us all the country information we need.

Remember that with two categories, you only need one dummy variable, where one category gets 1s and another category gets 0s. In this way both categories are uniquely identified. Here with three categories we also have unique codes for every category. Similarly, if you have 4 categories, you can code this with 3 dummy variables. In general, when you have a variable with K categories, you can code them with $K - 1$ dummy variables.

6.7 Analysing categorical predictor variables in R

R contains a data set (`PlantGrowth`) on yield on a sample of thirty observations ($n = 30$), under three different conditions. We already saw part of the data in Figure 6.4. The complete data consists of weight in three different groups: treatment 1 (trt1), treatment 2 (trt2) and a control group (ctrl). Now we want to model `weight` as a function of the categorical variable `group` using a linear model. Again, we discuss two options how to do that in R. First by creating your own dummy variables, second by letting R create dummy variables automatically.

6.7.1 Creating your own dummy variables

Suppose you want to compare treatments 1 and 2 to your control condition. Then it makes most sense to create two dummy variables, that code for the treatment 1 group and the treatment 2 group, respectively. Thus, we create a new variable `treatment_1` and code every specimen that belongs to treatment 1 as 1, and all other specimens as 0. Next, we create a new variable `treatment_2` and code every specimen that belongs to treatment 2 as 1, and all other specimens as 0. The code is as follows:

```
PlantGrowth <- PlantGrowth %>%
  mutate(treatment_1 = ifelse(group == "trt1", 1, 0),
         treatment_2 = ifelse(group == "trt2", 1, 0))
```

Next, we use these new dummy variables in a multiple regression analysis:

```
PlantGrowth %>%
  lm(weight ~ treatment_1 + treatment_2, data = .) %>%
  tidy()

## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    5.03      0.197     25.5  1.94e-20
```

```
## 2 treatment_1   -0.371     0.279     -1.33 1.94e- 1
## 3 treatment_2    0.494     0.279      1.77 8.77e- 2
```

You now have two numeric variables that you use in an ordinary multiple regression analysis. We see the effects (the 'slopes') of the two dummy variables. Based on these slopes and the intercept, we can construct the linear equation for the relationship between treatment and weight (yield):

$$\widehat{\texttt{weight}} = 5.03 - 0.37 \times \texttt{treatment\_1} + 0.49 \times \texttt{treatment\_2} \tag{6.13}$$

Based on this we can make predictions for the mean height in the control group, the treatment 1 group and the treatment 2 group.

Control group specimens score 0 on variable `treatment_1` and 0 on variable `treatment_2`. Therefore, their predicted weight equals:

$$5.03 - 0.37 \times 0 + 0.49 \times 0 = 5.03$$

.

Thus, the expected weight in the control group is equal to the intercept, as we used the control group as the reference group.

Specimens in the treatment 1 group score 1 on the `treatment_1` variable but 0 on the `treatment_2` variable. Therefore, their predicted weight equals:

$$5.03 - 0.37 \times 1 + 0.49 \times 0 = 5.03 - 0.37 = 4.66$$

.

Specimens in the treatment 2 group score 0 on the `treatment_1` variable but 1 on the `treatment_2` variable. Therefore, their predicted weight equals:

$$5.03 - 0.37 \times 0 + 0.49 \times 1 = 5.03 + 0.49 = 5.52$$

.

6.7.2 Let R create dummy variables automatically

The alternative is that R creates dummy variables automatically. The upside is that it is less work for you, the downside is that the first category (in terms of internally numbered category) always ends up as the reference category. The only thing that is required is that the independent variable in question is stored as a factor variable. Thus, in this case we want R to make dummy variables for our categorical variable `group`. But first, let's check whether the `group` variable is indeed a factor variable. If we look at the structure of it, we can see the types of variables.

```
PlantGrowth %>%
  str()
```

```
## 'data.frame': 30 obs. of  5 variables:
##  $ weight     : num  4.17 5.58 5.18 6.11 4.5 4.61 5.17 4.53 5.33 5.14 ...
##  $ group      : Factor w/ 3 levels "ctrl","trt1",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ grouptrt1  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ treatment_1: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ treatment_2: num  0 0 0 0 0 0 0 0 0 0 ...
```

Yes, the `group` variable is a factor (`Factor`). If we use a factor in a linear model, R will automatically code this variable into a set of dummy variables.

```
##   lm(weight ~ group, data = .)
out %>%
  tidy()

## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    5.53      0.203     27.2  4.52e-16
## 2 grouptrt1     -0.865     0.287     -3.01 7.52e- 3
```

The regression table now looks slightly different: all values are the same as in the analysis where we created our own dummies, except that R has used different names for the dummy variables it created. The parameter values are exactly the same as in the previous analysis, because in both analyses the control condition is the reference category. This control condition is chosen by R as the reference category, because alphabetically, `ctrl` comes before `trt1` and `trt2`.

Thus, the new variable `grouptrt1` codes 1s for all observations where `group = trt1` and 0s otherwise, and the new variable `grouptrt2` codes 1s for all observations where `group = trt2` and 0s otherwise.

6.7.3 Interpreting the regression table

The intercept is the expected mean for the reference group, that is, the group for which there is no dummy variable. Each slope is the difference between the group to which the slope belongs to and the reference group. For instance, in the yield example with manual coding of the dummy variables, the slope for the `treatment_1` dummy variable is the estimated population difference between the mean for treatment 1 minus the mean for control condition. Similarly, the slope for the `treatment_2` dummy variable is actually the estimated population difference between the mean for treatment 2 minus the mean for the control condition. The confidence intervals that belong to these parameter effects are also to be seen in this light: they are intervals for probable values for the *population* difference between the respective group means and the mean of the reference group. The t-values and p-values are related to null-hypothesis tests regarding these differences to be 0 in the population.

As an example, suppose that we want to estimate the difference in mean weight between plants from the treatment 1 group relative to the control group. From the output, we see that our best guess for this difference (the least square estimate) equals -0.37, where the yield is less with treatment 1 than under control conditions. The standard error for this difference equals 0.28. So a rough indication for the 95% confidence interval would be from $-0.37 - 2 \times 0.28$ to $-0.37 + 2 \times 0.28$, that is, from -0.93 to 0.19. Therefore, we infer that in the population, our best guess for the difference is somewhere between -0.93 and 0.19.

If we would want to, we could perform three null-hypothesis tests based on this output: 1) whether the population intercept equals 0, that is, whether the population mean of the control group equals 0; 2) whether the slope of the treatment 1 dummy variable equals 0, that is, whether the difference between the population means of treatment 1 group and the control group is 0; and 3) whether the slope of the treatment 2 group dummy variable equals 0, that is, whether the difference between the population means of the treatment 2 group and the control group is 0.

Obviously, the first hypothesis is not very interesting: we're not interested to know whether the average weight in the control group equals 0. But the other two null-hypotheses could be interesting in some scenarios. What is missing from the table is a test for the null-hypothesis that the means of the two treatments conditions are equal. This could be solved by manually creating two other dummy variables, where either treatment 1 or 2 is the reference group, or by looking at tricks in Chapter **??**. But what is also missing is a test for the null-hypothesis that all three population means are equal. In order to do that, we first need to explain *analysis of variance*.

6.8 Analysis of Variance

Since we know that applying a linear model to a categorical independent variable is the same as modelling group means, we can test the null-hypothesis that all group means are equal in the population. Let μ_{t1} be the mean yield in the population of the treatment 1 group, μ_{t2} be the mean yield in the treatment 2 group, and μ_c be the mean yield in the control group. Then we can specify the null-hypothesis using symbols in the following way:

$$H_0 : \mu_{t1} = \mu_{t2} = \mu_c \tag{6.14}$$

If all group means are equal in the population, then all population slopes would be 0. We want to test this null-hypothesis with a linear model in R. We then have only one independent variable, `group`, and if we let R do the dummy coding for us, R can give us an Analysis of Variance. We do that in the following way:

```
out <- PlantGrowth %>%
  lm(weight ~ group, data = .)
out %>%
  anova() %>%
  tidy()

## # A tibble: 2 x 6
##   term         df sumsq meansq statistic p.value
##   <chr>     <int> <dbl>  <dbl>     <dbl>   <dbl>
## 1 group         2  3.77  1.88       4.85  0.0159
## 2 Residuals    27 10.5   0.389     NA    NA
```

We don't see a regression table, but output based on a so-called Analysis Of VAriance, or ANOVA for short. This table is usually called an ANOVA table. The function `anova()` is used after fitting a linear model with `lm()`. ANOVA is in fact an alternative way of presenting the results of a linear model.

In the output, you see a column `statistic`, with the value 4.85 for the `group` variable. It looks similar to the column with the t-statistic in a regression table, but it isn't. The statistic is an F-statistic.

The F-statistic is constructed on the basis of Sum of Squares (SS, `sumsq` in the R table). Sums of squares we already encountered in Chapter 1, where they form the basis of variances and standard deviations. We also saw sums of squares in Chapter 4 where the sum of squared residuals (SSR) was minimised to get the least squares estimator of regression coefficients. In Chapter 4 we also saw that the sum of squared residuals (SSR) and the total sum of squares (SST) were used to compute the R-squared and the adjusted R-squared. Actually, the sum of squares that we see in the ANOVA table here in the row named `Residuals` is exactly the SSR: the sum of the squared residuals. Here we see that the sum of the squared residuals equals 10.5.

In the ANOVA table, we also see degrees of freedom (`df`). The degrees of freedom in the row named `Residuals` are the residual degrees of freedom that we already use when doing linear regression (Chapter 5). Here we see the residual degrees of freedom equals 27. This is so because we have 30 data points, and for a linear model the number of degrees of freedom is $n - K - 1 = 30 - 2 - 1 = 27$, with K being the number of independent variables (two dummy variables).

Continuing, we see Mean Squares (`meansq`). These numbers are nothing but the sum of squares (`sumsq`) divided by the respective degrees of freedom (`df`). For instance, in the row for Residuals, the Sum of Squares equals 10.5, the degrees of freedom equals 27, and the Mean square equals $\frac{10.5}{27} = 0.389$.

Then there is a column with F-values (`statistic`). F-values are test statistics and are used in the same way as t-statistics. Under the null-hypothesis they have a known distribution, and if they have very extreme values, you can reject the null-hypothesis. Whether or not the null-hypothesis can be rejected, depends on your pre-set α level and whether the p-value, reported in the last column (`p.value`) is equal or smaller than your α level.

The F-value is computed based on the Mean squares values (`meansq`). Let's

look at the F-value for the `group` variable. The F-value equals 4.85. This is the ratio of the mean square of the `group` effect, which is 1.88, and the mean square of the residuals (error), which is 0.389. Thus, the F-value for country is computed as $\frac{1.88}{0.389} = 4.85$. Under the null-hypothesis that all three population means are equal, this ratio is around 1. Why this is so, we will explain later. Here we see that the F-value based on these sample data is larger than 1. But is it large enough to reject the null-hypothesis? That depends on the degrees of freedom. The F-value is based on two mean squares, and these in turn are based on two separate numbers of degrees of freedom. The one for the effect of country was 2 (3 countries so 2 degrees of freedom), and the one for the residual mean square was 27 (27 residual degrees of freedom). We therefore have to look up in a table whether an F-value of 4.85 is significant at 2 and 27 degrees of freedom for a specific α. Such a table is displayed in Table 6.9. It shows critical values if your α is 0.05. In the columns we look up our *model degrees of freedom*. Model degrees of freedom is computed based on the number of independent variables. Here we have a categorical variable `group`. But because this categorical variable is represented in the analysis as two dummy variables, the number of variables is actually 2. The model degrees of freedom is therefore 2.

In the rows of Table 6.9 we look up our residual degrees of freedom: 27. For 2 and 27 degrees of freedom we find a critical F-value of 3.35. It means that if we have an α of 0.05, an F-value of 3.35 or larger is large enough to reject the null-hypothesis. Here we found an F-value of 4.85, so we reject the null-hypothesis that the three population means are equal. Therefore, the mean weight is not the same in the three experimental conditions.

Table 6.9: Critical values for the F-value if $\alpha = 0.05$, for different model degrees of freedom (columns) and error degrees of freedom (rows).

	1	2	3	4	5	10	25	50
5	6.61	5.79	5.41	5.19	5.05	4.74	4.52	4.44
6	5.99	5.14	4.76	4.53	4.39	4.06	3.83	3.75
10	4.96	4.10	3.71	3.48	3.33	2.98	2.73	2.64
27	4.21	3.35	2.96	2.73	2.57	2.20	1.92	1.81
50	4.03	3.18	2.79	2.56	2.40	2.03	1.73	1.60
100	3.94	3.09	2.70	2.46	2.31	1.93	1.62	1.48

Note that our null-hypothesis that all group means are equal in the population cannot be answered based on a regression table. If the population means are all equal, then the slope parameters should consequently be 0 in the population. Let's have a look again at the regression table, plotting also the 95% confidence intervals:

```
out <- PlantGrowth %>%
  lm(weight ~ group, data = .)
out %>%
  tidy(conf.int = 0.95)
```

```
## # A tibble: 3 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)    5.03      0.197     25.5  1.94e-20   4.63       5.44
## 2 grouptrt1     -0.371     0.279     -1.33 1.94e- 1  -0.943      0.201
## 3 grouptrt2      0.494     0.279      1.77 8.77e- 2  -0.0780     1.07
```

Looking at the 95% confidence intervals for `grouptrt1` and `grouptrt2`, we see that 0 is a reasonable value for the difference between the control group (the reference category) and treatment 1, and that 0 is also a reasonable value for the difference between the control group and treatment 2. But what does that tell us about the difference between the two treatment groups? How can we rigorously test the null-hypothesis, with one clear statistic, that all three group means are the same? In the regression table we have two t-statistics and two p-values, one for the difference between control and treatment 1 ($t = -1.33, p = 0.194$) and one of the difference between treatment 2 and the control group ($t = 1.77, p = 0.0.0877$), but we actually need one p-value for the null-hypothesis of three equal means.

The ANOVA table looks slightly different: instead of two separate effects for two dummy variables, we only see one row for the original categorical variable `group`. And in the column `df` (degrees of freedom): instead of 1 degree of freedom for a specific dichotomous dummy variable, we see 2 degrees of freedom for the nominal `group` variable. So this suggests that *the effects of the two dummy variables are now combined into one effect*, with one particular F-statistic, and one p-value that is also different from those of the two separate dummy variables. This is actually the p-value associated with the test of the null-hypothesis that all 3 means are equal:

$$H_0 : \mu_{t1} = \mu_{t2} = \mu_c \tag{6.15}$$

This hypothesis test is very different from the t-tests in the regression table. The t-test for the `grouptrt1` effect specifically tests whether the average weight in treatment 1 group is different from the average weight in the control group (the reference country). The t-test for the `grouptrt2` effect specifically tests whether the average weight in the treatment 2 group is different from the average weight in the control group (the reference country). Since these two hypotheses do not refer to our original research question regarding *overall* differences across all three groups, we do not report these t-tests, but we report the overall F-test from the ANOVA table.

In general, the rule is that if you have a specific research question that addresses a particular null-hypothesis, you only report the statistical results regarding that null-hypothesis. All other p-values that your software happens to show in its output should be ignored. We will come back to this issue in Chapter **??**.

6.9 The logic of the F-statistic

As stated earlier, the ANOVA is an alternative way of representing the linear model. Suppose we have a dependent variable Y, and three groups: A, B and C. In the usual linear model, we have an intercept b_0, and we use two dummy variables. Suppose we use C as our reference group, then we need two dummy variables for groups A and B. We could model the data then using the following equation, with normally distributed errors:

$$\begin{aligned} Y &= b_0 + b_1\texttt{dummy}_A + b_2\texttt{dummy}_B + e && (6.16)\\ e &\sim N(0, \sigma^2) && (6.17) \end{aligned}$$

This is the linear model as we know it. The linear equation has three unknown parameters that need to be estimated: one intercept and two dummy effects. The dummy effects are the differences between the means of groups A and B relative to reference group C.

Alternatively, we could represent the same data as follows:

$$\begin{aligned} Y &= b_1\texttt{dummy}_A + b_2\texttt{dummy}_B + b_3\texttt{dummy}_C + e && (6.18)\\ e &\sim N(0, \sigma^2) && (6.19) \end{aligned}$$

That is, instead of estimating one intercept and two dummy effects, we simply estimate the three population means directly! We leave out the intercept, and we estimate three population means.

Next, we focus on the variance of the dependent variable, Y in this case, that is split up into two parts: one part that is explained by the independent variable (groups in this case) and one part that is not explained (cf. Chapters 4). The unexplained part is easiest of course: that is simply the part shown by the residuals, hence σ^2.

The logic of the F-statistic is entirely based on this σ^2. As stated earlier, under the null-hypothesis the F-statistic should have a value of around 1. This is because F is a ratio and under the null-hypothesis, the numerator and the denominator of this ratio should be more or less equal. This is so because both the numerator and the denominator are estimators of σ^2. Under the null-hypothesis, these estimators should result in more or less the same numbers, and then the ratio is more or less 1. If the null-hypothesis is *not* true, then the numerator becomes larger than the denominator and hence the F-value becomes larger than 1.

In the previous section we saw that the numerator of the F-statistic was computed by taking the sum of squares of the group variable and dividing it by the degrees of freedom. What is actually being done is the following: If the null-hypothesis is really true, then the three population means are equal, and you simply have three independent samples from the *same* population. Each sample mean shows simply a slight deviation from the population mean.

This variance of sample means should remind us of something. If we go back to Chapter 2, we saw there that if we have a population with mean μ and variance σ^2, and if we draw many many random samples of size n and compute sample means for each sample, their distribution is a sampling distribution (Fig. 2.2). We also saw in Chapter 2 that on average the sampling distribution will show a mean that is the same as the population mean: the sample mean is an unbiased estimator of the population mean. And, important for ANOVA, the standard deviation of the sampling distribution, known as the standard error, will be equal to $\sigma_{\overline{Y}} = \sqrt{\frac{s^2}{n}}$ (Chapter 2). If we take the square, we see that the variance of the sample means is equal to

$$\sigma^2_{\overline{Y}} = \frac{s^2}{n} \tag{6.20}$$

where s^2 is the unbiased estimator of the variance of Y.

In the ANOVA model above, we have three group means. Now, suppose we have an alternative model, under the null-hypothesis, that there is really only one population mean μ, and that the observed different group means in groups A, B and C are only the result of chance (random sampling). Then the variance of the group means is nothing but the square of the standard error, and the number of observations per group is the sample size n. If that is the case, then we can flip the equation of the standard error around and say:

$$\widehat{\sigma^2} = \sigma^2_{\overline{Y}} \times n = \frac{SS}{2} \times n \tag{6.21}$$

or in words: our estimate of the total variance of Y in the population is the estimated variance of the group means in the population times the number of observations per group.

So the numerator is one estimator of the variance of the residuals. For that estimator we only used information about the group means: we looked at variation between groups. Now let's look at the denominator. For that estimator we use information from the raw data and how they deviate from the sample group means, that is we look at within-group variation. Similar to regression analysis, for each observed value, we compute the difference between the observed value and the group mean. We then compute the sum of squared residuals, SSR. If we want the variance, we need to divide this by sample size, n. However, if we want to estimate the variance in the population, we need to divide by a corrected n. In Chapter 2 we saw that if we wanted to estimate a variance in the population on the basis of one sample with one sample mean, we used $s^2 = \frac{SS}{n-1}$. The $n-1$ was in fact due to the loss of 1 degree of freedom because by computing the sample variance, we used the sample mean, which was only an *estimate* of the population mean. Here, because we have three groups, we need to estimate three population means, and the degrees of freedom is therefore $n-3$. The estimated variance in the population that is *not* explained by the independent variable is therefore $SSR/(n-3)$.

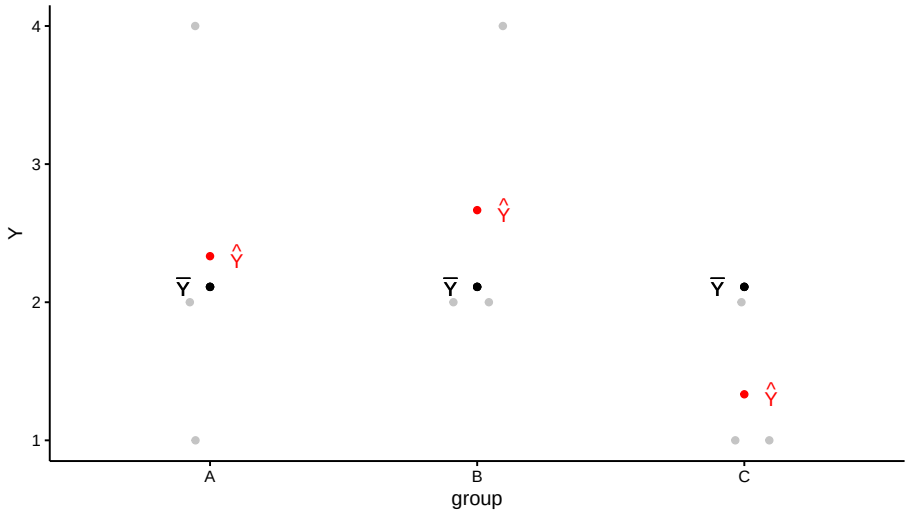


Figure 6.6: Illustration of ANOVA using a very small data set. In grey the raw data, in black the overall sample mean, and in red the sample group means.

Thus, if we want to estimate σ^2 related to the model, we can either do that by looking at the model residuals, computing the sum of squared residuals and dividing it by the degrees of freedom (in this case $SSR/(n-3)$), but we can also do it by looking at the variation of the group means and multiplying it by the group size. Now, the method of looking at the residuals will generally yield a good estimate of σ^2, whether the null-hypothesis is true or not. However, only if the null-hypothesis is true, the method of looking at the variation of group means will yield a good estimate of σ^2. Only if the null-hypothesis is true, both estimates will be more or less the same. But if the null-hypothesis is *not* true, if the means are really different in the population, then the method of looking at the variation of group means will yield an estimate of σ^2 that is larger than an estimate based on residuals. Then, if you compute a ratio, the ratio will become larger than 1. Therefore, an F-statistic larger than 1 is evidence that the population means might not be equal. How much larger than 1 an F-value should be to regard it as evidence against H_0 depends on your level of significance α and the degrees of freedom.

6.10 Small ANOVA example

To illustrate the idea of ANOVA and the computation of the F-statistic, let's assume we have a very small data set, involving height data from three countries A, B and C. From each country, we only have 3 data points. The data are plotted in Figure 6.6. In grey, we see the raw data values for variable Y. In group A, we see the values 4, 2 and 1; in group B, we see the values 4, 2 and

2, and in group C, we see the values 2, 1 and 1. When we sum all these values and divide by 9, we get the overall mean (the grand mean), which is equal to $\overline{Y} = 2.1111111$, denoted in black in Figure 6.6. In red, we see the sample group means. For group A, that is equal to $(4+2+1)/3 = 2.3333333$, for group B this is $(4+2+2)/3 = 2.6666667$, and for group C this equals $(2+1+1)/3 = 1.3333333$.

Thus, our ANOVA model for these data is the following:

$$\begin{aligned} Y &= b_1 \texttt{dummy}_A + b_2 \texttt{dummy}_B + b_3 \texttt{dummy}_C + e \\ e &\sim N(0, \sigma^2) \end{aligned} \tag{6.22}$$

Our OLS estimates for the parameters are the sample means, so that we have the linear equation

$$\widehat{Y} = 2.3333333 \texttt{dummy}_A + 2.6666667 \texttt{dummy}_B + 1.3333333 \texttt{dummy}_C \tag{6.23}$$

Based on this linear equation we can determine the predicted values for each data point. Table 6.10 shows the Y-values, the `group` variable, the dummy variables from the ANOVA model equation (Equation 6.22) and the predicted values. We see that the predicted value for each observed value is equal to the sample group mean.

Table 6.10: Small data example for illustrating ANOVA and the F-statistic.

Y	group	dummy_A	dummy_B	dummy_C	predicted	residual
1	A	1	0	0	2.33	-1.33
2	A	1	0	0	2.33	-0.33
4	A	1	0	0	2.33	1.67
2	B	0	1	0	2.67	-0.67
2	B	0	1	0	2.67	-0.67
4	B	0	1	0	2.67	1.33
2	C	0	0	1	1.33	0.67
1	C	0	0	1	1.33	-0.33
1	C	0	0	1	1.33	-0.33

Using these predicted values, we can compute the residuals, also displayed in Table 6.10, and these help us to compute the first estimate of σ^2, the one based on residuals, namely the SSR divided by the degrees of freedom. If we square the residuals in Table 6.10 and sum them, we obtain $SSR = 8$. To obtain the Mean Squared Error (MSE or `meansq` for Residuals), we divide the SSR by the degrees of freedom. Because the linear model with 2 dummy variables has $n - K - 1 = 9 - 2 - 1 = 6$ residuals degrees of freedom (see Chapter 5), we also have only 6 residual degrees of freedom. Thus we get $MSE = 8/6 = 1.3333333$. We can see these numbers in the bottom row in the ANOVA table, displayed in Table 6.11.

For our second estimate of σ^2, the one based on the group means, we look at the squared deviations of the group means from the overall mean (the grand mean). We saw that the grand mean equals 2.11. The sample mean for group A was 2.3333333, so the squared deviation equals 0.0493828. The sample mean for group B was 2.6666667, so the squared deviation equals 0.3086421. Lastly, the sample mean for group C was 1.3333333, so the squared deviation equals 0.3086421. Adding these squared deviations gives a sum of squares of 0.962963. To obtain an unbiased estimate for the population variance of these means, we have to divide this sum of squares by the number of groups minus 1 (model degrees of freedom), thus we get $0.962963/2 = 0.6604939$. This we must multiply by the sample size per group to obtain an estimate of σ^2 (see Equation 6.21), thus we obtain 1.4444444.

Table 6.11: ANOVA table for small data example.

term	df	sumsq	meansq	statistic	p.value
group	2	2.89	1.44	1.08	0.40
Residuals	6	8.00	1.33		

Table 6.12: Small data example for illustrating ANOVA and the F-statistic.

Y	group	predicted	grand_mean	deviation	sq_deviation
1	A	2.33	2.11	0.22	0.05
2	A	2.33	2.11	0.22	0.05
4	A	2.33	2.11	0.22	0.05
2	B	2.67	2.11	0.56	0.31
2	B	2.67	2.11	0.56	0.31
4	B	2.67	2.11	0.56	0.31
2	C	1.33	2.11	-0.78	0.60
1	C	1.33	2.11	-0.78	0.60
1	C	1.33	2.11	-0.78	0.60

Obtaining the estimate of σ^2 based on the group means can also be illustrated using Table 6.12. There again we see the raw data values for variable Y, the predicted values (the group means), but now also the grand mean, the deviations of the sample means from the grand mean, and their squared values. If we simply add the squared deviations, we no longer have to multiply by sample size. Thus we have as the sum of squares 2.8888889. Then we only have to divide by the number of groups minus 1, so we have $2.8888889/2 = 1.4444444$. This sum of squares, the degrees of freedom of 2, and the resulting MS can also be seen in the ANOVA table in Table 6.11.

Hence we have two estimates of σ^2, the one called the Mean Squared Error (MSE) that is based on the residuals (sometimes also called the MS within or MSW), and the other one called the Mean Squared Between groups (MSB), that is based on the sum of squares of group mean differences. For the F-statistic, we

use the MS Between (MSB) as the numerator and the MSE as the denominator,

$$F = \frac{MSB_{\texttt{group}}}{MSE} = \frac{1.4444444}{1.3333333} = 1.0833333 \tag{6.24}$$

We see that the F-statistic is larger than 1. That means that the estimate for σ^2, MSB_{group}, based on the sample means is larger than the estimate based on the residuals, MSE. This could indicate that the null-hypothesis, that the three population means are equal, is not true. However, is the F-value really large enough to justify such a conclusion?

To answer that question, we need to know what values the F-statistic would take for various data sets if the null-hypothesis were true (the sampling distribution of F). If for each data set we have three groups, each consisting of three observed values, then we have 2 degrees of freedom for the `group` effect, and 6 residual degrees of freedom. Table 6.9 shows critical values if we want to use an α of 0.05. If we look up the column with a 2 (for the number of model degrees of freedom) and the row with a 6 (for the residual degrees of freedom), we find a critical F-value of 5.14. This means that if the null-hypothesis is true and we repeatedly take random samples, we find an F-value equal to or larger than 5.14 only 5% of the time. If we want to reject the null-hypothesis, therefore, at an alpha of 5%, the F-value has to be equal or larger than 5.14. Here we found an F-value of only 1.0833333, which is much smaller, so *we cannot reject the null-hypothesis that the means are equal.*

For illustration, Figure 6.7 shows the distribution of the F-statistic with 2 and 6 degrees of freedom under the null-hypothesis. The figure shows it happens quite a lot under the null-hypothesis that the F-statistic is equal to 1.0833333 or larger.

6.11 Reporting ANOVA

In all cases where you have a categorical predictor variable with more than two categories, and where the null-hypothesis is about the equality of all group means, you have to use the factor variable in R associated with the original nominal variable. That is, don't make dummy variables yourself, but let R do it for you. You then always report the corresponding F-statistic from the ANOVA table.

For this particular example, you report the results of the analysis of variance in the following way:

> "The null-hypothesis that all 3 population means are equal was tested with an analysis of variance. The results showed that the null-hypothesis cannot be rejected, $F(2,\ 6) = 1.08, MSE = 1.33, p = 0.40$".

Always check the degrees of freedom for your F-statistic carefully. The first number refers to the degrees of freedom for the Mean Square Between: this is the number of groups minus 1 ($K - 1$). This is equal to the number of dummy

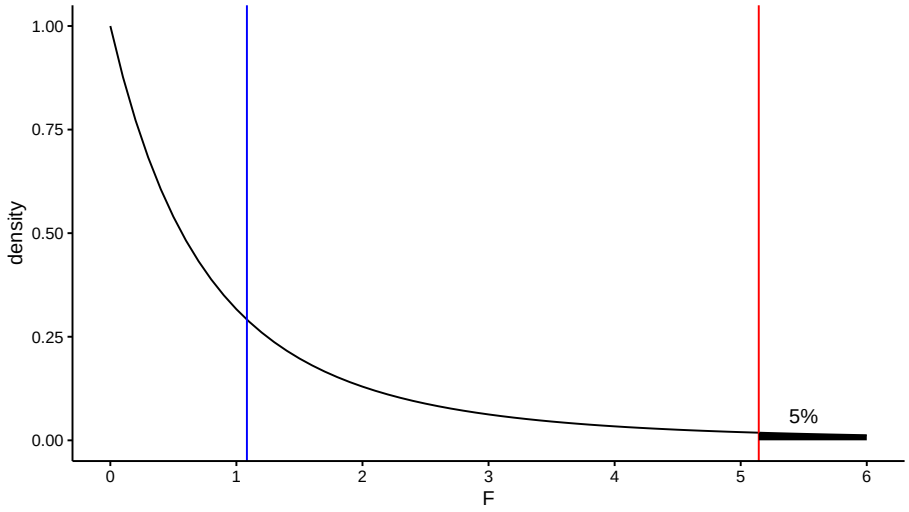


Figure 6.7: Density plot of the F-distribution with 2 and 6 degrees of freedom. In blue the observed F-statistic in the small data example, in red the critical value for an α of 0.05. The blackened area under the curve is 5%.

variables are used in the linear model. This is also called the *model degrees of freedom*. The second number refers to the residual degrees of freedom: this is $n - K - 1$ as we saw Chapter 5, where K is the number of dummy variables. In this ANOVA model you have 9 data points and you have 2 dummy variables for the three groups. So your residual degrees of freedom is $9 - 2 - 1 = 6$. This residual degrees of freedom is equal to that of the t-statistic for multiple regression.

6.12 Relationship between F- and t-distributions

The t-distribution and the F-distribution have much in common. Here we will illustrate this. Suppose that we test the null-hypothesis that a certain population slope is 0. We perform a regression analysis and obtain a t-statistic of -2.40. Suppose our sample size was 42, so that our residual degrees of freedom equals $42 - 2 = 40$. Figure 6.8 shows the theoretical t-distribution with 40 degrees of freedom. It also shows our value of -2.40. The shaded area represents the values for t that would be significant at an $\alpha = 0.05$.

Now look closely at Figure 6.8. The density says something about the probability of drawing certain values. Imagine that you randomly pick numbers from this t-distribution. The density plot tells you that values around zero are more probable than values around 2 or -2, and that values around 2 or -2 are more probable than values around 3 or -3. Imagine that you pick a million values for t, randomly from this t-distribution. Then imagine that you take the square of

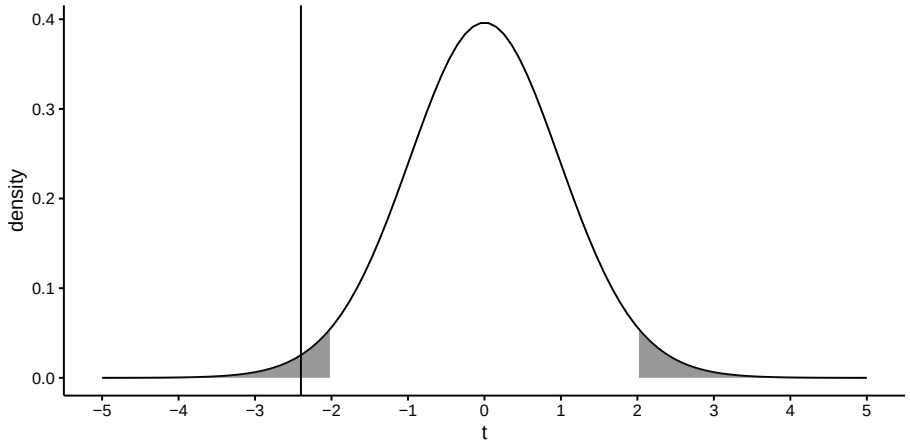


Figure 6.8: The vertical line represents a t-value of -2.40. The shaded area represents the extreme 5% of the possible t-values

each value (thus, suppose as the first 3 randomly drawn t-values you get -3.12, 0.14, and -1.6, you then square these numbers to get the numbers 9.73, 0.02, and 2.79). If you then make a density plot of these one million squared numbers, you get the density plot in Figure 6.9. It turns out that this density is an F-distribution with 1 model degrees of freedom and 40 residual degrees of freedom.

If we also square the observed test statistic t-value of -2.40, we obtain an F-value of 5.76. From online tables, we know that, with 1 model degrees of freedom and 40 residual degrees of freedom, the proportion of F-values larger than 5.76 equals 0.02. The proportion of t-values, with 40 (residual) degrees of freedom, larger than 2.40 or smaller than -2.40 is also 0.02. Thus, the two-sided p-value associated with a certain t-value, is equal to the p-value associated with an F-value that is the square of the t-value.

$$F(1, x) = t^2(x) \tag{6.25}$$

This means that if you see a t-statistic of say -2.40 reported with a residual degrees of freedom of 40, $t(40) = -2.40$, you can equally report this as an $F(1,\ 40) = (-2.40)^2 = 5.76$. Similarly, if you see a reported F-value of $F(1, 67) = 49$, you could without problems turn this into a $t(67) = 7$. Note however that this is only the case if the *model* degrees of freedom of the F-statistic is equal to 1. This means you cannot do this if you are comparing more than two groups means.

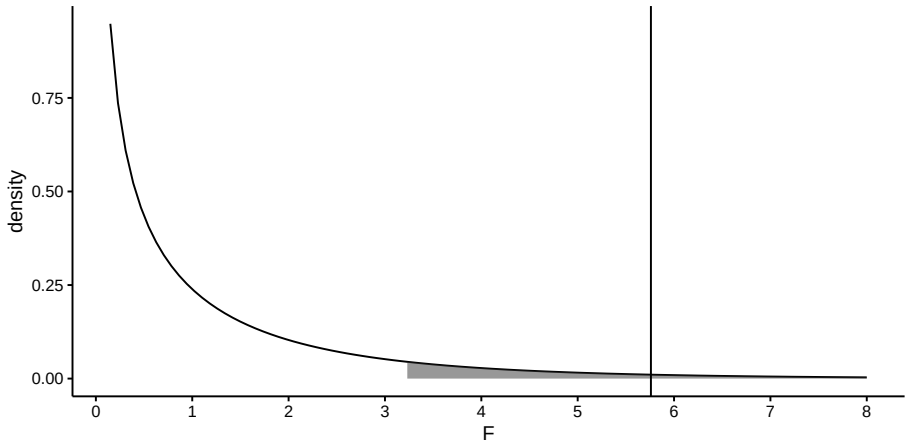


Figure 6.9: The F-distribution with 1 model degrees of freedom and 40 error degrees of freedom. The shaded area is the upper 5% of the distribution. The vertical line represents the square of -2.40: 5.76

Chapter 7

Assumptions of linear models

7.1 Introduction

Linear models are models. A model describes the relationship between two or more variables. A good model gives a valid summary of what the relationship between the variables looks like. Let's look at a very simple example of two variables: height and weight. In a sample of 100 children from a distant country, we find 100 combinations of height in centimetres and weight in kilograms that are depicted in the scatter plot in Figure 7.1.

We'd like to find a linear model for these data, so we determine the least squares regression line. We also determine the standard deviation of the residuals so that we have the following statistical model:

$$\texttt{weight} = -104.83 + 1.04 \times \texttt{height} + e \tag{7.1}$$

$$e \sim N(0, \sigma = 4.04) \tag{7.2}$$

This model, defined above, is depicted in Figure 7.2. The blue line is the regression line, and the dots are the result of simulating (inventing) independent normal residuals with standard deviation 4.04. The figure shows how the data would like according to the model.

The actual data, displayed in Figure 7.1 might have arisen from this model in Figure 7.2. The data is only different from the simulated data because of the randomness of the residuals.

A model should be a good model for two reasons. First, a good model is a summary of the data. Instead of describing all 100 data points on the children, we could summarise these data with the linear equation of the regression line and the standard deviation (or variance) of the residuals. The second reason is that you would like to *infer* something about the relationship between height and

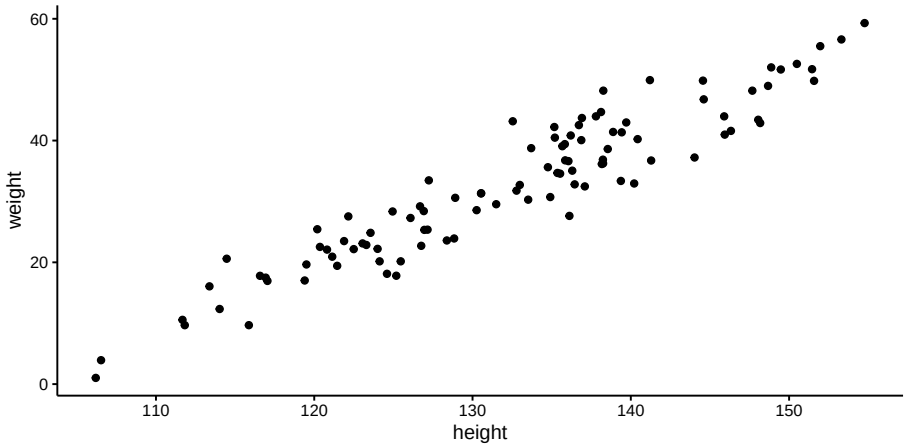


Figure 7.1: Data set on height and weight in 100 children.

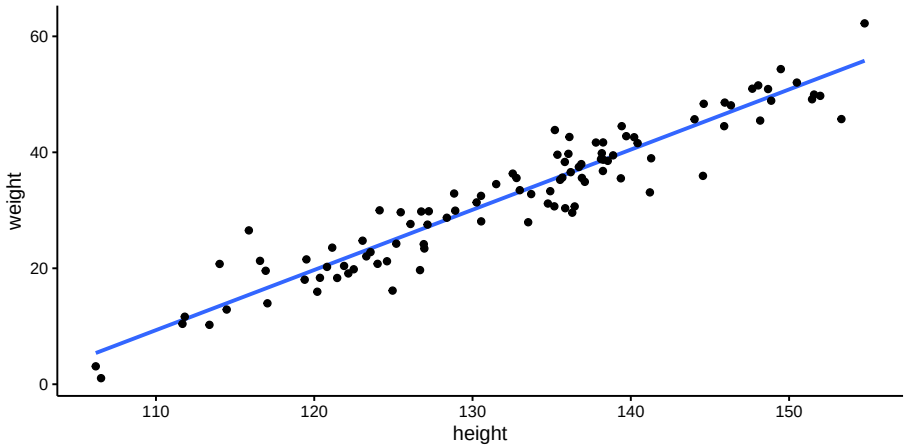


Figure 7.2: Data set on height and weight in 100 children and the least squares regression line.

weight in all children from that distant country. It turns out that the standard error, and hence the confidence intervals and hypothesis testing, are only valid if the model describes the data well. This means that if the model is not a good description of your sample data, then you draw the wrong conclusions about the population.

For a linear model to be a good model, there are four conditions that need to be fulfilled.

1. **linearity** The relationship between the variables can be described by a linear equation (also called additivity)
2. **independence** The residuals are independent of each other
3. **equal variance** The residuals have equal variance (also called homoskedasticity)
4. **normality** The distribution of the residuals is normal

If these conditions (often called assumptions) are not met, the inference with the computed standard error is invalid. That is, if the assumptions are not met, the standard error should not be trusted, or should be computed using alternative methods.

Below we will discuss these four assumptions briefly. For each assumption, we will show that the assumption can be checked by looking at the residuals. We will see that if the residuals do not look right, one or more of the assumptions are violated. But what does it mean that the residuals 'look right'?

Well, the linear model says that the residuals have a *normal distribution*. So for the height and weight data, let's apply regression, compute the residuals for all 100 children, and plot their distribution with a histogram, see Figure 7.3. The histogram shows a bell-shaped distribution with one peak that is more or less symmetric. The symmetry is not perfect, but you can well imagine that if we had measured more children, the distribution could more and more resemble a normal distribution.

Another thing the model implies is that the residuals are *random*: they are random draws from a normal distribution. This means, if we would plot the residuals, we should see no systematic pattern in the residuals. The scatter plot in Figure 7.4 plots the residuals in the order in which they appear in the data set. The figure seems to suggest a random scatter of dots, *without any kind of system or logic*. We could also plot the residuals as a function of the predicted height (the dependent variable). This is the most usual way to check for any systemetic pattern. Figure 7.5 shows there is no systematic relationship between the predicted height of a child and the residual.

When it looks like this, it shows that the residuals are randomly scattered around the regression line (the predicted heights). Taken together, Figures 7.3, 7.4 and 7.5 suggest that the assumptions of the linear model are met.

Let's have a look at the same kinds of residual plots when each of the assumptions of the linear model are violated.

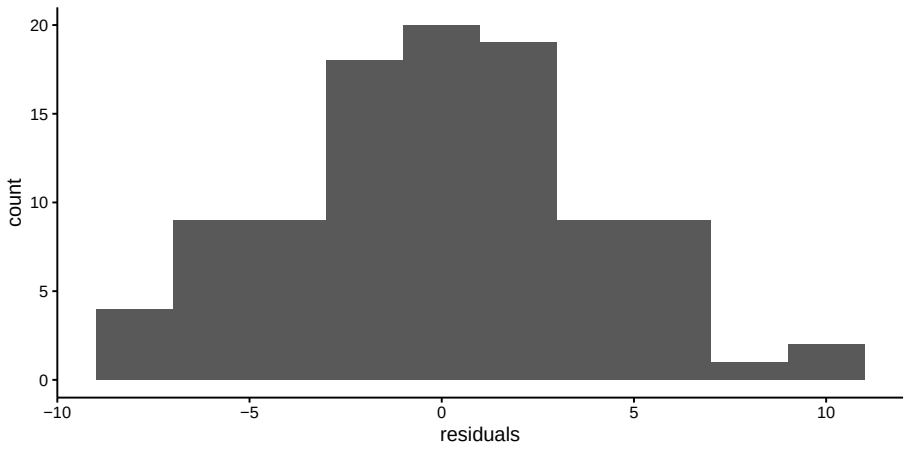


Figure 7.3: Histogram of the residuals after regressing weight on height.

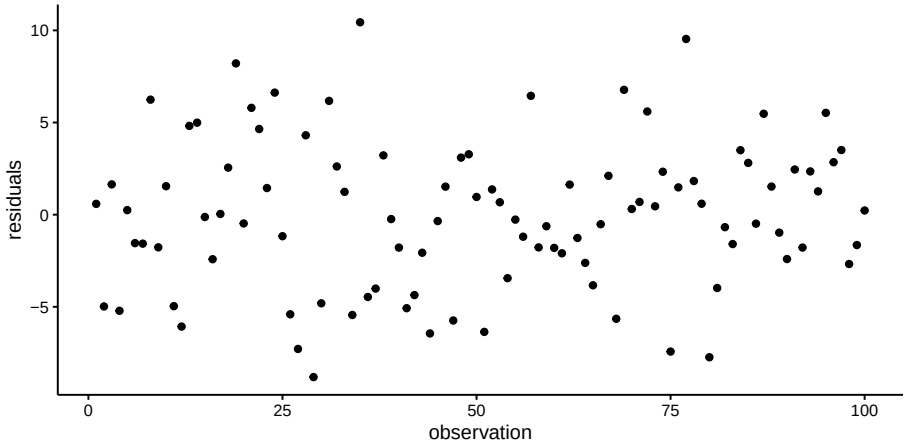


Figure 7.4: Residual plot after regressing weight on height.

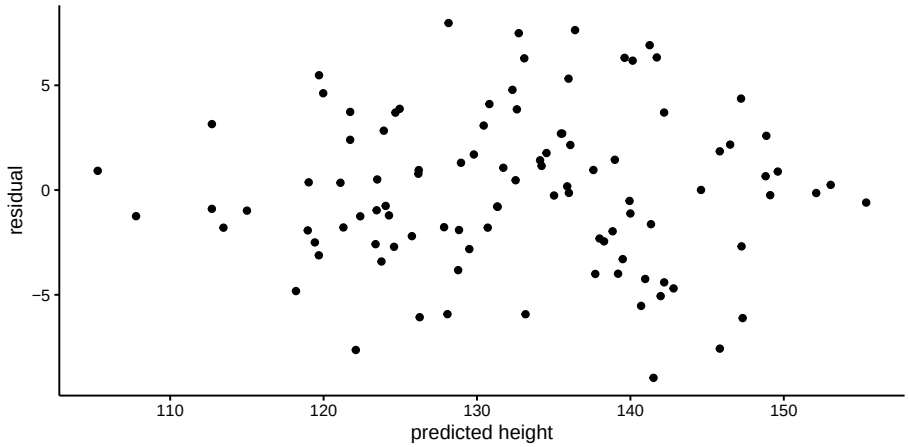


Figure 7.5: Residuals as a function of height.

7.2 Independence

The assumption of independence is about the way in which observations are similar and dissimilar *from each other*. Take for instance the following regression equation for children's height predicted by their age:

$$\texttt{height} = 100 + 5 \times \texttt{age} + e \tag{7.3}$$

This regression equation predicts that a child of age 5 has a height of 125 and a child of age 10 has a height of 150. In fact, all children of age 5 have the same predicted height of 125 and all children of age 10 have the same predicted height of 150. Of course, in reality, children of the same age will have very different heights: they differ. According to the above regression equation, children are similar in height because they have the same age, but they differ because of the random term e that has a normal distribution: predictor `age` makes them similar, residual e makes them dissimilar. Now, if this is all there is, then this is a good model. But let's suppose that we're studying height in an international group of 50 Ethiopian children and 50 Vietnamese children. Their heights are plotted in Figure 7.6.

From this graph, we see that heights are similar because of age: older children are taller than younger children. But we see that children are also similar because of their national background: Ethiopian children are systematically taller than Vietnamese children, irrespective of age. So here we see that a simple regression of height on age is not a good model. We see that, when we estimate the simple regression on age and look at the residuals in Figure 7.7.

As our model predicts random residuals, we expect a random scatter of residuals. However, what we see here is a systematic order in the residuals:

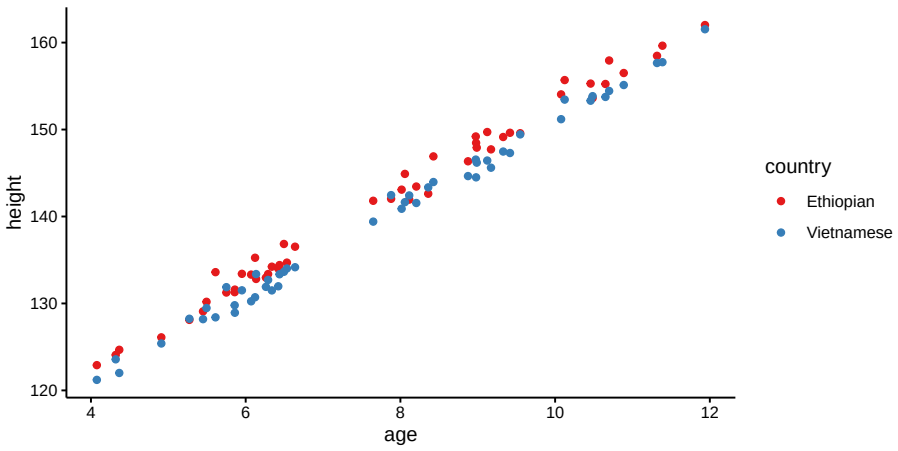


Figure 7.6: Data on age and height in children from two countries.

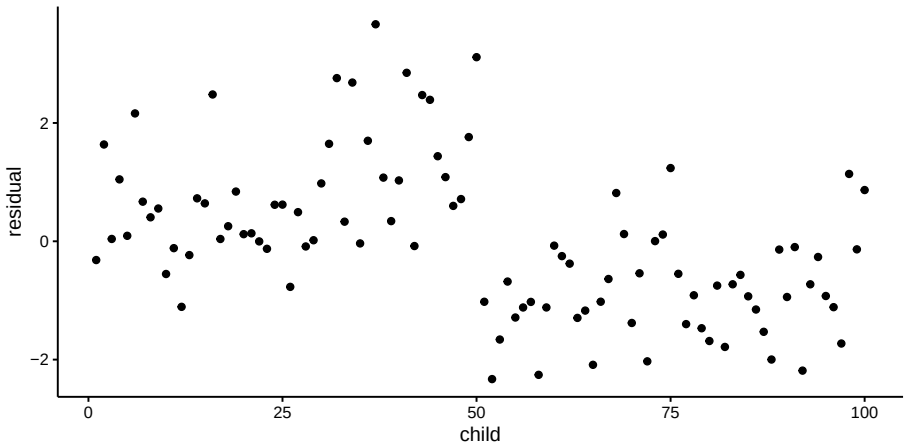


Figure 7.7: Residual plot after regressing height on age.

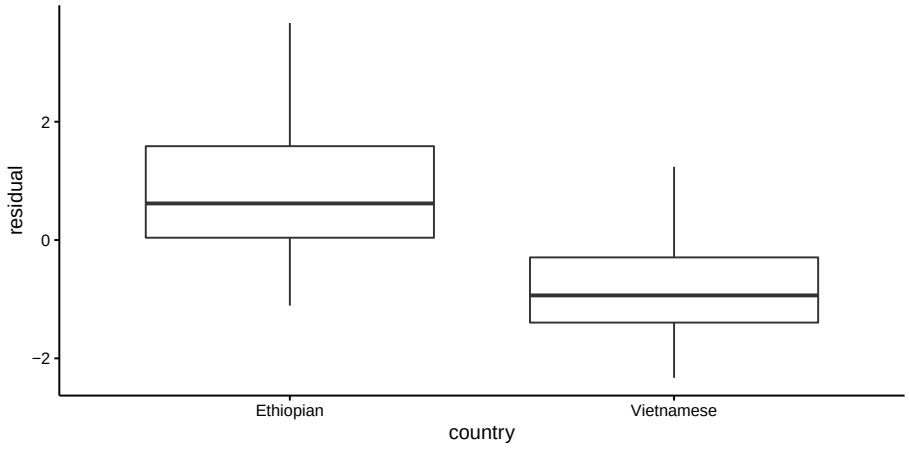


Figure 7.8: Residual plot after regressing height on age.

they tend to be positive for the first 50 children and negative for the last 50 children. These turn out to be the Ethiopian and the Vietnamese children, respectively. This systematic order in the residuals is a violation of independence: the residuals should be random, and they are not. The residuals are dependent on country: positive for Ethiopians, negative for Vietnamese children. We see that clearly when we plot the residuals as a function of country, in Figure 7.8.

Thus, there is more than just age that makes children similar. That means that the model is not a good model: if there is more than just age that makes children more alike, then that should be incorporated into our model. If we use multiple regression, including both age and country, and we do the analysis, then we get the following regression equation:

$$\widehat{\texttt{height}} = 102.641 + 5.017 \times \texttt{age} - 1.712 \times \texttt{countryViet} \tag{7.4}$$

When we now plot the residuals we see that there is no longer a clear country difference, see Figure 7.9.

Another typical example of non-random scatter of residuals is shown in Figure 7.10. They come from an analysis of reaction times, done on 10 students where we also measured their IQ. Each student was measured on 10 trials. We predicted reaction time on the basis of student's IQ using a simple regression analysis. The residuals are clearly not random, and if we look more closely, we see some clustering if we give different colours for the data from the different students, see Figure 7.11.

We see the same information if we draw a boxplot, see Figure 7.12. We see that residuals that are close together come from the same student. So, reaction time are not only similar because of IQ, but also because they come from the same student: clearly something other than IQ also explains why

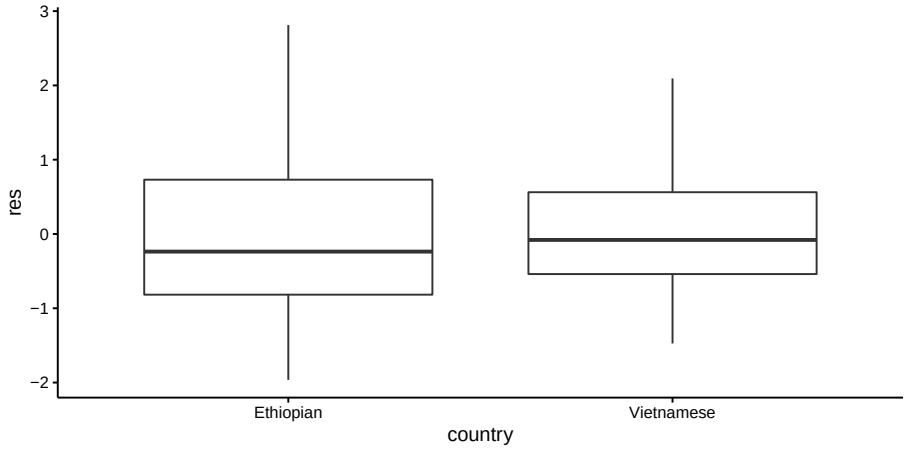


Figure 7.9: Residual plot after regressing height on age and country.

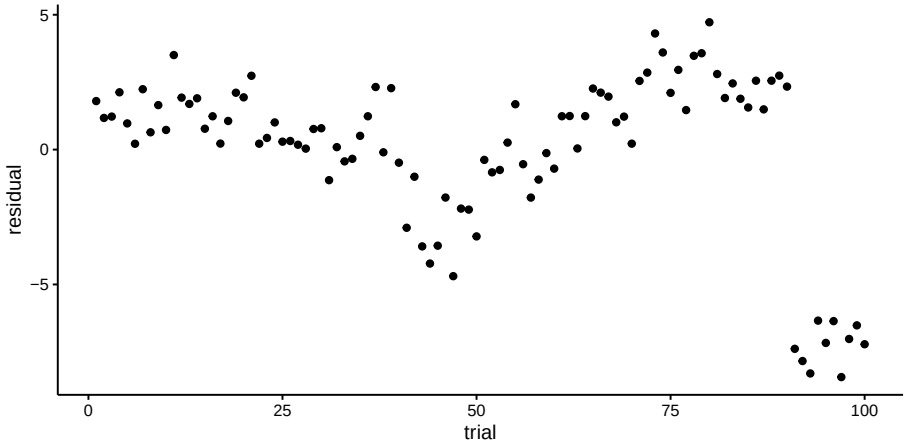


Figure 7.10: Residual plot after regressing reaction time on IQ.

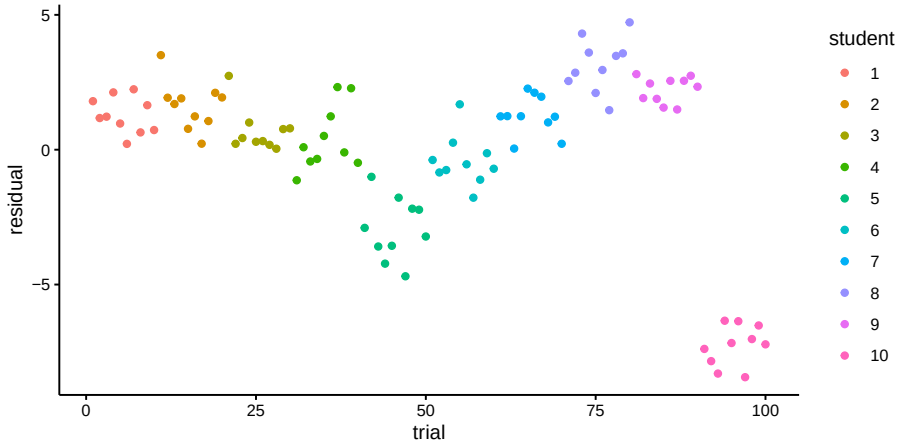


Figure 7.11: Residual plot after regressing reaction time on IQ, with separate colours for each student.

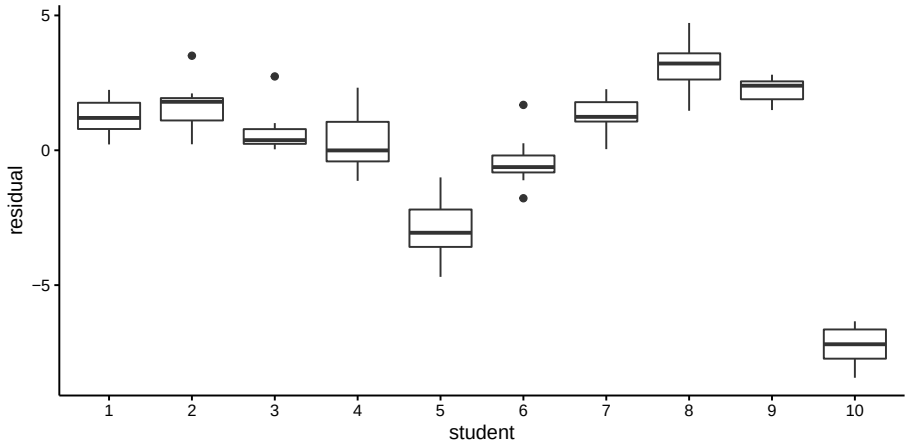


Figure 7.12: Box plot after regressing reaction time on IQ.

reaction times are different across individuals. The residuals in this analysis based on IQ are not independent: they are dependent on the student. This may be because of a number of factors: dexterity, left-handedness, practice, age, motivation, tiredness, or any combination of such factors. You may or may not have information about these factors. If you do, you can add them to your model and see if they explain variance and check if the residuals become more randomly distributed. But if you don't have any extra information, or if do you but the residuals remain clustered, you might either consider adding the categorical variable `student` to the model or use linear mixed models, discussed in Chapter **??**.

The assumption of independence is the most important assumption in linear models. Just a small amount of dependence among the observations causes your actual standard error to be much larger than reported by your software. For example, you may think that a confidence interval is [0.1, 0.2], so you reject the null-hypothesis, but in reality the standard error is much larger, with a much wider interval, say [-0.1, 0.4] so that in reality you are not allowed to reject the null-hypothesis. The reason that this happens can be explained when we look again at Figure 7.11. Objectively, there are 100 observations, and this is fed into the software: $n = 100$. This sample size is then used to compute the standard error (see Chapter 5). However, because the reaction times from the same student are so much alike, *effectively* the number of observations is much smaller. The reaction times from one student are in fact so much alike, you could almost say that there are only 10 different reaction times, one for each student, with only slight deviations within each student. Therefore, the real number of observations is somewhere between 10 and 100, and thus the reported standard error is underestimated when there is dependence in your residuals (standard errors are inversely related to sample size, see Chapter 5).

7.3 Linearity

The assumption of linearity is often also referred to as the assumption of *additivity*. Contrary to intuition, the assumption is not that the relationship between variables should be linear. The assumption is that there is linearity or additivity in the parameters. That is, *the effects of the variables in the model* should add up.

Suppose we gather data on height and fear of snakes in 100 children from a different distant country. Figure 7.13 plots these two variables, together with the least squares regression line.

Figure 7.14 shows a pattern in the residuals: the positive residuals seem to be smaller than the negative residuals. We also clearly see a problem when we plot residuals against the predicted fear (see Fig. 7.15). The same problem is reflected in the histogram in Figure 7.16, that does not look symmetric at all. What might be the problem?

Take another look at the data in Figure 7.13. We see that for small heights, the data points are all below the regression line, and the same pattern we see

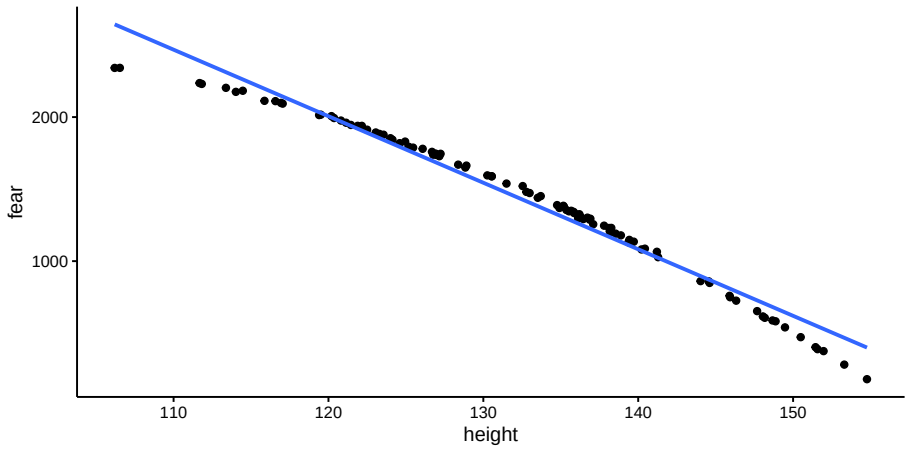


Figure 7.13: Least squares regression line for fear of snakes on height in 100 children.

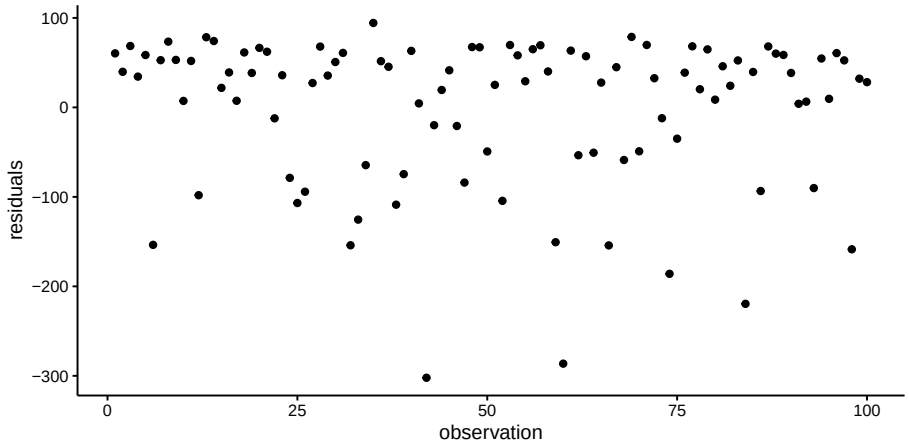


Figure 7.14: Residual plot after regressing fear of snakes on height.

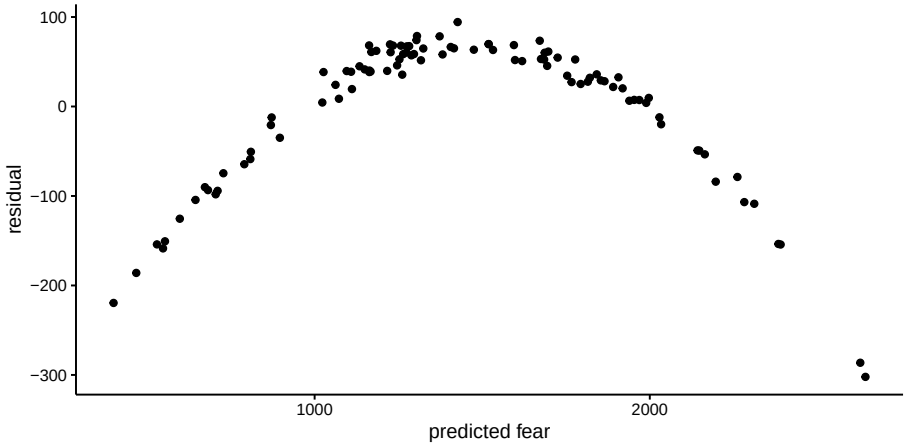


Figure 7.15: Residual plot after regressing fear of snakes on height.

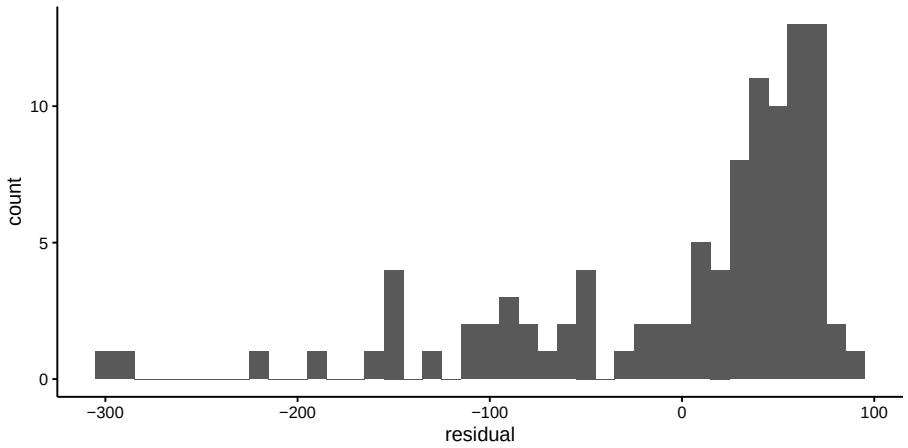


Figure 7.16: Histogram of the residuals after regressing fear of snakes on height.

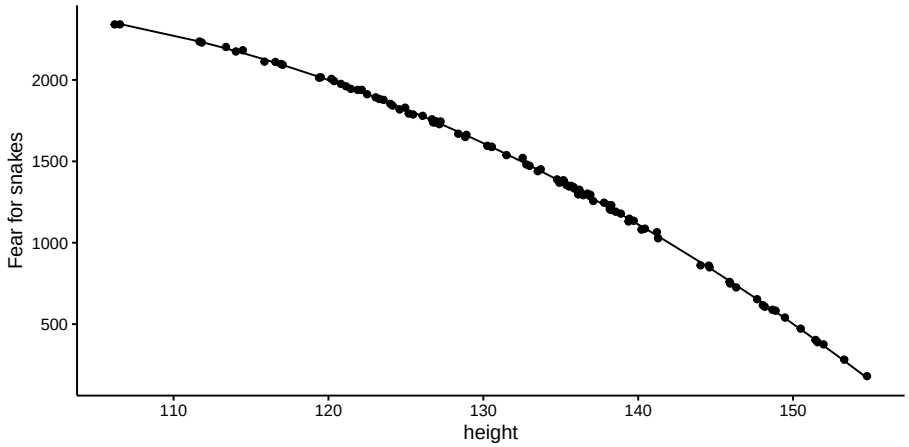


Figure 7.17: Observed and predicted fear based on a linear model with height and height squared

for large heights. For average heights, we see on the contrary all data points above the regression line. Somehow the data points do not suggest a completely linear relationship, but a curved one.

This problem of model misfit could be solved by not only using `height` as the predictor variable, but also the *square* of `height`, that is, $\texttt{height}^2$. For each observed height we compute the square. This new variable, let's call it `height2`, we add to our regression model. The least squares regression equation then becomes:

$$\widehat{\texttt{fear}} = -2000 + 100 \times \texttt{height} - 0.56 \times \texttt{height2} \tag{7.5}$$

If we then plot the data and the regression line, we get Figure 7.17. There we see that the regression line goes straight through the points. Note that the regression line when plotted against `height` is non-linear, but equation 7.5 itself is linear, that is, there are only two effects added up, one from variable `height` and one from variable `height2`. We also see from the histogram (Figure 7.18) and the residuals plot (Figure 7.19) that the residuals are randomly drawn from a normal distribution and are not related to predicted fear. Thus, our additive model (our linear model) with effects of height and height squared results in a nice-fitting model with random normally scattered residuals.

In sum, the relationship between two variables need not be linear in order for a linear model to be appropriate. A transformation of an independent variable, such as taking a square, can result in normally randomly scattered residuals. The linearity assumption is that the effects of a number of variables (transformed or untransformed) add up and lead to a model with normally and independently, randomly scattered residuals.

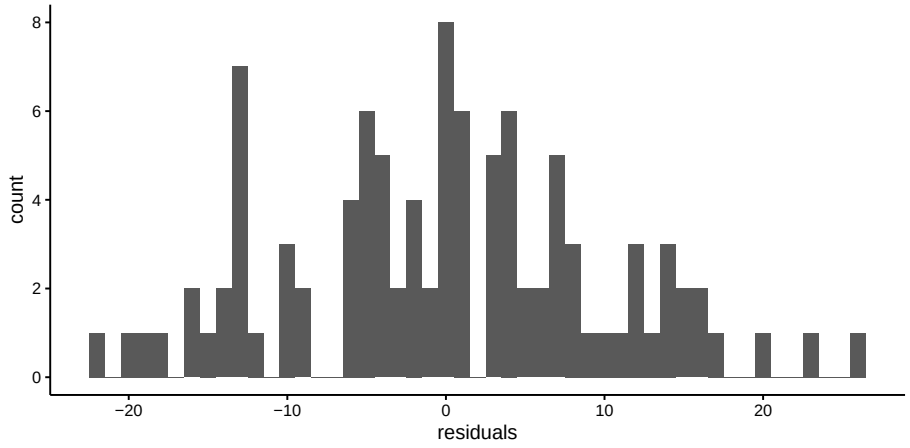


Figure 7.18: Histogram of the residuals of the fear of snakes data with height squared introduced into the linear model.

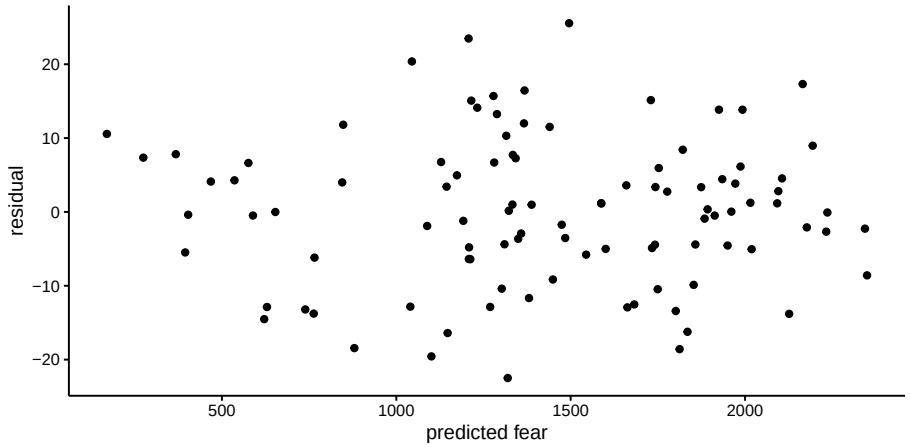


Figure 7.19: Residuals plot of the fear of snakes data with height squared introduced into the linear model.

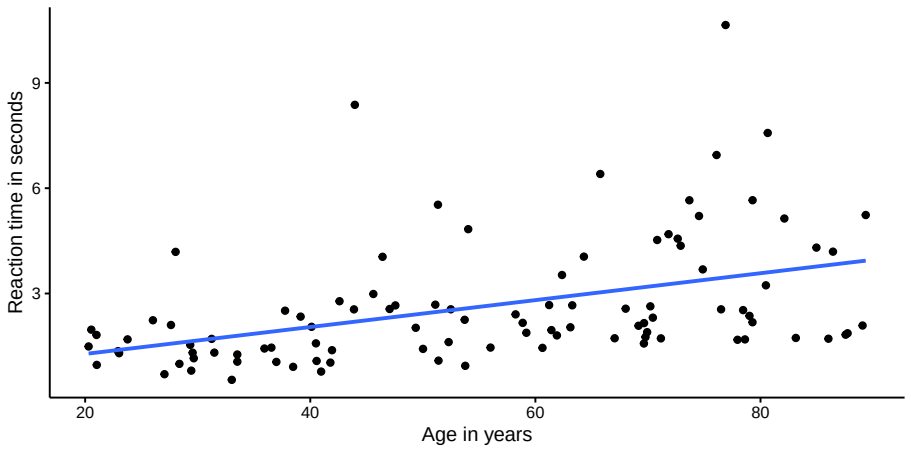


Figure 7.20: Least squares regression line for reaction time on age in 100 adults.

7.4 Equal variances

Suppose we measure reaction times in both young and older adults. Older persons tend to have longer reaction times than young adults. Figure 7.20 shows a data set on 100 persons. Figure 7.21 shows the residuals as a function of age, and shows something remarkable: it seems that the residuals are much more varied for older people than for young people. There is more variance at older ages than at younger ages. This is a violation of the equal variance assumption. Remember that a linear model goes with a normal distribution for the residuals with a certain variance. In a linear model, there is only mention of one variance of the residuals σ^2, not several!

The equal variance assumption is an important one: if the data show that the variance is different for different subgroups of individuals in the data set, then the standard errors of the regression coefficients cannot be trusted.

We often see an equal variance violation in reaction times. An often used strategy of getting rid of such a problem is to work not with the reaction time, but the *logarithm* of the reaction time. Figure 7.22 shows the data with the computed logarithms of reaction time, and Figure 7.23 shows the residuals plot. You can see that the log-transformation of the reaction times resulted in a much better model.

Note that the assumption is not about the variance in the sample data, but about the residuals in the population data. It might well be that there are slight differences in the sample data of the older people than in the sample data of the younger people. These could well be due to chance. The important thing to know is that the assumption of equal variance is that in the population of older adults, the variation in residuals is the same as the variation in residuals in the population of younger adults.

The equal variance assumption is often referred to as the *homogeneity of*

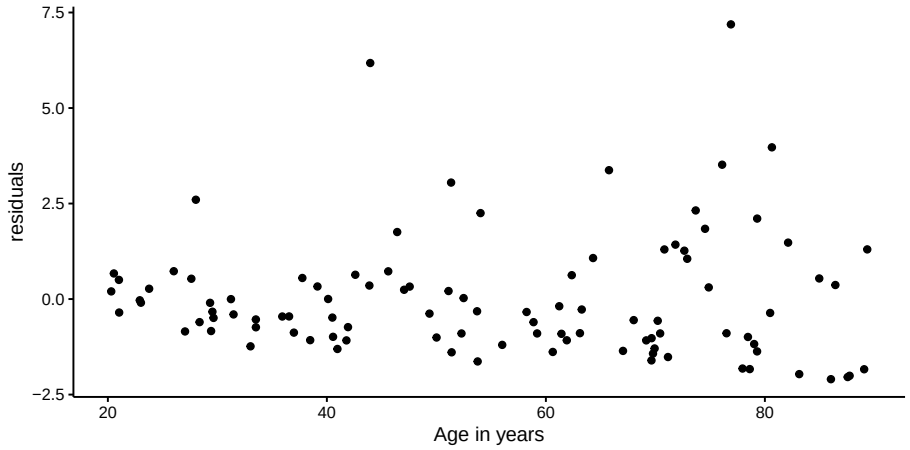


Figure 7.21: Residual plot after regressing reaction time on age.

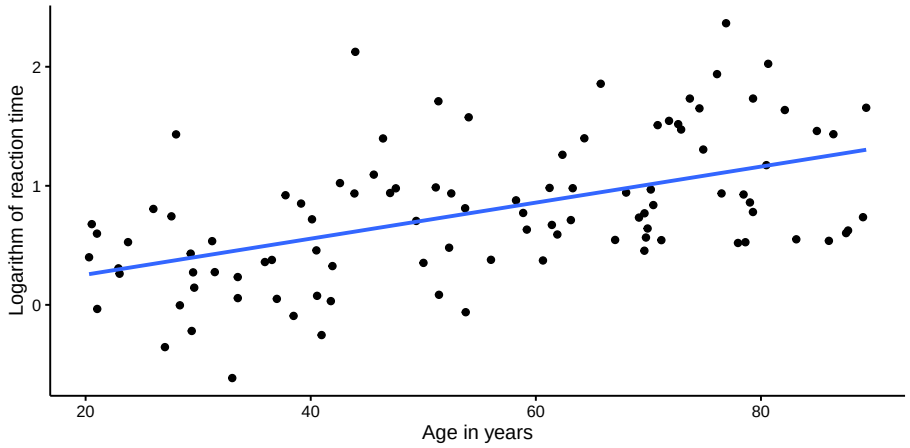


Figure 7.22: Least squares regression line for log reaction time on age in 100 adults.

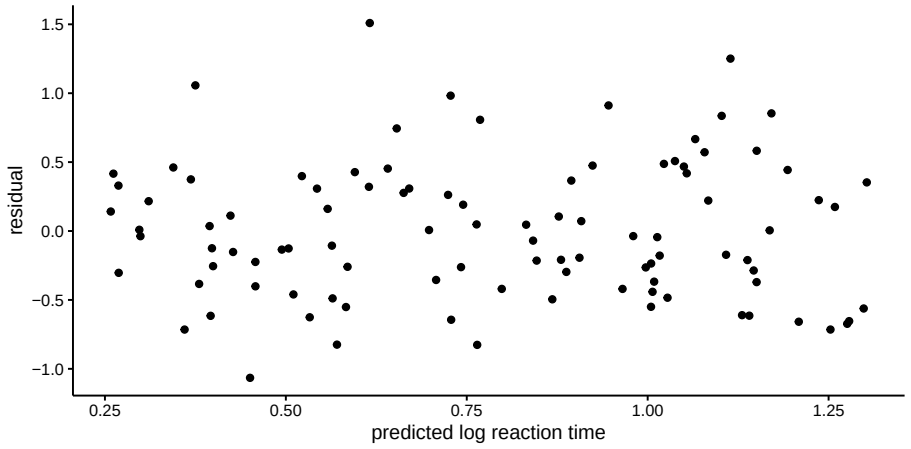


Figure 7.23: Residual plot after regressing log reaction time on age.

variance assumption or *homoscedasticity*. It is the assumption that variance is homogeneous (of equal size) across all levels and subgroups of the independent variables in the population. The computation of the standard error is highly dependent on the size of the variance of the residuals. If the size of this variance differs across levels and subgroups of the data, the standard error also varies and the confidence intervals cannot be easily determined. This in turn has an effect on the computation of p-values, and therefore inference. Having no homogeneity of variance therefore leads to wrong inference, with inflated or deflated type I and type II error rates.

The inflation or deflation of type I and type II error rates are limited in the case that group sizes are more or less equal. For example, suppose you have an age variable with about an equal number of older persons and younger persons, but unequal variances of the residuals. In that case you should not worry too much about the precision of your p-values and your confidence intervals: they are more or less correct. However, if you have more than 1.5 times more elderly in your sample than youngsters (or vice versa), with unequal variances of the residuals, then you should worry. Briefly: if the greater error variance is associated with the greater group size, then the reported p-value is too small, and if the greater error variance is associated with the smaller group size, then the reported p-value is too large. If the p-value is around your pre-chosen α-level and you're unsure whether to reject or not to reject your null-hypothesis, look for more robust methods of computing standard errors.

7.5 Residuals normally distributed

As we've already seen, the assumption of the linear model is that the residuals are normally distributed. Let's look at the reaction time data again and see

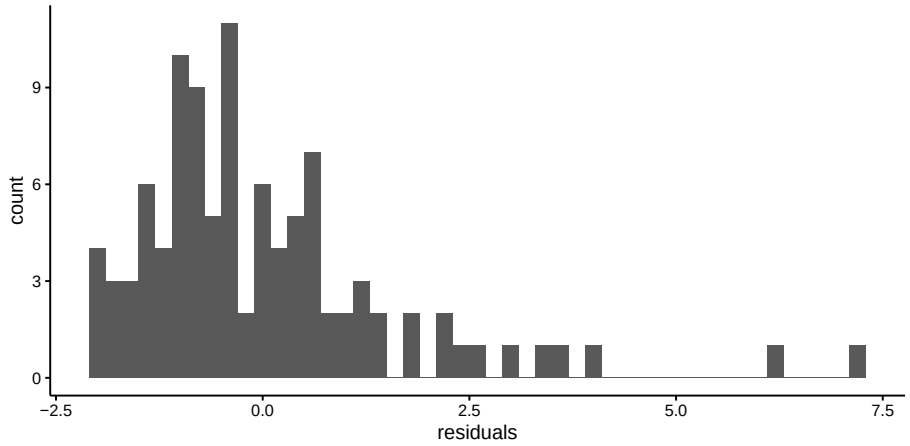


Figure 7.24: Histogram of the residuals after a regression of reaction time on age.

what the histogram of the residuals looks like if we use reaction time as our dependent variable. Figure 7.24 shows that in that case the distribution is not symmetric: it is clearly skewed.

After a logarithmic transformation of the reaction times, we get the histogram in Figure 7.25, which looks more symmetric.

Remember that if your sample size is of limited size, a distribution will never look completely normal, even if it is sampled from a normal distribution. It should however be *likely* to be sampled from a *population* of data that seems normal. That means that the histogram should not be too skewed, or too peaked, or have two peaks far apart. Only if you have a lot of observations, say 1000, you can reasonably say something about the shape of the distribution.

If you have categorical independent variables in your linear model, it is best to look at the various subgroups separately and look at the histogram of the residuals: the residuals e are defined as residuals given the rest of the linear model. For instance, if there is a model for height, and country is the only predictor in the model, all individuals from the same country are given the same expected height based on the model. They only differ from each other because of the normally distributed random residuals. Therefore look at the residuals for all individuals from one particular country to see whether the residuals are indeed normally distributed. Then do this for all countries separately. Think about it: the residuals might look non-normal from country A, and non-normal from country B, but put together, they might look very normal! This is illustrated in Figure 7.26. Therefore, when checking for the assumption of normality, do this for every subgroup separately.

It should be noted that the assumption of normally distributed residuals as checked with a histogram is the least important assumption. Even when the

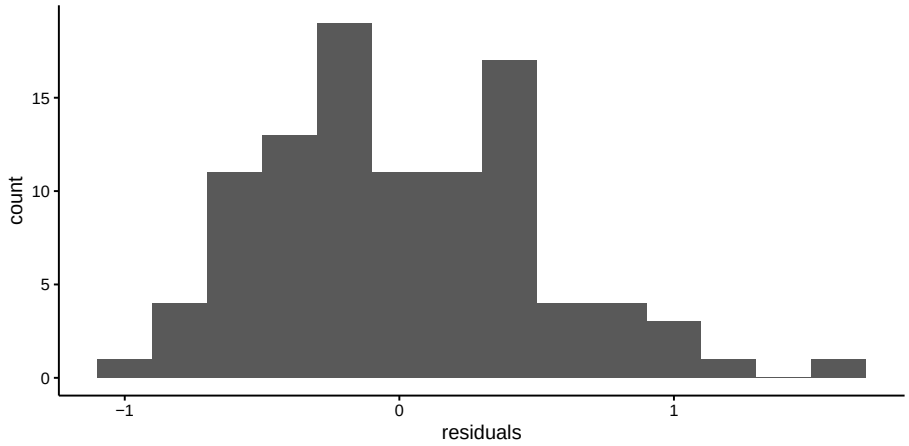


Figure 7.25: Histogram of the residuals after a regression of log reaction time on age.

distribution is skewed, your standard errors are more or less correct. Only in severe cases, like with the residuals in Figure 7.24, the standard errors start to be somewhat incorrect.

7.6 General approach to testing assumptions

It is generally advised to always check the residuals. All four assumptions mentioned above can be checked by looking at the residuals. We advise to do this with three types of plots.

The first is the histogram of the residuals: this shows whether the residuals are more or less normally distributed. The histogram should show a more or less symmetric distribution. If the plot does not look asymmetric at all, try to find a transformation of the dependent variable that makes the residuals more normal. An example of this is to log-transform reaction times.

The second type of plot that you should look at is a plot where the residuals are on the y-axis and the predicted values for the dependent variable ($\widehat{Y}$) is on the x-axis. Such a plot can reveal systematic deviation from normality, but also non-equal variance.

The third type of plot that you should study is one where the residuals are on the vertical axis and one of the predictor variables is on the horizontal axis. In this plot, you can spot violations of the equal variance assumption. You can also use such a plot for candidate predictor variables that are not in your model yet. If you notice a pattern, this is indicative of dependence, which means that this variable should probably be included in your model.

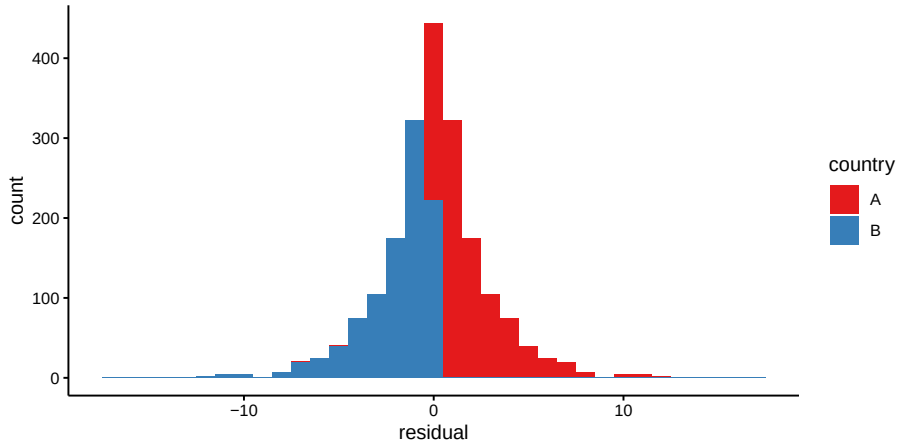


Figure 7.26: Two distributions might be very non-normal, but when taken together, might look normal nevertheless. Normality should therefore always be checked for each subgroup separately.

7.7 Checking assumptions in R

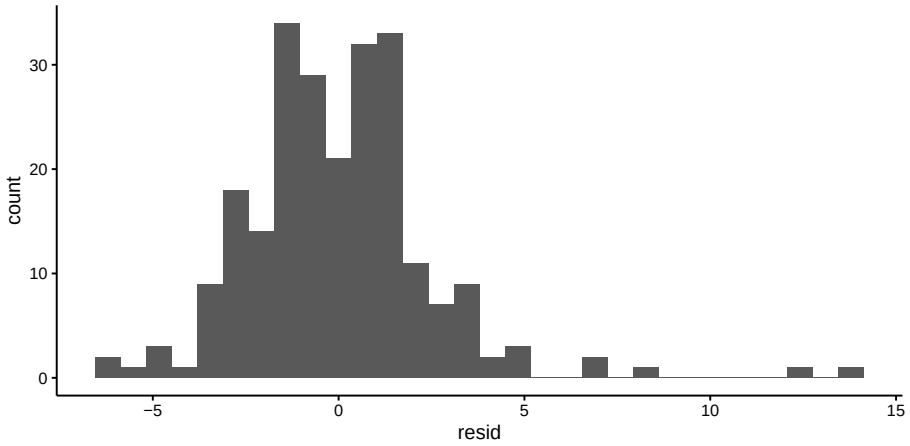
In this section we show the general code for making residual plots in R. We will look at how to make the three types of plots of the residuals to check the four assumptions.

When you run a linear model with the `lm()` function, you can use the package `modelr` to easily obtain the residuals and predicted values that you need for your plots. Let's use the `mpg` data to illustrate the general approach. This data set contains data on 234 cars. First we model the number of city miles per gallon (`cty`) as a function of the number of cylinders (`cyl`).

```
out <- mpg %>%
  lm(cty ~ cyl, data = .)
```

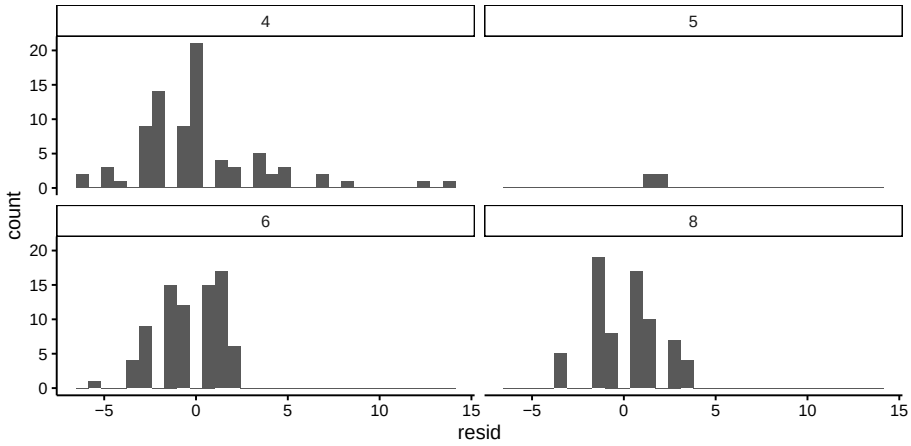
Next, we use the function `add_residuals` from the `modelr` package to add residuals to the data set and plot a histogram.

```
library(modelr)
mpg %>%
  add_residuals(out) %>%
  ggplot(aes(x = resid)) +
  geom_histogram()
```



As stated earlier, it's even better to do this for the different subgroups separately:

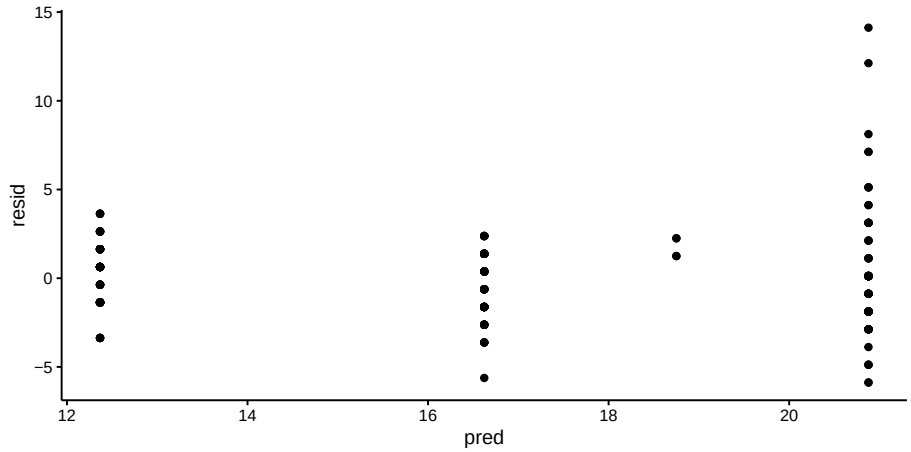
```
mpg %>%
  add_residuals(out) %>%
  ggplot(aes(x = resid)) +
  geom_histogram() +
  facet_wrap(. ~ cyl)
```



For the second type of plot, we use two functions from the `modelr` package to add predicted values and residuals to the data set, and use these to make a residual plot:

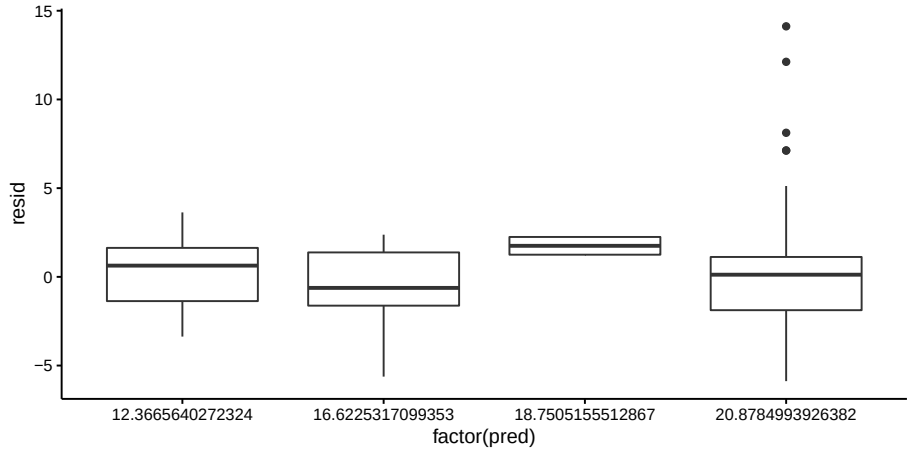
```
mpg %>%
  add_residuals(out) %>%
  add_predictions(out) %>%
  ggplot(aes(x = pred, y = resid)) +
```

```
  geom_point()
```



When there are few values for the predictions, or when you have a categorical predictor, it's better to make a boxplot:

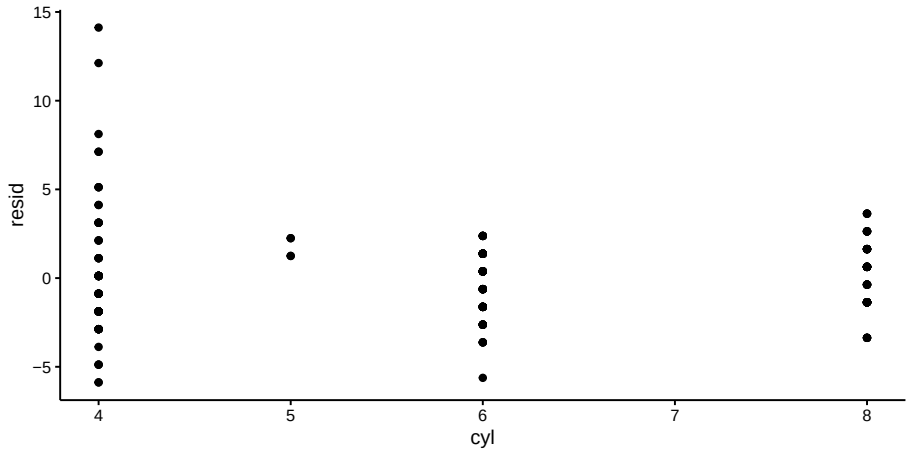
```
mpg %>%
  add_residuals(out) %>%
  add_predictions(out) %>%
  ggplot(aes(x = factor(pred), y = resid)) +
  geom_boxplot()
```



For the third type of plot, we put the predictor on the x-axis and the residual on the y-axis.

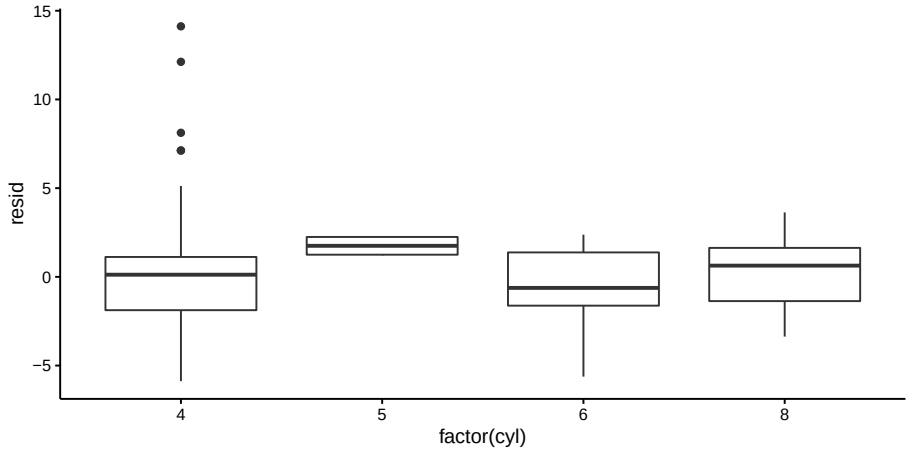
```
mpg %>%
  add_residuals(out) %>%
```

```
  ggplot(aes(x = cyl, y = resid)) +
  geom_point()
```



Again, with categorical variables or variables with very few categories, it is sometimes clearer to use a boxplot:

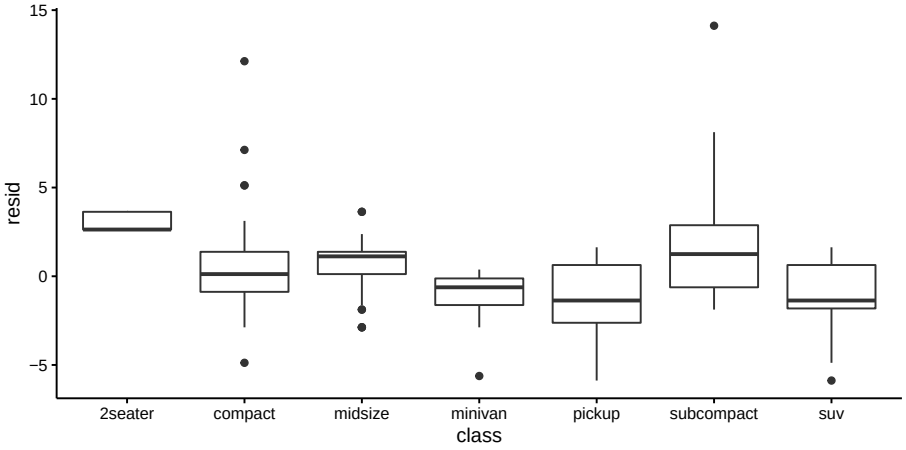
```
mpg %>%
  add_residuals(out) %>%
  ggplot(aes(x = factor(cyl), y = resid)) +
  geom_boxplot()
```



To check for independence you can also put variables on the x-axis that are not in the model yet, for example the type of the car (`class`):

```
mpg %>%
  add_residuals(out) %>%
```

```
ggplot(aes(x = class, y = resid)) +
geom_boxplot()
```



Chapter 8

When assumptions are not met: non-parametric alternatives

8.1 Introduction

Linear models do not apply to every data set. As discussed in Chapter 7, sometimes the assumptions of linear models are not met. One of the assumptions is linearity or additivity. Additivity requires that one unit change in variable X leads to the same amount of change in Y, no matter what value X has. For bivariate relationships this leads to a linear shape. But sometimes you can only expect that Y will change in the same direction, but you don't believe that this amount is the same for all values of X. This is the case for example with an ordinal dependent variable. Suppose we wish to model the relationship between the age of a mother and an aggression score in her 7-year-old child. Suppose aggression is measured on a three-point ordinal scale: 'not aggressive', 'sometimes aggressive', 'often aggressive'. Since we do not know the quantitative differences between these three levels, there are many graphs we could draw for a given data set.

Suppose we have the data set given in Table 8.1. If we want to make a scatter plot, we could arbitrarily choose the values 1, 2, and 3 for the three categories, respectively. We would then get the plot in Figure 8.1. But since the aggression data are ordinal, we could also choose the arbitrary numeric values 0, 2, and 3, which would yield the plot in Figure 8.2.

As you can see from the least squares regression lines in Figures 8.1 and 8.2, when we change the way in which we code the ordinal variable into a numeric one, we also see the best fitting regression line changing. This does not mean though, that ordinal data cannot be modelled linearly. Look at the example data in Table 8.2 where aggression is measured with a 7-point scale. Plotting these

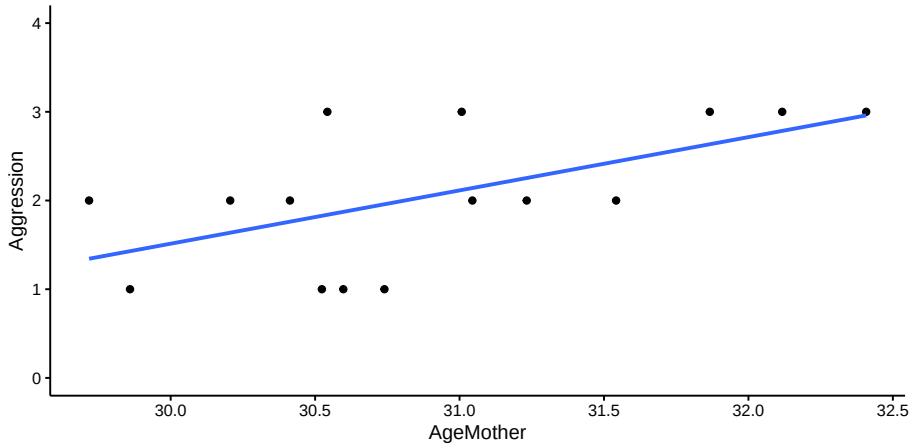


Figure 8.1: Regression of the child's aggression score (1,2,3) on the mother's age.

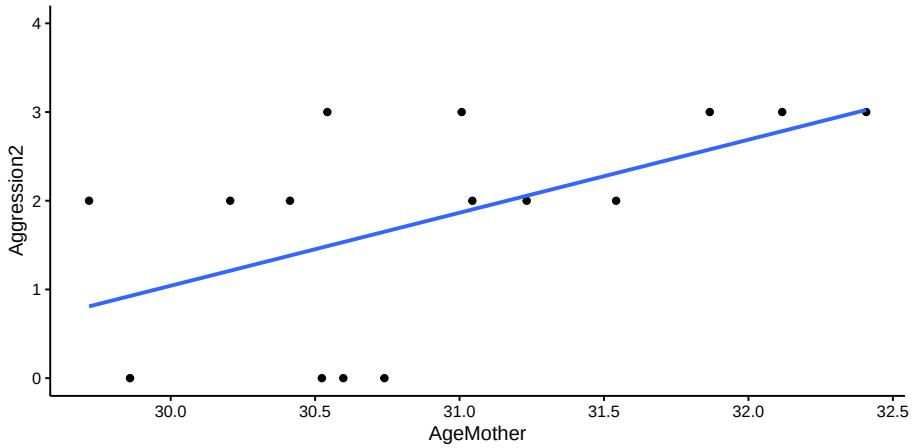


Figure 8.2: Regression of the child's aggression score (0,2,3) on the mother's age.

Table 8.1: Aggression in children and age of the mother.

AgeMother	Aggression
32.00	Sometimes aggressive
31.00	Often aggressive
32.00	Often aggressive
30.00	Not aggressive
31.00	Sometimes aggressive
30.00	Sometimes aggressive
31.00	Not aggressive
31.00	Often aggressive
31.00	Not aggressive
30.00	Sometimes aggressive
32.00	Often aggressive
32.00	Often aggressive
31.00	Sometimes aggressive
30.00	Sometimes aggressive
31.00	Not aggressive

data in Figure 8.3 using the values 1 through 7, we see a nice linear relationship. So even when the values 1 thru 7 are arbitrarily chosen, a linear model can be a good model for a given data set with one or more ordinal variables. Whether the interpretation makes sense is however up to the researcher.

So with ordinal data, always check that your data indeed conform to a linear model, but realise at the same time that you're assuming a *quantitative* and additive relationship between the variables that may or may not make sense. If you believe that a quantitative analysis is meaningless then you may consider a non-parametric analysis that we discuss in this chapter.

Another instance where we favour a non-parametric analysis over a linear model one, is when the assumption of normally distributed residuals is not tenable. For instance, look again at Figure 8.1 where we regressed aggression in the child on the age of its mother. Figure 8.4 shows a histogram of the residuals. Because of the limited number of possible values in the dependent variable (1, 2 and 3), the number of possible values for the residuals is also very restricted, which leads to a very discrete distribution. The histogram looks therefore far removed from a continuous symmetric, bell-shaped distribution, which is a violation of the normality assumption.

Every time we see a distribution of residuals that is either very skew, or has very few different values, we should consider a non-parametric analysis. Note that the shape of the distribution of the residuals is directly related to what scale values we choose for the ordinal categories. By changing the values we change the regression line, and that directly affects the relative sizes of the residuals.

First, we will discuss a non-parametric alternative for two numeric variables. We will start with Spearman's ρ (rho, pronounced 'row'), also called Spearman's rank-order correlation coefficient r_s. Next we will discuss an alternative to r_s,

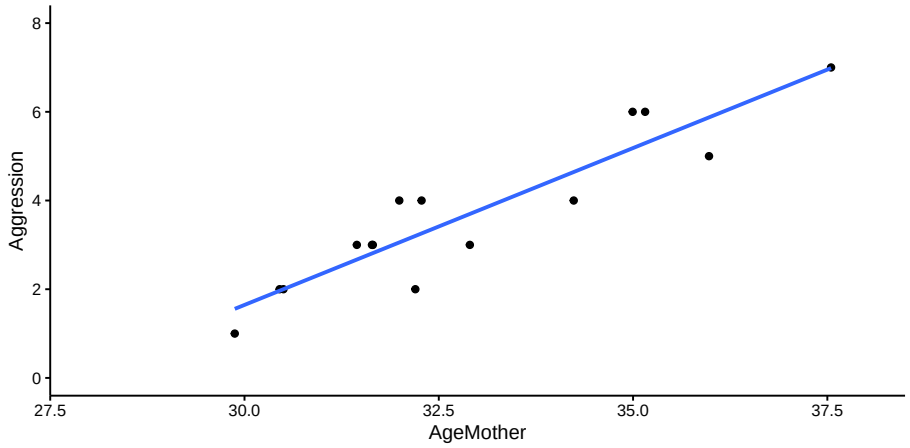


Figure 8.3: Regression of the child's aggression 1 thru 7 Likert score on the mother's age.

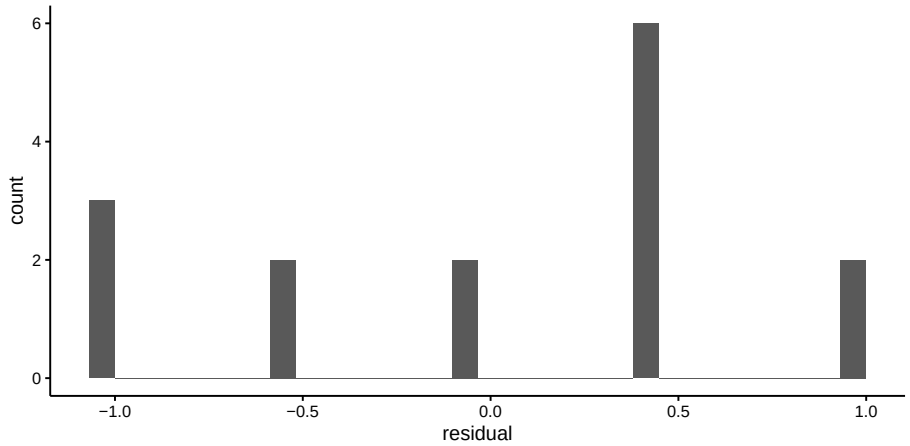


Figure 8.4: Histogram of the residuals after the regression of a child's aggression score on the mother's age.

Table 8.2: Aggression in children on a 7-point Likert scale and age of the mother.

AgeMother	Aggression
35.0	6
32.0	4
35.2	6
36.0	5
32.9	3
29.9	1
32.3	4
32.2	2
34.2	4
30.5	2
31.6	3
30.5	2
31.7	3
31.4	3
37.5	7

Kendall's τ (tau, pronounced 'taw'). After that, we will discuss the combination of numeric and categorical variables, when comparing groups.

8.2 Spearman's ρ (rho)

Suppose we have 10 students and we ask their teachers to rate them on their performance. One teacher rates them on geography and the other teacher rates them on history. We only ask them to give *rankings*: indicate the brightest student with a 1 and the dullest student with a 10. Then we might see the data set in Table 8.3. We see that student 9 is the brightest student in both geography and history, and student 7 is the dullest student in both subjects.

Table 8.3: Student rankings on geography and history.

student	rank.geography	rank.history
1	5	4
2	4	5
3	6	7
4	7	8
5	8	6
6	9	9
7	10	10
8	2	3
9	1	1
10	3	2

Now we acknowledge the ordinal nature of the data by only having rankings: a person with rank 1 is brighter than a person with rank 2, but we do not how large the difference in brightness really is. Now we want to establish to what extent there is a relationship between rankings on geography and the rankings on history: the higher the ranking on geography, the higher the ranking on history?

By eye-balling the data, we see that the brightest student in geography is also the brightest student in history (rank 1). We also see that the dullest student in history is also the dullest student in geography (rank 10). Furthermore, we see relatively small differences between the rankings on the two subjects: high rankings on geography seem to go together with high rankings on history. Let's look at these differences between rankings more closely by computing them, see Table 8.4.

Table 8.4: Student rankings on geography and history.

student	rank.geography	rank.history	difference
1	5	4	-1
2	4	5	1
3	6	7	1
4	7	8	1
5	8	6	-2
6	9	9	0
7	10	10	0
8	2	3	1
9	1	1	0
10	3	2	-1

So theoretically the difference could be as large as 9, but here we see a biggest difference of -2. When all differences are small, this says something about how the two rankings overlap: they are related. We could compute an average difference: the average difference is the sum of these differences, divided by 10, so we get 0. This is because we have both plus and minus values. It would be better to take the square of the differences, so that we would get positive values, see Table 8.5.

Now we can compute the average squared difference, which is equal to 10/10 = 1. Generally, the smaller this value, the closer the rankings of the two teachers are together, and the more correlation there is between the two subjects.

A clever mathematician like Spearman showed that is even better to use a somewhat different measure for a correlation between ranks. He showed that it is wiser to compute the following statistic, where d is the difference in rank and d^2 is the squared difference (and n is sample size):

$$r_s = 1 - \frac{6 \sum d^2}{n^3 - n} \tag{8.1}$$

Table 8.5: Student rankings on geography and history.

rank.geography	rank.history	difference	squared.difference
5	4	-1	1
4	5	1	1
6	7	1	1
7	8	1	1
8	6	-2	4
9	9	0	0
10	10	0	0
2	3	1	1
1	1	0	0
3	2	-1	1

because then you get a value between -1 and 1, just like a Pearson correlation, where a value close to 1 describes a high positive correlation (high rank on one variable goes together with a high rank on the other variable) and a value close to -1 describes a negative correlation (a high rank on one variable goes together with a low rank on the other variable). So in our case the sum of the squared differences is equal to 10, and n is the number of students, so we get:

$$r_s = 1 - \frac{6 \times 10}{10^3 - 10} = 1 - \frac{60}{990} = 0.94 \tag{8.2}$$

This is called the Spearman rank-order correlation coefficient r_s, or Spearman's rho (the Greek letter ρ). It can be used for any two variables of which at least one is ordinal. The trick is to convert the scale values into ranks, and then apply the formula above. For instance, if we have the variable `Grade` with the following values (C, B, D, A, F), we convert them into rankings by saying the A is the highest value (1), B is the second highest value (2), C is the third highest value (3), D is the fourth highest value (4) and F is the fifth highest value (5). So transformed into ranks we get (3, 2, 4, 1, 5). Similarly, we could turn numeric variables into ranks. Table 8.6 shows how the variables `grade`, `shoesize` and `height` are transformed into their respective ranked versions. Note that the ranking is alphanumerically by default: the first alphanumeric value gets rank 1. You could also do the ranking in the opposite direction, if that makes more sense.

Table 8.6: Ordinal and numeric variables and their ranked transformations.

student	grade	rank.grade	shoesize	rank.shoesize	height	rank.height
1	A	1	6	1	1.70	1
2	D	4	8	3	1.82	2
3	C	3	9	4	1.92	4
4	B	2	7	2	1.88	3

8.3 Spearman's rho in R

When we let R compute r_s for us, it automatically ranks the data for us. Let's look at the `mpg` data on 234 cars from the `ggplot2` package again. Suppose we want to treat the variables `cyl` (the number of cylinders) and `year` (year of the model) as ordinal variables, and we want to look whether the ranking on the `cyl` variable is related to the ranking on the `year` variable. We use the function `rcorr()` from the `Hmisc` package to compute Pearson's rho:

```
library(Hmisc)
rcorr(mpg$cty, mpg$year, type = "spearman")
```

```
##       x     y
## x  1.00 -0.01
## y -0.01  1.00
##
## n= 234
##
##
## P
##   x      y
## x        0.9169
## y 0.9169
```

In the output you will see a correlation matrix very similar the one for a Pearson correlation. Spearman's rho is equal -0.01. You will also see whether the correlation is significantly different from 0, indicated by a p-value. If the p-value is very small, you may conclude that on the basis of these data, the correlation in the population is not equal to 0, ergo, in the population there is a relationship between the year a car was produced and the number of cylinders.

8.4 Kendall's rank-order correlation coefficient τ

If you want to study the relationship between two variables, of which at least one is ordinal, you can either use Spearman's r_s or Kendall's τ (tau, pronounced 'taw' as in 'law'). However, if you have three variables, and you want to know whether there is a relationship between variables A and B, over and above the effect of variable C, you can use an extension of Kendall's τ. Note that this is very similar to the idea of multiple regression: a coefficient for variable X_1 in multiple regression with two predictors is the effect of X_1 on Y over and above the effect of X_2 on Y. The logic of Kendall's τ is also based on rank orderings, but it involves a different computation. Let's look at the student data again with the teachers' rankings of ten students on two subjects in Table 8.4.

Table 8.7: Student rankings on geography and history, now ordered according to the ranking for geography.

student	rank.geography	rank.history
9	1	1
8	2	3
10	3	2
2	4	5
1	5	4
3	6	7
4	7	8
5	8	6
6	9	9
7	10	10

From this table we see that the history teacher disagrees with the geography teacher that student 8 is brighter than student 10. She also disagrees with her colleague that student 1 is brighter than student 2. If we do this for all possible pairs of students, we can count the number of times that they agree and we can count the number of times they disagree. The total number of possible pairs is equal to $\binom{10}{2} = n(n-1)/2 = 90/2 = 45$ (see Chapter 3). This is a rather tedious job to do, but it can be made simpler if we reshuffle the data a bit. We put the students in a new order, such that the brightest student in geography comes first, and the dullest last. This also changes the order in the variable history. We then get the data in Table 8.7. We see that the geography teacher believes that student 9 outperforms all 9 other students. On this, the history teacher agrees, as she also ranks student 9 first. This gives us 9 agreements. Moving down the list, we see that the geography teacher believes student 8 outperforms 8 other students. However, we see that the history teacher believes student 8 only outperforms 7 other students. This results in 7 agreements and 1 disagreement. So now in total we have $9 + 7 = 16$ agreements and 1 disagreements. If we go down the whole list in the same way, we will find that there are in total 41 agreements and 4 disagreements.

The computation is rather tedious. There is a trick to do it faster. Now focus on Table 8.7 but start in the column of the history teacher. Start at the top row and count the number of rows beneath it with a rank higher than the rank in the first row. The rank in the first row is 1, and all other ranks beneath it are higher, so the number of ranks is 9. We plug that value in the last column in Table 8.8. Next we move to row 2. The rank is 3. We count the number of rows below row 2 with a rank higher than 3. Rank 2 is lower, so we are left with 7 rows and we again plug 7 in the last column of Table 8.8. Then we move on to row 3, with rank 2. There are 7 rows left, and all of them have a higher rank. So the number is 7. Then we move on to row 4. It has rank 5. Of the 6 rows below it, only 5 have a higher rank. Next, row 5 shows rank 4. Of the 5 rows below it, all 5 show a higher rank. Row 6 shows rank 7. Of the 4 rows

Table 8.8: Student rankings on geography and history, now ordered according to the ranking for geography, with number of agreements.

student	rank.geography	rank.history	number
9	1	1	9
8	2	3	7
10	3	2	7
2	4	5	5
1	5	4	5
3	6	7	3
4	7	8	2
5	8	6	2
6	9	9	1
7	10	10	0

below it, only 3 show a higher rank. Row 7 shows rank 8. Of the three rows below it, only 2 show a higher rank. Row 8 shows rank 6. Both rows below it show a higher rank. And row 9 shows rank 9, and the row below it shows a higher rank so that is 1. Finally, when we add up the values in the last column in Table 8.8, we find 41. This is the number of agreements. The number of disagreements can be found by reasoning that the total number of pairs equals the number of pairs that can be formed using a total number of 10 objects: $\binom{10}{2}$ $= 10(10-1)/2 = 45$. In this case we have 45 possible pairs. Of these there are 41 agreements, so there must be $45 - 41 = 4$ disagreements. We can then fill in the formula to compute Kendall's τ:

$$\tau = \frac{agreements - disagreements}{totalnumberofpairs} = \frac{37}{45} = 0.82 \tag{8.3}$$

This τ-statistic varies between -1 and 1 and can therefore be seen as a non-parametric analogue of a Pearson correlation. Here, the teachers more often agree than disagree, and therefore the correlation is positive. A negative correlation means that the teachers more often disagree than agree on the relative brightness of their students.

As said, the advantage of Kendall's τ over Spearman's r_s is that Kendall's τ can be extended to cover the case that you wish to establish the strength of the relationships of two variables A and B, over and above the relationship with C. The next section shows how to do that in R.

8.5 Kendall's τ in R

Let's again use the `mpg` data on 234 cars. We can compute Kendall's τ for the variables `cyl` and `year` using the `Kendall` package:

```r
library(Kendall)
Kendall(mpg$cyl, mpg$year)

## tau = 0.112, 2-sided pvalue =0.068798
```

As said, Kendall's τ can also be used if you want to control for a third variable (or even more variables). This can be done with the `ppcor` package. Because this package has its own function `select()`, you need to be explicit about which function from which package you want to use. Here you want to use the `select()` function from the `dplyr` package (part of the tidyverse suite of packages).

```r
library(ppcor)
mpg %>%
  dplyr::select(cyl, year, cty) %>%
  pcor(method = "kendall")

## $estimate
##             cyl      year        cty
## cyl   1.0000000 0.1642373 -0.7599993
## year  0.1642373 1.0000000  0.1210952
## cty  -0.7599993 0.1210952  1.0000000
##
## $p.value
##                                                                            cyl
## cyl  0.000000000000000000000000000000000000000000000000000000000000000000000000
## year 0.000189654827628346264994235736978112072392832487821578979492187500000000
## cty  0.000000000000000000000000000000000000000000000000000000000000000000770657
##             year
## cyl  0.0001896548
## year 0.0000000000
## cty  0.0059236967
##                                                                            cty
## cyl  0.000000000000000000000000000000000000000000000000000000000000000000770657
## year 0.005923696652933938683327497187747212592512369155883789062500000000000000
## cty  0.000000000000000000000000000000000000000000000000000000000000000000000000
##
## $statistic
##             cyl     year        cty
## cyl    0.000000 3.732412 -17.271535
## year   3.732412 0.000000   2.751975
## cty  -17.271535 2.751975   0.000000
##
## $n
## [1] 234
##
```

```
## $gp
## [1] 1
##
## $method
## [1] "kendall"
```

In the output, we see that the Kendall correlation between `cyl` and `year`, controlled for `cty`, equals 0.16, with an associated p-value of 0.00019.

8.6 Kruskall-Wallis test for group comparisons

Now that we have discussed relationships between ordinal and numeric variables, let's have a look at the case where we also have categorical variables.

Suppose we have three groups of students that go on a field trip together: mathematicians, psychologists and engineers. Each can pick a rain coat, with five possible sizes: 'extra small', 'small', 'medium', 'large' or 'extra large'. We want to know if preferred size is different in the three populations, so that teachers can be better prepared in the future. Now we have information about size, but this knowledge is not numeric: we do not know the difference in size between 'medium' and 'large', only that 'large' is larger than 'medium.' We have ordinal data, so computing a mean is impossible here. Even if we would assign values like 1= 'extra small', 2='small', 3= 'medium', etcetera, the mean would be rather meaningless as these values are arbitrary. So instead of focussing on means, we can focus on medians: the middle values. For instance, the median value for our sample of mathematicians could be 'medium', for our sample of psychologists 'small', and for our sample of engineers 'large.' Our question might then be whether the median values in the three populations are really different.

This can be assessed using the Kruskall-Wallis test. Similar to Spearman's r_s and Kendall's τ, the data are transformed into ranks. This is done for all data at once, so for all students together.

For example, if we had the data in Table 8.6, we could transform the variable `size` into ranks, from smallest to largest. Student 1 has size 'extra small' so he or she gets the rank 1. Next, both student 4 and student 6 have size 'small', so they should get ranks 2 and 3. However, because there is no reason to prefer one student over the other, we give them both the *mean* of ranks 2 and 3, so they both get the rank 2.5. Next in line is student 3 with size 'medium' and (s)he gets rank 4. Next in line is student 5 with size 'large' (rank 5) and last in line is student 2 with size 'extra large' (rank 6).

Next, we could compute the average rank per group. The group with the smallest sizes would have the lowest average rank, etcetera. Under the null-hypothesis, if the distribution of size were the same in all three groups, the average ranks would be about the same. If the average rank is very different across groups, this is an indication that size is not distributed equally among the three groups. In order to have a proper statistical test, a rather complex formula is used to compute the so-called KW-statistic, see Castellan & Siegel

Table 8.9: Field trip data.

student	group	size	rank
001	math	extra small	1
002	math	extra large	6
003	psych	medium	4
004	psych	small	2.5
005	engineer	large	5
006	math	small	2.5

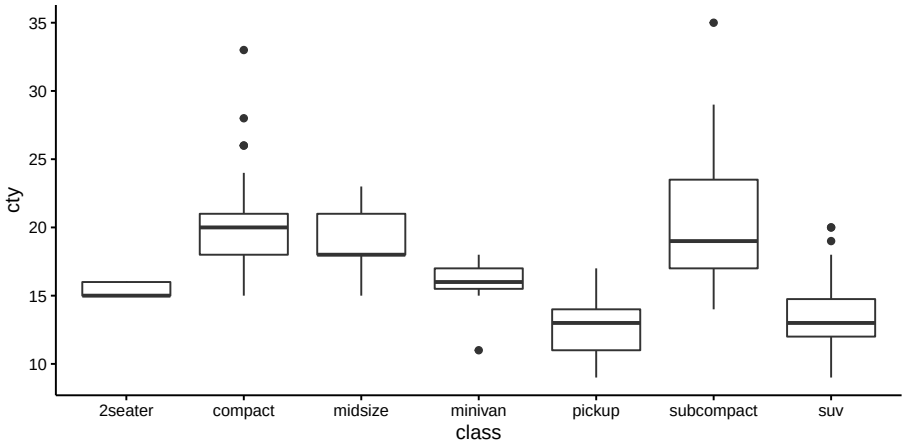


Figure 8.5: Distributions of city mileage (`city`) as a function of car type (`class`).

(1988), that you don't need to know. The distribution of this KW-statistic under the null-hypothesis is known, so we know what extreme values are, and consequently can compute p-values. This tedious computation can be done in R.

8.7 Kruskall-Wallis test in R

Let's look at the `mpg` data again. It contains data on cars from 7 different types. Suppose we want to know whether the different types show difference in the median city miles per gallon. The medians are plotted in Figure 8.5: it seems that indeed the distributions of `cty` are very different for different car types. But these differences could be due to sampling: maybe by accident, the pick-up trucks in this sample happened to have a relatively low mileage, and that the differences in the population of all cars are non-existing. To test this null-hypothesis, we run the following R code:

```
mpg %>%
  kruskal.test(cty ~ class, data = .)

##
##  Kruskal-Wallis rank sum test
##
## data:  cty by class
## Kruskal-Wallis chi-squared = 149.53, df = 6, p-value <
## 0.00000000000000022
```

The output gives us the ability to give the following report:

> "The null-hypothesis that city miles per gallon is distributed equally for all types of cars was tested using a Kruskall-Wallis test with an α of 0.05. Results showed that the null-hypothesis could be rejected, $X^2 = 149.53$, $df = 6$, $p < 0.0001$."

Chapter 9

Moderation: testing interaction effects

9.1 Interaction with one numeric and one dichotomous variable

Suppose there is a linear relationship between age (in years) and vocabulary (the number of words one knows): the older you get, the more words you know. Suppose we have the following linear regression equation for this relationship:

$$\widehat{\texttt{vocab}} = 205 + 500 \times \texttt{age} \tag{9.1}$$

According to this equation, the expected number of words for a newborn baby (`age = 0`) equals 205. This may sound silly, but suppose this model is a very good prediction model for vocabulary size in children between 2 and 5 years of age. Then this equation tells us that the expected increase in vocabulary size is 500 words per year.

This model is meant for everybody in the Netherlands. But suppose that one researcher expects that the increase in words is much faster in children from high socio-economic status (SES) families than in children from low SES families. He believes that vocabulary will be larger in higher SES children than in low SES children. In other words, he expects an effect of SES, over and above the effect of age:

$$\widehat{\texttt{vocab}} = b_0 + b_1 \times \texttt{age} + b_2 \times \texttt{SES} \tag{9.2}$$

This *main effect* of `SES` is yet unknown and denoted by b_2. Note that this linear equation is an example of multiple regression.

Let's use some numerical example. Suppose age is coded in years, and SES is dummy coded, with a 1 for high SES and a 0 for low SES. Let b_2, the effect

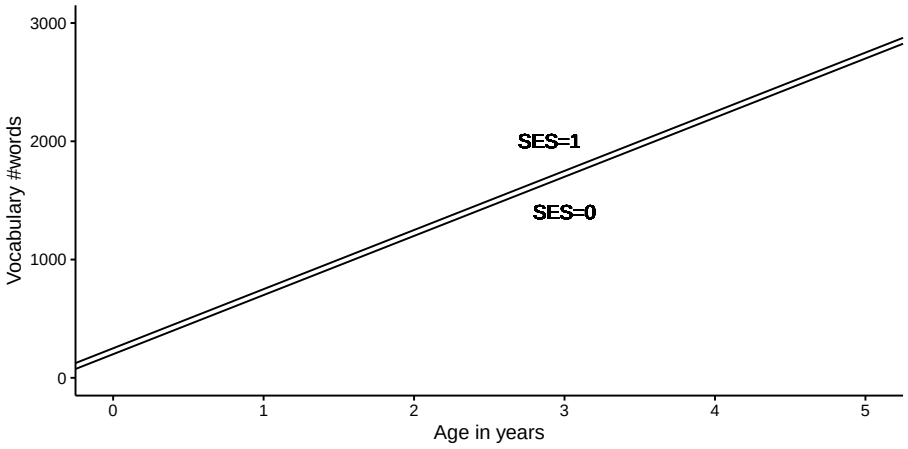


Figure 9.1: Two regression lines: one for low SES children and one for high SES children.

of SES over and above age, be 10. Then we can write out the linear equation for low SES and high SES separately.

$$
\begin{aligned}
lowSES: \widehat{\texttt{vocab}} &= 200 + 500 \times \texttt{age} + 10 \times 0 && (9.3)\\
&= 200 + 500 \times \texttt{age} && (9.4)\\
highSES: \widehat{\texttt{vocab}} &= 200 + 500 \times \texttt{age} + 10 \times 1 && (9.5)\\
&= (200 + 10) + 500 \times \texttt{age} && (9.6)\\
&= 210 + 500 \times \texttt{age} && (9.7)
\end{aligned}
$$

Figure 9.1 depicts the two regression lines for the high and low SES children separately. We see that the effect of SES involves a change in the intercept: the intercept equals 200 for low SES children and the intercept for high SES children equals 210. The difference in intercept is indicated by the coefficient for SES. Note that the two regression lines are parallel: for every age, the difference between the two lines is equal to 10. For every age therefore, the predicted number of words is 10 words more for high SES children than for low SES children.

So far, this is an example of multiple regression that we already saw in Chapter 4. But suppose that such a model does not describe the data that we actually have, or does not make the right predictions based on on our theories. Suppose our researcher also expects that the *yearly increase* in vocabulary is a bit lower than 500 words in low SES families, and a little bit higher than 500 words in high SES families. In other words, he believes that `SES` might *moderate* (affect or change) the slope coefficient for `age`. Let's call the slope coefficient in this case b_1. In the above equation this slope parameter is equal to 500, but let's now let itself have a linear relationship with `SES`:

$$b_1 = a + b_3 \times \texttt{SES} \tag{9.8}$$

In words: the slope coefficient for the regression of `vocab` on `age`, is itself linearly related to `SES`: we predict the slope on the basis of `SES`. We model that by including a slope b_3, but also an intercept a. Now we have *two* linear equations for the relationship between `vocab`, `age` and `SES`:

$$\begin{aligned} \widehat{\texttt{vocab}} &= b_0 + b_1 \times \texttt{age} + b_2 \times \texttt{SES} && (9.9) \\ b_1 &= a + b_3 \times \texttt{SES} && (9.10) \end{aligned}$$

We can rewrite this by plugging the second equation into the first one (substitution):

$$\widehat{\texttt{vocab}} = b_0 + (a + b_3 \times \texttt{SES}) \times \texttt{age} + b_2 \times \texttt{SES} \tag{9.11}$$

Multiplying this out gets us:

$$\widehat{\texttt{vocab}} = b_0 + a \times \texttt{age} + b_3 \times \texttt{SES} \times \texttt{age} + b_2 \times \texttt{SES} \tag{9.12}$$

If we rearrange the terms a bit, we get:

$$\widehat{\texttt{vocab}} = b_0 + a \times \texttt{age} + b_2 \times \texttt{SES} + b_3 \times \texttt{SES} \times \texttt{age} \tag{9.13}$$

Now this very much looks like a regression equation with one intercept and *three* slope coefficients: one for `age` (a), one for `SES` (b_2) and one for `SES` $\times$ `age` (b_3).

We might want to change the label a into b_1 to get a more familiar looking form:

$$\widehat{\texttt{vocab}} = b_0 + b_1 \times \texttt{age} + b_2 \times \texttt{SES} + b_3 \times \texttt{SES} \times \texttt{age} \tag{9.14}$$

So the first slope coefficient is the increase in vocabulary for every year that `age` increases (b_1), the second slope coefficient is the increase in vocabulary for an increase of 1 on the `SES` variable (b_2), and the third slope coefficient is the increase in vocabulary for every increase of 1 on the *product* of `SES` and `age` (b_3).

What does this mean exactly?

Suppose we find the following parameter values for the regression equation:

$$\widehat{\texttt{vocab}} = 200 + 450 \times \texttt{age} + 125 \times \texttt{SES} + 100 \times \texttt{SES} \times \texttt{age} \tag{9.15}$$

If we code low SES children as `SES = 0`, and high SES children as `SES = 1`, we can write the above equation into two regression equations, one for low SES children (`SES = 0`) and one for high SES children (`SES = 1`):

$$\begin{aligned} lowSES: \widehat{\texttt{vocab}} &= 200 + 450 \times \texttt{age} && (9.16) \\ highSES: \widehat{\texttt{vocab}} &= 200 + 450 \times \texttt{age} + 125 + 100 \times \texttt{age} && (9.17) \\ &= (200 + 125) + (450 + 100) \times \texttt{age} \\ &= 325 + 550 \times \texttt{age} \end{aligned}$$

Then for low SES children, the intercept is 200 and the regression slope for age is 450, so they learn 450 words per year. For high SES children, we see the same intercept of 200, with an extra 125 (this is the main effect of SES). So effectively their intercept is now 325. For the regression slope, we now have $450 \times \texttt{age} + 100 \times \texttt{age}$ which is of course equal to $550 \times \texttt{age}$. So we see that the high SES group has both a different intercept, and a different slope: the increase in vocabulary is 550 per year: somewhat steeper than in low SES children. So yes, the researcher was right: vocabulary increase per year is faster in high SES children than in low SES children.

These two different regression lines are depicted in Figure 9.2. It can be clearly seen that the lines have two different intercepts and two different slopes. That they have two different slopes can be seen from the fact that the lines are not parallel. One has a slope of 450 words per year and the other has a slope of 550 words per year. This difference in slope of 100 is exactly the size of the slope coefficient pertaining to the product `SES` $\times$ `age`, b_3. Thus, the interpretation of the regression coefficient for a product of two variables is that it represents *the difference in slope*.

The observation that the slope coefficient is different for different groups is called an *interaction effect*, or *interaction* for short. Other words for this phenomenon are *modification* and *moderation*. In this case, `SES` is called the *modifier variable*: it modifies the relationship between `age` on vocabulary. Note however that you could also interpret `age` as the modifier variable: the effect of `SES` is larger for older children than for younger children. In the plot you see that the difference between vocabulary for high and low SES children of age 6 is larger than it is for children of age 2.

9.2 Interaction effect with a dummy variable in R

Let's look at some example output for an R data set where we have a categorical variable that is not dummy-coded yet. The data set is on chicks and their weight during the first days of their lives. Weight is measured in grams. The chicks were given one of four different diets. Here we use only the data from chicks on two different diets 1 and 2. We select only the Diet 1 and 2 data. We store the

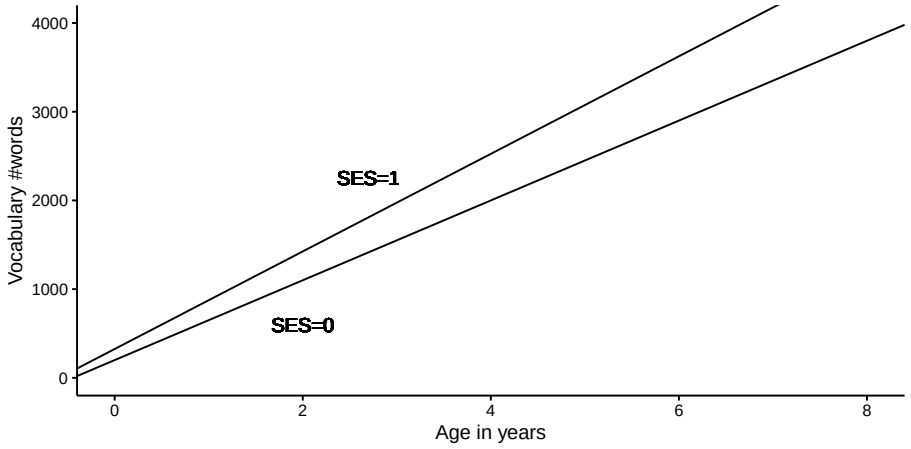


Figure 9.2: Two regression lines for the relationship between `age` and `vocab`, one for low SES children (`SES = 0`) and one for high SES children (`SES = 1`).

Diet 1 and 2 data under the name `chick_data`. When we have a quick look at the data with `glimpse()`, we see that `Diet` is a factor (`<fct>`).

```
chick_data <- ChickWeight %>%
  filter(Diet == 1 | Diet == 2)

chick_data %>%
    glimpse()

## Rows: 340
## Columns: 4
## $ weight <dbl> 42, 51, 59, 64, 76, 93, 106, 125, 149, 171, 199, 205, 40, 49...
## $ Time   <dbl> 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 21, 0, 2, 4, 6, 8, 10...
## $ Chick  <ord> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, ...
## $ Diet   <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
```

The general regression of `weight` on `Time` is shown in Figure 9.3. This regression line for the entire sample of chicks has a slope of around 8 grams per day. Now we want to know whether this slope is the same for chicks in the Diet 1 and Diet 2 groups, in other words, do chicks grow as fast with Diet 1 as with Diet 2? We might expect that `Diet` *moderates* the effect of `Time` on `weight`. We use the following code to study this `Diet` × `Time` interaction effect, by having R automatically create a dummy variable for the factor `Diet`. In the model we specify that we want a main effect of `Time`, a main effect of `Diet`, and an interaction effect of `Time` by `Diet`:

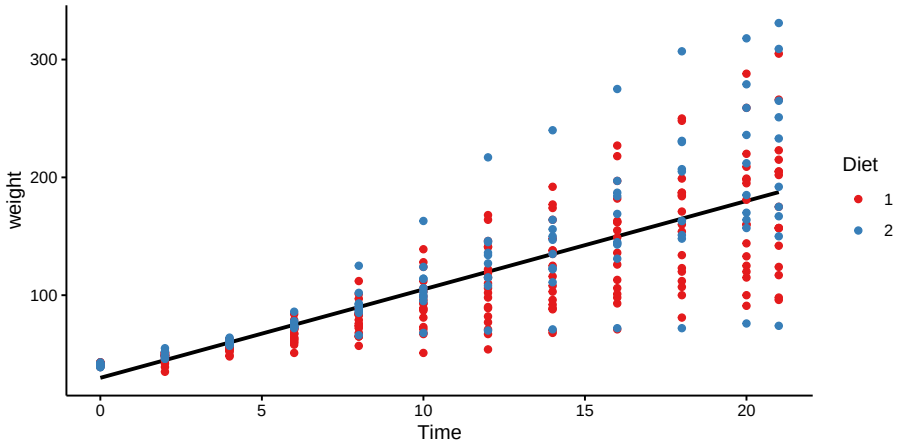


Figure 9.3: The relationship between `Time` and `weight` in all chicks with either Diet 1 or Diet 2.

```
out <- chick_data %>%
  lm(weight ~ Time + Diet + Time:Diet, data = .)
out %>%
  tidy(conf.int = 0.95)

## # A tibble: 4 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)    30.9      4.50      6.88  2.95e-11   22.1       39.8
## 2 Time            6.84     0.361    19.0   4.89e-55    6.13       7.55
## 3 Diet2          -2.30     7.69     -0.299 7.65e- 1  -17.4       12.8
## 4 Time:Diet2      1.77     0.605     2.92  3.73e- 3    0.577      2.96
```

In the regression table, we see the effect of the numeric `Time` variable, which has a slope of 6.84. For every increase of 1 in `Time`, there is a corresponding expected increase of 6.84 grams in weight. Next, we see that R created a dummy variable `Diet2`. That means this dummy codes 1 for Diet 2 and 0 for Diet 1. From the output we see that if a chick gets Diet 2, its weight is -2.3 grams heavier (that means, Diet 2 results in a lower weight).

Next, R created a dummy variable `Time:Diet2`, by multiplying the variables `Time` and `Diet2`. Results show that this interaction effect is 1.77.

These results can be plugged into the following regression equation:

$$\widehat{\texttt{weight}} = 30.93 + 6.84 \times \texttt{Time} - 2.3 \times \texttt{Diet2} + 1.77 \times \texttt{Time} \times \texttt{Diet2} \quad (9.18)$$

If we fill in 1s for the `Diet2` dummy variable, we get the equation for chicks with Diet 2:

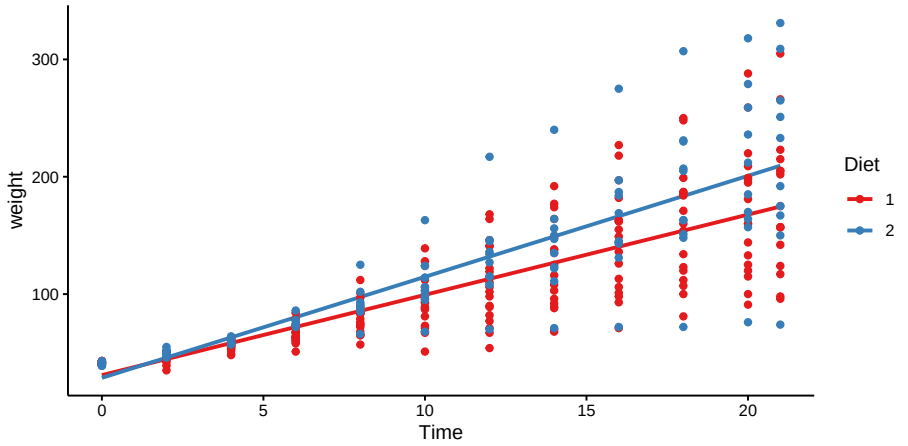


Figure 9.4: The relationship between `Time` and `weight` in chicks, separately for Diet 1 and Diet 2.

$$\begin{aligned} \widehat{\texttt{weight}} &= 30.93 + 6.84 \times \texttt{Time} - 2.3 \times 1 + 1.77 \times \texttt{Time} \times 1 \\ &= 28.63 + 8.61 \times \texttt{Time} \end{aligned} \tag{9.19}$$

If we fill in 0s for the `Diet2` dummy variable, we get the equation for chicks with Diet 1:

$$\widehat{\texttt{weight}} = 30.93 + 6.84 \times \texttt{Time} \tag{9.20}$$

When comparing these two regression lines for chicks with Diet 1 and Diet 2, we see that the slope for `Time` is 1.77 steeper for Diet 2 chicks than for Diet 1 chicks. In this particular random sample of chicks, the chicks on Diet 1 grow 6.84 grams per day (on average), but chicks on Diet 2 grow $6.84 + 1.77 = 8.61$ grams per day (on average).

We visualised these results in Figure 9.4. There we see two regression lines: one for the red data points (chicks on Diet 1) and one for the blue data points (chicks on Diet 2). These two regression lines are the same as those regression lines we found when filling in either 1s and 0s in the general linear model. Note that the lines are not parallel, like in Chapter 6. Each regression line is the least squares regression line for the subsample of chicks on a particular diet.

We see that the difference in slope is 1.77 grams per day. This is what we observe in *this* particular sample of chicks. However, what does that tell us about the difference in slope for chicks in general, that is, the population of all chicks? For that, we need to look at the confidence interval. In the regression table above, we also see the 95% confidence intervals for all model parameters. The 95% confidence interval for the `Time` $\times$ Diet2 interaction effect is (0.58,

2.96). That means that plausible values for this interaction effect are those values between 0.58 and 2.96.

It is also possible to do null-hypothesis testing for interaction effects. One could test whether this difference of 1.77 is possible *if the value in the entire population of chicks equals 0*? In other words, is the value of 1.77 significantly different from 0?

The null-hypothesis is

$$H_0 : \beta_{\texttt{Time}\times\texttt{Diet2}} = 0 \tag{9.21}$$

The regression table shows that the null-hypothesis for the interaction effect has a t-value of $t = 2.92$, with a p-value of $3.73 \times 10^{-3} = 0.00373$. For research reports one always also reports the degrees of freedom for a statistical test. The (residual) degrees of freedom can be found in R by typing

```
out$df.residual

## [1] 336
```

We can report that

> "we reject the null-hypothesis and conclude that there is evidence that the `Time` × `Diet2` interaction effect is not 0, $t(336) = 2.92, p = 0.004$."

Summarising, in this section, we established that `Diet` moderates the effect of `Time` on `weight`: we found a significant diet by time interaction effect. The difference in growth rate is 1.77 grams per day, with a 95% confidence interval from 0.58 to 2.96. In more natural English: diet has an effect on the growth rate in chicks.

In this section we discussed the situation that regression slopes might be different in two groups: the regression slope might be steeper in one group than in the other group. So suppose that we had a numerical predictor X for a numerical dependent variable Y, we said that a particular dummy variable Z *moderated* the effect of X on Y. This moderation was quantified by an *interaction* effect. So suppose we have the following linear equation:

$$Y = b_0 + b_1 \times X + b_2 \times Z + b_3 \times X \times Z + e$$

Then, we call b_0 the intercept, b_1 the main effect of X, b_2 the main effect of Z, and b_3 the interaction effect of X and Z (alternatively, the X by Z interaction effect).

9.3 Interaction effects with a categorical variable in R

In the previous section, we looked at the difference in slopes between two groups. But what we can do for two groups, we can do for multiple groups. The data set on chicks contains data on chicks with 4 different diets. When we perform the same analysis using all data in `ChickWeight`, we obtain the regression table

```
out <- ChickWeight %>%
  lm(weight ~ Time + Diet + Time:Diet, data = .)
out %>%
  tidy(conf.int = 0.95)

## # A tibble: 8 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)   30.9       4.25     7.28   1.09e-12   22.6       39.3
## 2 Time           6.84      0.341   20.1    3.31e-68    6.17       7.51
## 3 Diet2         -2.30      7.27    -0.316  7.52e- 1  -16.6       12.0
## 4 Diet3        -12.7       7.27    -1.74   8.15e- 2  -27.0        1.59
## 5 Diet4         -0.139     7.29    -0.0191 9.85e- 1  -14.5       14.2
## 6 Time:Diet2     1.77      0.572    3.09   2.09e- 3    0.645      2.89
## 7 Time:Diet3     4.58      0.572    8.01   6.33e-15    3.46       5.70
## 8 Time:Diet4     2.87      0.578    4.97   8.92e- 7    1.74       4.01
```

The regression table for four diets is substantially larger than for two diets. It contains one slope parameter for the numeric variable `Time`, three different slopes for the factor variable `Diet` and three different interaction effects for the `Time` by `Diet` interaction.

The full linear model equation is

$$\begin{aligned}\widehat{\texttt{weight}} = 30.93 + 6.84 \times \texttt{Time} - 2.3 \times \texttt{Diet2} - 12.68 \times \texttt{Diet3} - 0.14 \times \texttt{Diet4} \\ +1.77 \times \texttt{Time} \times \texttt{Diet2} + 4.58 \times \texttt{Time} \times \texttt{Diet3} + 2.87 \times \texttt{Time} \times \texttt{Diet4}\end{aligned} \tag{9.22}$$

You see that R created dummy variables for Diet 2, Diet 3 and Diet 4. We can use this equation to construct a separate linear model for the Diet 1 data. Chicks with Diet 1 have 0s for the dummy variables `Diet2`, `Diet3` and `Diet4`. If we fill in these 0s, we obtain

$$\widehat{\texttt{weight}} = 30.93 + 6.84 \times \texttt{Time} \tag{9.23}$$

For the chicks on Diet 2, we have 1s for the dummy variable `Diet2` and 0s for the other dummy variables. Hence we have

$$\begin{aligned}\widehat{\texttt{weight}} &= 30.93 + 6.84 \times \texttt{Time} - 2.3 \times 1 + 1.77 \times \texttt{Time} \times 1 \\ &= 30.93 + 6.84 \times \texttt{Time} - 2.3 + 1.77 \times \texttt{Time} \\ &= (30.93 - 2.3) + (6.84 + 1.77) \times \texttt{Time} \\ &= 28.63 + 8.61 \times \texttt{Time} \end{aligned} \tag{9.24}$$

Here we see exactly the same equation for Diet 2 as in the previous section where we only analysed two diet groups. The difference between the two slopes in the Diet 1 and Diet 2 groups is again 1.77. The only difference for this interaction effect is the standard error, and therefore the confidence interval is also slightly different. We will come back to this issue in Chapter **??**.

For the chicks on Diet 3, we have 1s for the dummy variable `Diet3` and 0s for the other dummy variables. The regression equation is then

$$\begin{aligned}\widehat{\texttt{weight}} &= 30.93 + 6.84 \times \texttt{Time} - 12.68 \times 1 + 4.58 \times \texttt{Time} \times 1 \\ &= (30.93 - 12.68) + (6.84 + 4.58) \times \texttt{Time} \\ &= 18.25 + 11.42 \times \texttt{Time} \end{aligned} \tag{9.25}$$

We see that the intercept is again different than for the Diet 1 chicks. We also see that the slope is different: it is now 4.58 steeper than for the Diet 1 chicks. This difference in slopes is exactly equal to the `Time` by `Diet3` interaction effect. This is also what we saw in the Diet 2 group. Therefore, we can say that an interaction effect for a specific diet group says something about how much steeper the slope is in that group, compared to the reference group. The reference group is the group for which all the dummy variables are 0. Here, that is the Diet 1 group.

Based on that knowledge, we can expect that the slope in the Diet 4 group is equal to the slope in the reference group (6.84) plus the `Time` by `Diet4` interaction effect, 2.87, so 9.71.

We can do the same for the intercept in the Diet 4 group. The intercept is equal to the intercept in the reference group (30.93) plus the main effect of the `Diet4` dummy variable, -0.14, which is 30.79.

The linear equation is then for the Diet 4 chicks:

$$\widehat{\texttt{weight}} = 30.79 + 9.71 \times \texttt{Time} \tag{9.26}$$

The four regression lines are displayed in Figure 9.5. The steepest regression line is the one for the Diet 3 chicks: they are the fastest growing chicks. The slowest growing chicks are those on Diet 1. The confidence intervals in the regression table tell us that the difference between the growth rate with Diet 4 compared to Diet 1 is somewhere between 1.74 and 4.01 grams per day.

Suppose we want to test the null hypothesis that all four slopes are the same. This implies that the `Time` by `Diet` interaction effects are all equal to 0. We

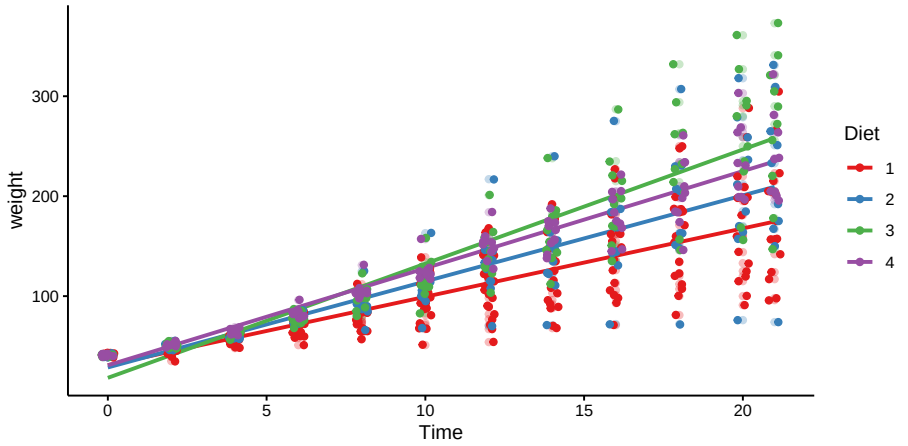


Figure 9.5: Four different regression lines for the four different diet groups.

can test this null hypothesis

$$H_0 : \beta_{\texttt{Time}\times\texttt{Diet2}} = \beta_{\texttt{Time}\times\texttt{Diet2}} = \beta_{\texttt{Time}\times\texttt{Diet4}} = 0 \tag{9.27}$$

by running an ANOVA. That is, we apply the `anova()` function to the results of an `lm()` analysis:

```
out_anova <- out %>%
  anova()
out_anova %>%
  tidy()

## # A tibble: 4 x 6
##   term         df    sumsq   meansq statistic    p.value
##   <chr>     <int>    <dbl>    <dbl>     <dbl>      <dbl>
## 1 Time          1 2042344. 2042344.    1760.   2.11e-176
## 2 Diet          3  129876.   43292.      37.3  5.07e- 22
## 3 Time:Diet     3   80804.   26935.      23.2  3.47e- 14
## 4 Residuals   570  661532.    1161.      NA    NA
```

In the output we see a `Time` by `Diet` interaction effect with 3 degrees of freedom. That term refers to the null-hypothesis that all three interaction effects are equal to 0. The F-statistic associated with that null-hypothesis equals 23.2. The residual degrees of freedom equals 570, so that we can report:

> "The slopes for the four different diets were significantly different from each other, $F(3, 570) = 23.2, MSE = 26935, p < 0.001$."

9.4 Interaction between two dichotomous variables in R

In the previous section we discussed the situation that regression slopes might be different in two four groups. In Chapter 6 we learned that we could also look at slopes for dummy variables. The slope is then equal to the difference in group means, that is, the slope is the increase in the group mean of one group compared to the reference group.

Now we discuss the situation where we have two dummy variables, and want to do inference on their interaction. Does one dummy variable moderate the effect of the other dummy variable?

Let's have a look at a data set on penguins. It can be found in the `palmerpenguins` package.

```
# install.packages("palmerpenguins")
library(palmerpenguins)
penguins %>%
  str ()

## tibble [344 x 8] (S3: tbl_df/tbl/data.frame)
##  $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1
##  $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 .
##  $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
##  $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
##  $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
##  $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 .
##  $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ..
##  $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007
```

We see there is a `species` factor with three levels, and a `sex` factor with two levels. Let's select only the Adelie and Chinstrap species.

```
penguin_data <- penguins %>%
  filter(species %in% c("Adelie", "Chinstrap"))
```

Suppose that we are interested in differences in flipper length across species. We then could run a linear model, with `flipper_length_mm` as the dependent variable, and `species` as independent variable.

```
out <- penguin_data %>%
  lm(flipper_length_mm ~ species, data = .)
out %>%
  tidy(conf.int = 0.95)

## # A tibble: 2 x 7
##   term            estimate std.error statistic   p.value conf.low conf.high
```

```
##   <chr>               <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 (Intercept)        190.       0.548    347.   8.31e-300   189.      191.
## 2 speciesChinstrap     5.87     0.983      5.97 9.38e-  9     3.93      7.81
```

The output shows that in this sample, the Chinstrap penguins have on average larger flippers than Adelie penguins. The confidence intervals tell us that this difference in flipper length is somewhere between 3.93 and 7.81. But suppose that this is not what we want to know. The real question might be whether this difference is different for male and female penguins. Maybe there is a larger difference in flipper length in females than in males?

This difference or change in one variable (`flipper_length_mm`) as a function of another variable (`sex`) should remind us of *moderation*: maybe sex moderates the effect of species on flipper length.

In order to study such moderation, we have to analyse the `sex` by `species` interaction effect. By now you should know how to do that in R:

```
out <- penguin_data %>%
  lm(flipper_length_mm ~ species + sex + species:sex, data = .)
out %>%
  tidy(conf.int = 0.95)

## # A tibble: 4 x 7
##   term                estimate std.error statistic   p.value conf.low conf.high
##   <chr>                  <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 (Intercept)            188.       0.707    266.   2.15e-267 186.        189.
## 2 speciesChinstrap         3.94     1.25       3.14 1.92e-  3   1.47        6.41
## 3 sexmale                  4.62     1.00       4.62 6.76e-  6   2.65        6.59
## 4 speciesChinstrap:se~     3.56     1.77       2.01 4.60e-  2   0.0638      7.06
```

In the output we see an intercept of 188. Next, we see an effect of a dummy variable, coding 1s for Chinstrap penguins (`speciesChinstrap`). We also see an effect of a dummy variable coding 1s for male penguins (`sexmale`). Then, we see an interaction effect of these two dummy effects. That means that this dummy variable codes 1s for the specific combination of Chinstrap penguins that are male (`speciesChinstrap:sexmale`).

$$\widehat{\texttt{flipper\_length}} = 188 + 3.94 \times \texttt{speciesChinstrap} + 4.62 \times \texttt{sexmale} + 3.56 \times \texttt{speciesChinstrap} \times \texttt{sexmale} \quad (9.28)$$

From this we can make the following predictions. The predicted flipper length for female Adelie penguins is

$$188 + 3.94 \times 0 + 4.62 \times 0 + 3.56 \times 0 \times 0 = 188 \quad (9.29)$$

The predicted flipper length for male Adelie penguins is

$$\begin{aligned} &188 + 3.94 \times 0 + 4.62 \times 1 + 3.56 \times 0 \times 1 \\ &= 188 + 4.62 = 192.62 \end{aligned} \tag{9.30}$$

The predicted flipper length for female Chinstrap penguins is

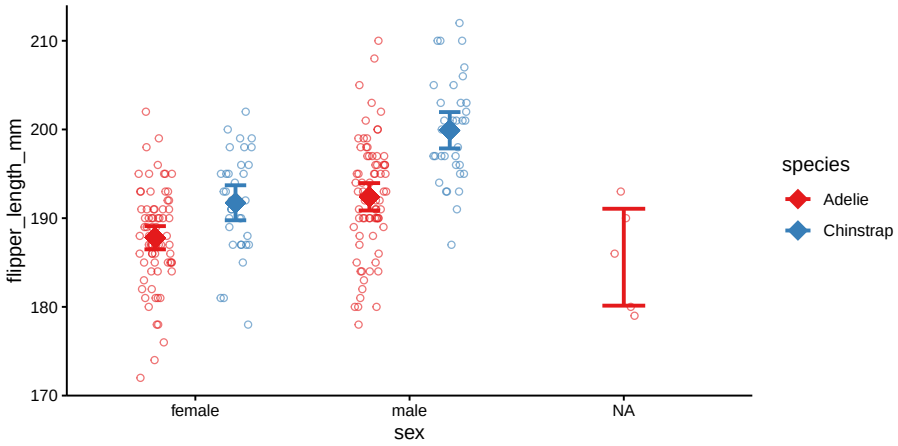
$$\begin{aligned} &188 + 3.94 \times 1 + 4.62 \times 0 + 3.56 \times 0 \times 0 \\ &= 188 + 3.94 = 191.94 \end{aligned} \tag{9.31}$$

and the predicted flipper length for male Chinstrap penguins is

$$\begin{aligned} &188 + 3.94 \times 1 + 4.62 \times 1 + 3.56 \times 1 \times 1 \\ &= 188 + 3.94 + 4.62 + 3.56 \\ &= 200.12 \end{aligned} \tag{9.32}$$

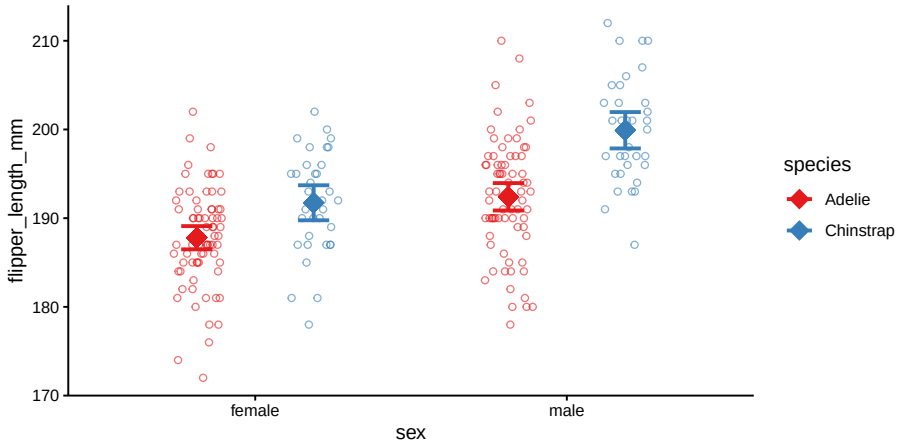
These predicted flipper length for each male/species combination are actually the group means. It is generally best to plot these means with a *means and errors plot*. For that we first need to compute means by R. With `left_join()` we add these means to the data set. These diamant-shaped means (`shape = 18`) are plotted with intervals that are twice (`mult = 2`) the standard error of those means (`geom = "errorbar"`).

```
left_join(penguin_data, # adding group means to the data set
          penguin_data %>%
            group_by(species, sex) %>%
            summarise(mean = mean(flipper_length_mm))
          ) %>%
  ggplot(aes(x = sex, y = flipper_length_mm, colour = species)) +
  geom_jitter(position = position_jitterdodge(), # the raw data
              shape = 1,
              alpha = 0.6) +
  geom_point(aes(y = mean),    # the groups means
             position = position_jitterdodge(jitter.width = 0),
             shape = 18,
             size = 5) +
  stat_summary(fun.data = mean_se,  # computing errorbars
               fun.args = list(mult = 2),
               geom = "errorbar",
               width = 0.2,
               position = position_jitterdodge(jitter.width = 0),
               size = 1) +
  scale_color_brewer(palette = "Set1")
```



This plot shows also the data on penguins with unknown sex (`sex = NA`). If we leave these out, we get

```
left_join(penguin_data, # adding group means to the data set
          penguin_data %>%
            group_by(species, sex) %>%
            summarise(mean = mean(flipper_length_mm))
          ) %>%
  filter(!is.na(sex)) %>%
  ggplot(aes(x = sex, y = flipper_length_mm, colour = species)) +
  geom_jitter(position = position_jitterdodge(), # the raw data
              shape = 1,
              alpha = 0.6) +
  geom_point(aes(y = mean),     # the groups means
             position = position_jitterdodge(jitter.width = 0),
             shape = 18,
             size = 5) +
  stat_summary(fun.data = mean_se,  # computing errorbars
               fun.args = list(mult = 2),
               geom = "errorbar",
               width = 0.2,
               position = position_jitterdodge(jitter.width = 0),
               size = 1) +
  scale_color_brewer(palette = "Set1")
```



Comparing the Adelie and the Chinstrap data, we see that for both males and females, the Adelie penguins have smaller flippers than the Chinstrap penguins. Comparing males and females, we see that the males have generally larger flippers than females. More interestingly in relation to this chapter, the means in the females are farther apart than the means in the males. Thus, in females the effect of species is larger than in males. This is the interaction effect, and this difference in the difference in means is equal to 3.56 in this data set. With a confidence level of 95% we can say that this difference in the effect of species is probably somewhere between 0.06 and 7.06 mm in the population of all penguins.

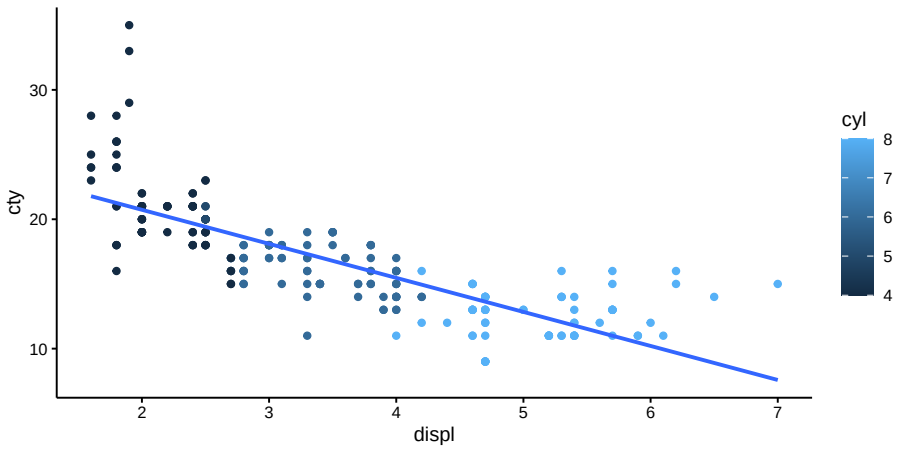
9.5 Moderation involving two numeric variables in R

In all previous examples, we saw at least one categorical variable. We saw that for different levels of a dummy variable, the slope of another variable varied. We also saw that for different levels of a dummy variable, the effect of another dummy variable varied. In this section, we look at how the slope of a numeric variable can vary, as a function of the level of another numeric variable.

As an example data set, we look at the `mpg` data frame, available in the `ggplot2` package. It contains data on 234 cars. Let's analyse the dependent variable `cty` (city miles per gallon) as a function of the numeric variables `cyl` (number of cylinders) and `displ` (engine displacement). First we plot the relationship between engine displacement and city miles per gallon. We use colours, based on the number of cylinders. We see that there is in general a negative slope: the higher the displacement value, the lower the city miles per gallon.

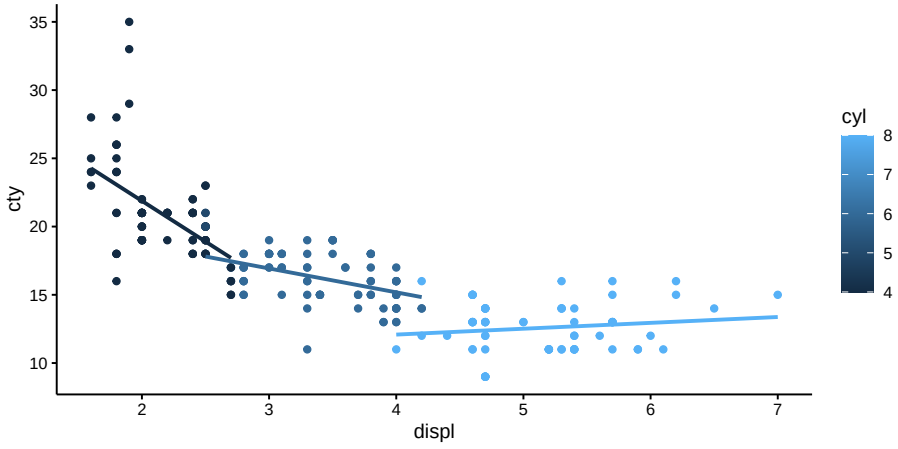
```
mpg %>%
  ggplot(aes(x = displ, y = cty)) +
  geom_point(aes(colour = cyl)) +
```

```
  geom_smooth(method = "lm", se = F)
```



When we run separate linear models for the different number of cylinders, we get

```
mpg %>%
  ggplot(aes(x = displ, y = cty, colour = cyl, group = cyl)) +
  geom_point() +
  geom_smooth(method = "lm", se = F)
```



We see that the slope is different, depending on the number of cylinders: the more cylinders, the less negative is the slope: very negative for cars with low number of cylinders, and slightly positive for cars with high number of cilinders. In other words, the slope increases in value with increasing number of cylinders. If we want to quantify this interaction effect, we need to run a linear model with an interaction effect.

```
out <- mpg %>%
  lm(cty ~ displ + cyl + displ:cyl, data = .)
out %>%
  tidy()

## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   38.6       2.03      19.0  3.69e-49
## 2 displ         -5.29      0.825     -6.40 8.45e-10
## 3 cyl           -2.70      0.376     -7.20 8.73e-12
## 4 displ:cyl      0.559     0.104      5.38 1.85e- 7
```

We see that the `displ` by `cyl` interaction effect is 0.559. It means that the slope of `displ` changes by 0.559 for every unit increase in `cyl`.

For example, when we look at the predicted city miles per gallon with `cyl = 2`, we get the following model equation:

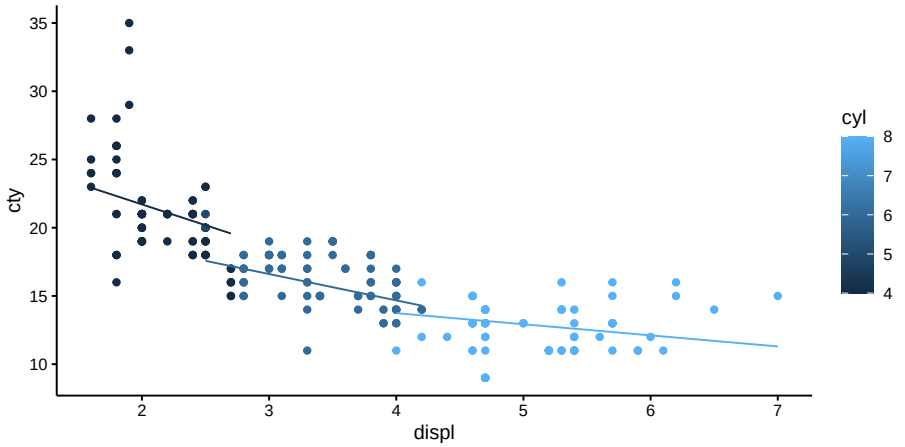
$$\begin{aligned}
\widehat{\texttt{cty}} &= 0.6 - 5.285 \times \texttt{displ} - 2.704\texttt{cyl} + 0.559 \times \texttt{displ} \times \texttt{cyl} \\
\widehat{\texttt{cty}} &= 0.6 - 5.285 \times \texttt{displ} - 2.704 \times 2 + 0.559 \times \texttt{displ} \times 2 \\
\widehat{\texttt{cty}} &= 0.6 - 5.285 \times \texttt{displ} - 5.408 + 1.118 \times \texttt{displ} \\
\widehat{\texttt{cty}} &= (0.6 - 5.408) + (1.118 - 5.285) \times \texttt{displ} \\
\widehat{\texttt{cty}} &= -4.808 - 4.167 \times \texttt{displ} \qquad (9.33)
\end{aligned}$$

If we increase the number of cylinders from 2 to 3, we obtain the equation:

$$\begin{aligned}
\widehat{\texttt{cty}} &= 0.6 - 5.285 \times \texttt{displ} - 2.704 \times 3 + 0.559 \times \texttt{displ} \times 3 \\
\widehat{\texttt{cty}} &= -7.512 - 3.608 \times \texttt{displ} \qquad (9.34)
\end{aligned}$$

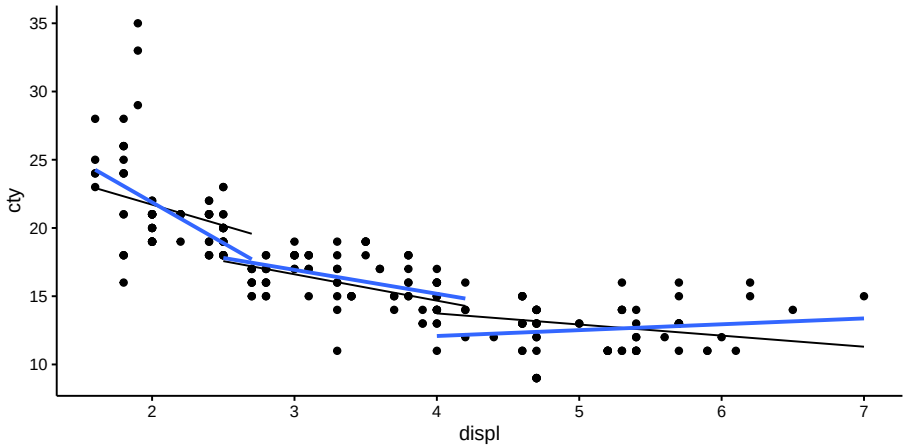
We see a different intercept and a different slope. The difference in the slope between 3 and 2 cylinders equals 0.559, which is exactly the interaction effect. If you do the same exercise with 4 and 5 cylinders, or 6 and 7 cylinders, you will always see this difference again. This parameter for the interaction effect just says that the best prediction for the change in slope when increasing the number of cylinders with 1, is 0.559. We can plot the predictions from this model in the following way:

```
library(modelr)
mpg %>%
  add_predictions(out) %>%
  ggplot(aes(x = displ, y = cty, colour = cyl)) +
  geom_point() +
  geom_line(aes(y = pred, group = cyl))
```



If we compare these predicted regression lines with those in the previous figure

```
mpg %>%
  add_predictions(out) %>%
  ggplot(aes(x = displ, y = cty, group = cyl)) +
  geom_point() +
  geom_line(aes(y = pred), colour = "black") +
  geom_smooth(method = "lm", se = F)
```



we see that they are a little bit different. That is because in the model we treat `cyl` as numeric: for every increase of 1 in `cyl`, the slope changes by a fixed amount. When you treat `cyl` as categorical, then you estimate the slope separately for all different levels. You would then see multiple parameters for the interaction effect:

```
out <- mpg %>%
  mutate(cyl = factor(cyl)) %>%
  lm(cty ~ displ + cyl + displ:cyl, data = .)
out %>%
  tidy()

## # A tibble: 8 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    33.8      1.70      19.9   1.50e-51
## 2 displ          -5.96     0.785     -7.60  7.94e-13
## 3 cyl5            1.60     1.17       1.37  1.72e- 1
## 4 cyl6          -11.6      2.50      -4.65  5.52e- 6
## 5 cyl8          -23.4      2.89      -8.11  3.12e-14
## 6 displ:cyl5     NA       NA         NA    NA
## 7 displ:cyl6      4.21     0.948      4.44  1.38e- 5
## 8 displ:cyl8      6.39     0.906      7.06  2.05e-11
```

When `cyl` is turned into a factor, you see that cars with 4 cylinders are taken as the reference category, and there are effects of having 5, 6, or 8 cylinders. We see the same for the interaction effects: there is a reference category with 4 cylinders, where the slope of `displ` equals -5.96. Cars with 6 and 8 cylinders have different slopes: the one for 6 cylinders is 5.96 + 4.21 and the one for 8 cylinders is 5.96 + 6.39. The slope for cars with 5 cylinders can't be separately estimated because there is no variation in `displ` in the `mpg` data set.

You see that you get different results, depending on whether you treat a variable as numeric or as categorical. Treated as numeric, you end up with a simpeler model with fewer parameters, and therefore a larger number of degrees of freedom. What to choose depends on the research question and the amount of data. In general, a model should be not too complex when you have relatively few data points. Whether the model is appropriate for your data can be checked by looking at the residuals and checking the assumptions.

Chapter 10

Generalised linear models: logistic regression

10.1 Introduction

In previous chapters we were introduced to the linear model, with its basic form

$$\begin{aligned} Y &= b_0 + b_1 X_1 + \cdots + b_n X_n + e && (10.1) \\ e &\sim N(0, \sigma_e^2) && (10.2) \end{aligned}$$

Two basic assumptions of this model are the additivity in the parameters, and the normally distributed residual e. Additivity in the parameters means that the effects of intercept and the independent variables $X_1, X_2, \ldots X_n$ are additive: the assumption is that you can sum these effects to come to a predicted value for Y. So that is also true when we include an interaction effect to account for a moderation,

$$\begin{aligned} Y &= b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2 + e && (10.3) \\ e &\sim N(0, \sigma_e^2) && (10.4) \end{aligned}$$

or when we use a quadratic term to account for another type of non-linearity in the data:

$$\begin{aligned} Y &= b_0 + b_1 X_1 + b_2 X_1 X_1 + e && (10.5) \\ e &\sim N(0, \sigma_e^2) && (10.6) \end{aligned}$$

In all these models, the assumption is that the effects of the parameters (b_0, b_1, b_2) can be summed.

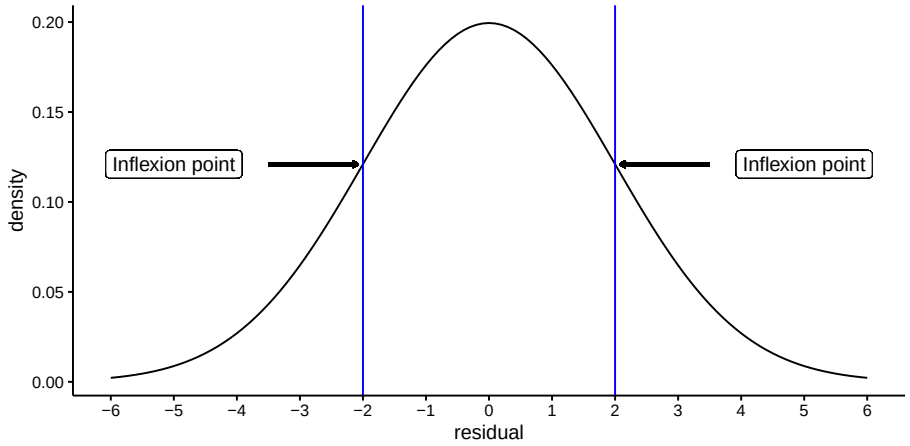


Figure 10.1: Density function of the normal distribution, with mean 0 and variance 4 (standard deviation 2). Inflexion points are positioned at residual values of minus 1 standard deviation and plus 1 standard deviation.

The other major assumption of linear (mixed) models is the normal distribution of the residuals. As we have seen in for instance Chapter 7, sometimes the residuals are not normally distributed. Remember that with a normal distribution $N(0, \sigma^2)$, in principle all values between $-\infty$ and $+\infty$ are possible, but they tend to concentrate around the value of 0, in the shape of the bell-curve. Figure 10.1 shows the normal distribution $N(0, \sigma^2 = 4)$: it is centred around 0 and has variance 4. Note that the inflexion point, that is the point where the decrease in density tends to decelerate, is exactly at the values -2 and +2. These are equal to the square root of the variance, which is the standard deviation, $-\sigma$ and $+\sigma$.

A normal distribution is suitable for continuous dependent variables. For most measured variables this is not true. Think for example of temperature measures: if the thermometer gives degrees Celsius with a precision of only 1 decimal, we can never have values of say 10.07 or -56.789. Our actual data will in fact be *discrete*, showing rounded values like 10.1, 10.2, 10.3, but never any values in between.

Nevertheless, the normal distribution can still be used in many such cases. Take for instance a data set where the temperature in Amsterdam in summer was predicted on the basis of a linear model. Fig 10.2 shows the distribution of the residuals for that model. The temperature measures were discrete with a precision of one tenth of a degree Celsius, but the distribution seems well approximated by a normal curve.

But let's look at an example where the discreteness is more prominent. In Figure 10.3 we see the residuals of an analysis of exam results. Students had to do an assignment that had to meet 4 criteria: 1) originality, 2) language, 3)

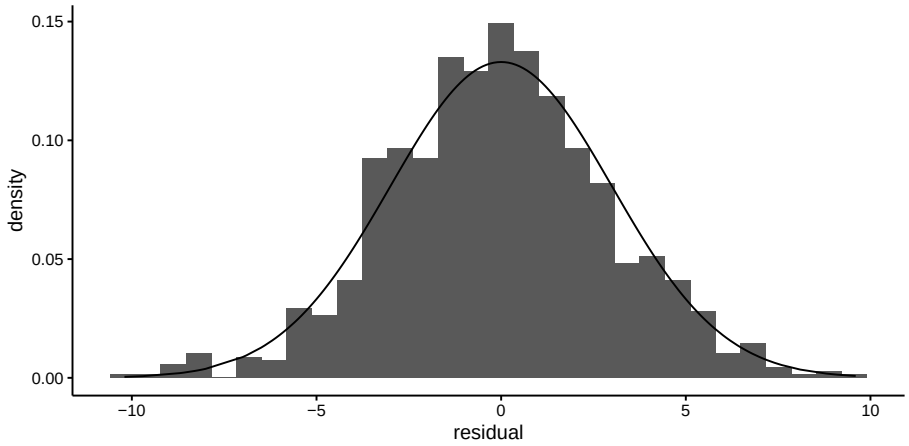


Figure 10.2: Even if residuals are really discrete, the normal distribution can be a good approximation of their distribution.

structure, and 4) literature review. Each criterion was scored as either fulfilled (1) or not fulfilled (0). The score for the assignment was determined on the basis of *the number of criteria* that were met, so the scores could be 0, 1, 2, 3 or 4. In an analysis, this score was predicted on the basis of the average exam score on previous assignments, using a linear model.

Figure 10.3 shows that the residuals are very discrete, and that the continuous normal distribution is a very bad approximation of the histogram. We often see this phenomenon when our data consist of *counts* with a limited maximum number.

An even more extreme case we observe when our dependent variable consists of whether or not students passed the assignment: only those assignments that fulfilled all 4 criteria are regarded as sufficient. If we score all students with a sufficient assignment as passed (scored as a value of 1) and all students with an insufficient assignment as failed (scored as a value of 0) and we predict this score by the average exam score on previous assignments using a linear model, we get the residuals displayed in Figure 10.4.

Here it is also evident that a normal approximation of the residuals will not do. When the dependent variable has only 2 possible values, a linear model will never work because the residuals can never have a distribution that is even remotely looking normal.

In this chapter and the next we will discuss how generalised linear models can be used to analyse data sets where the assumption of normally distributed residuals is not tenable. First we discuss the case where the dependent variable has only 2 possible values (dichotomous dependent variables like yes/no or pass/fail, heads/tails, 1/0). In Chapter **??**, we will discuss the case where the dependent variable consists of counts $(0, 1, 2, 3, 4, \dots)$.

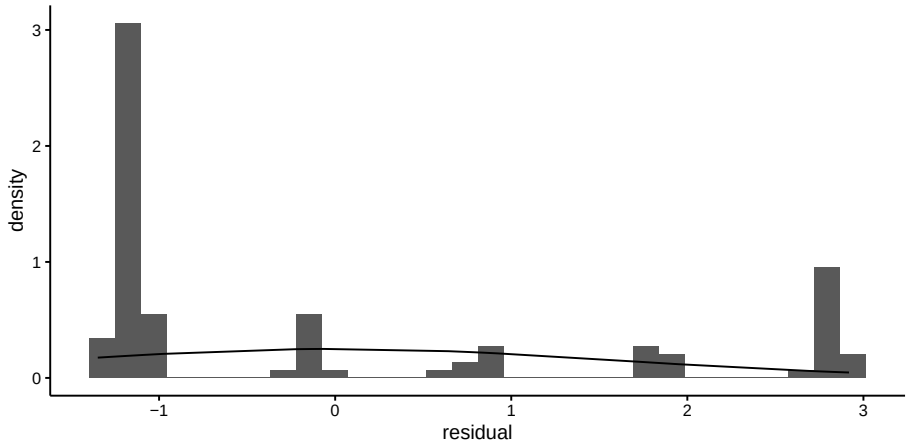


Figure 10.3: Count data example where the normal distribution is not a good approximation of the distribution of the residuals.

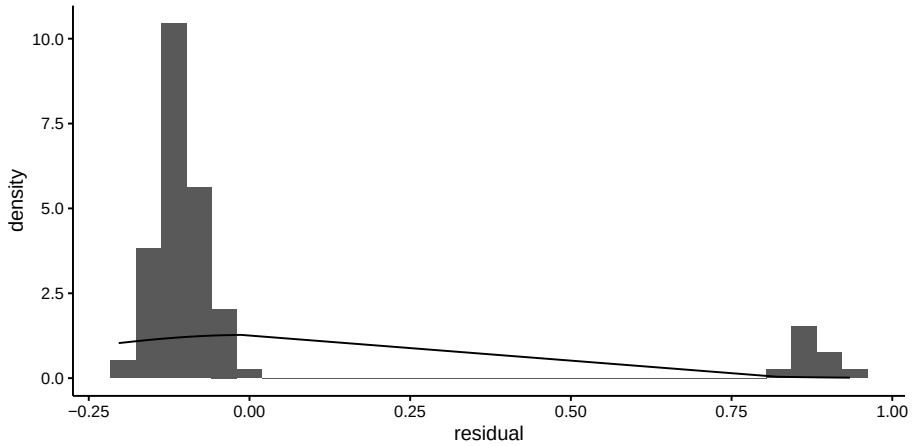


Figure 10.4: Dichotomous data example where the normal distribution is not a good approximation of the distribution of the residuals.

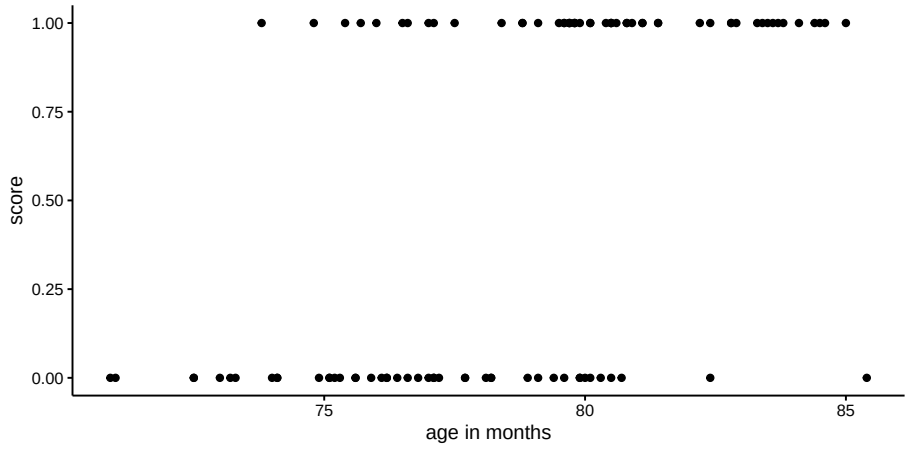


Figure 10.5: Data example: Exam outcome (score) as a function of age, where 1 means pass and 0 means fail.

10.2 Logistic regression

Imagine that we analyse results on an exam for third grade children. These children are usually either 6 or 7 years old, depending on what month they were born in. The exam is on February 1st. A researcher wants to know whether the age of the child can explain why some children pass the test and others fail. She computes the age of the child in months. Each child that passes the exam gets a score of 1 and all the others get a score of 0. Figure 10.5 plots the data.

She wants to use the following linear model:

$$\begin{aligned} \texttt{score} &= b_0 + b_1 \texttt{age} + e && (10.7) \\ e &\sim N(0, \sigma_e^2) && (10.8) \end{aligned}$$

Figure 10.6 shows the data with the estimated regression line and Figure 10.7 shows the distribution of the residuals as a function of `age`.

Clearly a linear model is not appropriate. Here, the assumption that the dependent variable, score in this case, is scattered randomly around the predicted value with a normal distribution is not reasonable. The main problem is that the dependent variable score can only have 2 values: 0 and 1. When we have a dependent variable that is categorical, so not continuous, we generally use *logistic regression*. In this chapter we cover the case when the dependent variable takes binary values, like 0 and 1.

10.2.1 Bernoulli distribution

Rather than using a normal distribution, we could try a Bernoulli distribution. The Bernoulli distribution is the distribution of a coin flip. For example, if the

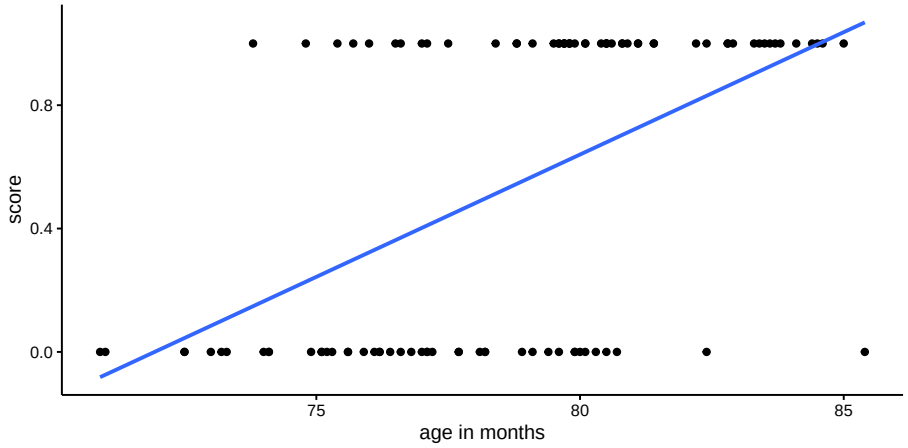


Figure 10.6: Example exam data with a linear regression line.

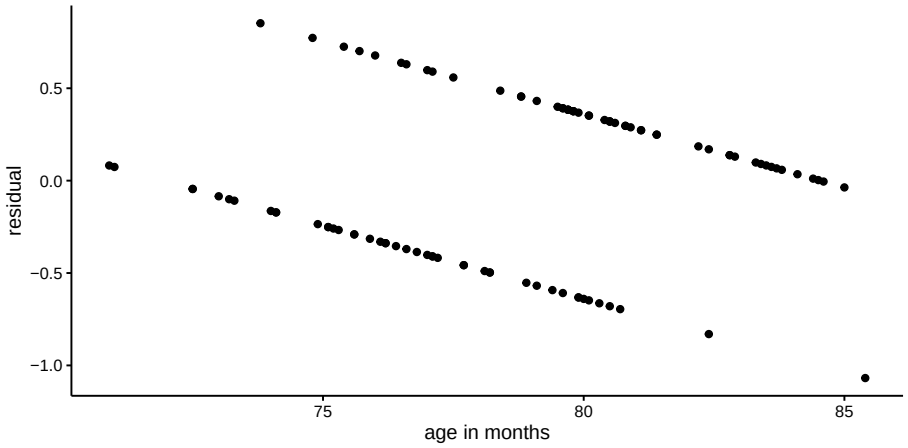


Figure 10.7: Residuals as a function of age, after a linear regression analysis of the exam data.

probability of heads is 0.1, we can expect that if we flip the coin 100 times, on average we expect to see 10 times heads and 90 times tails. Our best bet then is that the outcome is tails. However, if we actually flip the coin, we might see heads anyway. There is some randomness to be expected. Let Y be the outcome of a coin flip: heads or tails. If we have a Bernoulli distribution for variable Y with probability p for heads, we *expect* to see heads p times, but we actually *observe* heads or tails (Y).

$$Y \sim Bern(p) \tag{10.9}$$

The same is true for the normal distribution in the linear model case: we *expect* that the observed value of Y is exactly equal to its predicted value ($b_0 + b_1X$), but we *observe* Y that it is most often different.

$$Y \sim N(\mu = b_0 + b_1X, \sigma_e^2) \tag{10.10}$$

In our example of passing the exam by the third graders, the pass rate could also be conceived as the outcome of a coin flip: pass instead of heads and fail instead of tails. So would it be an idea to predict the *probability* of passing the exam on the basis of age? And then for every predicted probability, we allow for the fact that actually the observed success can differ. Our linear model could then look like this:

$$p_i = b_0 + b_1\texttt{age}_i \tag{10.11}$$

$$\texttt{score}_i \sim Bern(p_i) \tag{10.12}$$

So for each child i, we predict the probability of success, p_i, on the basis of her/his age. Next, the randomness in the data comes from the fact that a probability is only a probability, so that the observed success of a child $\texttt{score}_i$, is like a coin toss with probability of p_i for success.

For example, suppose that we have a child with an age of 80 months, and we have $b_0 = -3.8$ and $b_1 = 0.05$. Then the predicted probability p_i is equal to $-3.8 + 0.05 \times 80 = 0.20$. The best bet for such a child would be that it fails the exam. But, although 0.20 is a small probability, there is a chance that the child passes the exam. This model also means that if we would have 100 children of age 80 months, we would *expect* that 20 of these children would pass the test and 80 would fail. But we can't make exact predictions for one individual alone: we don't know exactly which child will pass and which child won't. Note that this is similar to the normally distributed residual in the linear model: in the linear model we expect a child to have a certain value for Y, but we know that there will be a deviation from this predicted value: the residual. For a whole group of children with the same predicted value for Y, we know that the whole group will show residuals that have a normal distribution. But we're not sure what the residual will be for each individual child.

Unfortunately, this model for probabilities is not very helpful. If we use a linear model for the probability, this means that we can predict probability

values of less than 0 and more than 1, and this is not possible for probabilities. If we use the above values of $b_0 = -3.8$ and $b_1 = 0.05$, we predict a probability of -.3 for a child of 70 months and a probability of 1.2 for a child of 100 months. Those values are meaningless, since probabilities are always between 0 and 1!

10.2.2 Odds and logodds

Instead of predicting probabilities, we could predict *odds*. The nice property of odds is that they can have very large values, much larger than 1.

What are odds again? Odds are a different way of talking about probability. Suppose the probability of winning the lottery is 1%. Then the probability of loosing is 99%. This is equal to saying that the odds of winning against loosing are 1 to 99, or $1 : 99$, because the probability of winning is 99 times smaller than the probability of loosing.

As another example, suppose the probability of being alive tomorrow is equal to 0.9999. Then the probability of not being alive tomorrow is $1 - 0.9999 = 0.0001$. Then the probability of being alive tomorrow is $0.9999/0.0001 = 9999$ times larger than the the probability of not being alive. Therefore the odds of being alive tomorrow against being dead is 9999 to 1 (9999:1).

If we have a slightly biased coin, the probability of heads might be 0.6. The probability of tails is then 0.4. Then the probability of heads is then 1.5 times larger than the probability of heads (0.6/0.4=1.5). So the odds of heads against tails is then 1.5 to 1. For the sake of clarity, odds are often multiplied by a constant to get integers, so we can also say the odds of heads against tails are 3 to 2. Similarly, if the probability of heads were 0.61, the odds of heads against tails would be 0.61 to 0.39, which can be modified into 61 to 39.

Now that we know how to go from probability statements to statements about odds, how do we go from odds to probability? If someone says the odds of heads against tails is 10 to 1, this means that for every 10 heads, there will be 1 tails. In other words, if there were 11 coin tosses, 10 would be heads and 1 would be tails. We can therefore transform odds back to probabilities by noting that 10 out of 11 coin tosses is heads, so $10/11 = 0.91$, and 1 out of 11 is tails, so $1/11 = 0.09$.

If someone says the odds of winning a gold medal at the Olympics is a thousand to one (1000:1), this means that if there were $1000 + 1 = 1001$ opportunities, there would be a gold medal in 1000 cases and failure in only one. This corresponds to a probability of 1000/1001 for winning and 1/1001 for failure.

As a last example, if at the horse races, the odds of Bruno winning against Sacha are four to five (4:5), this means that for every 4 winnings by Bruno, there would be 5 winnings by Sacha. So out of a total of 9 winnings, 4 will be by Bruno and 5 will be by Sacha. The probability of Bruno outrunning Sacha is then $4/9 = 0.44$.

If we would summarise the odds by doing the division, we have just one number. For example, if the odds are 4 to 5 (4:5), the odds are $4/5 = 0.8$, and if the odds

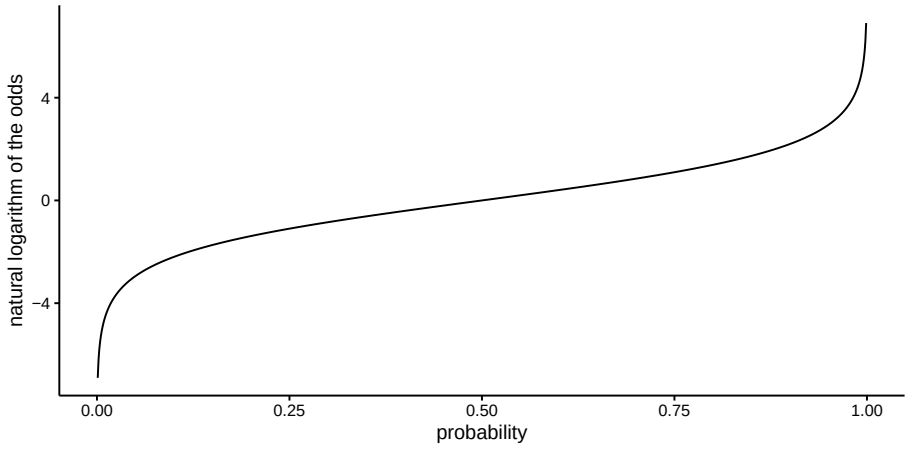


Figure 10.8: The relationship between a probability and the natural logarithm of the corresponding odds.

are a thousand to one (1000:1), then we can also say the odds are 1000. Odds, unlike probabilities, can have values that are larger than 1.

However, note that odds can never be negative: a very small odds is one to a thousand (1:1000). This can be summarised as an odds of 0.000999001, but that is still larger than 0. In summary: probabilities range from 0 to 1, and odds from 0 to infinity.

Because odds can never be negative, mathematicians have proposed to use the *natural logarithm*[1] of the odds as the preferred transformation of probabilities. For example, suppose we have a probability of heads of 0.42. This can be transformed into an odds by noting that in 100 coin tosses, we would expect 42 times heads and 58 times tails. So the odds are 42:58, which is equal to $\frac{42}{58} = 0.7241379$. The *natural* logarithm of 0.7241379 equals -0.3227734 (use the ln button on your calculator!). If we have a value between 0 and 1 and we take the logarithm of that value, we always get a value smaller than 0. In short: a probability is never negative, but the corresponding logarithm of the odds can be negative.

Figure 10.8 shows the relationship between a probability (with values between 0 and 1) and the natural logarithm of the corresponding odds (the *logodds*). The result is a mirrored S-shaped curve on its side. For large probabilities close to one, the equivalent logodds becomes infinitely positive, and for very small probabilities close to zero, the equivalent logodds becomes infinitely

[1]The natural logarithm of a number is its logarithm to the base of the constant e, where e is approximately equal to 2.7. The natural logarithm of x is generally written as lnx or $log^e x$. The natural logarithm of x is the power to which e needs to be raised to equal x. For example, $ln(2)$ is 0.69, because $e^{0.69} = 2$, and $ln(0.2) = -1.6$ because $e^{-1.6} = 0.2$. The natural logarithm of e itself, $ln(e)$, is 1, because $e^1 = e$, while the natural logarithm of 1, $ln(1)$, is 0, since $e^0 = 1$.

negative. A logodds of 0 is equal to a probability of 0.5. If a logodds is larger than 0, it means the probability is larger than 0.5, and if a logodds is smaller than 0 (negative), the probability is smaller than 0.5.

In summary, if we use a linear model to predict probabilities, we have the problem of predicted probabilities smaller than 0 and larger than 1 that are meaningless. If we use a linear model to predict odds we have the problem of predicted odds smaller than 0 that are meaningless: they are impossible! If on the other hand we use a linear model to predict *the natural logarithm of odds* (logodds), we have no problem whatsoever. We therefore propose to use a linear model to predict *logodds*: the natural logarithm of the odds that correspond to a particular probability.

Returning back to our example of the children passing the exam, suppose we have the following linear equation for the relationship between `age` and the logarithm of the odds of passing the exam

$$\texttt{logodds} = -33.15 + 0.42 \times \texttt{age},$$

This equation predicts that a child aged 70 months has a logodds of $-33.15 + 0.42 \times 70 = -3.75$. In order to transform that logodds back to a probability, we first have to take the exponential of the logodds[2] to get the odds:

$$\texttt{odds} = \exp(\texttt{logodds}) = e^{\texttt{logodds}} = e^{-3.75} = 0.02$$

An odds of 0.02 means that the odds of passing the exam is 0.02 to 1 (0.02:1). So out of $1 + 0.02 = 1.02$ times, we expect 0.02 successes and 1 failure. The probability of success is therefore $\frac{0.02}{1+0.02} = 0.02$. Thus, based on this equation, the expected probability of passing the exam for a child of 70 months equals 0.02.

If you find that easier, you can also memorise the following formula for the relationship between a logodds of x and the corresponding probability:

$$p_x = \frac{\exp(x)}{1 + \exp(x)} \tag{10.13}$$

Thus, if you have a logodds x of -3.75, the odds equals $\exp(-3.75) = 0.02$, and the corresponding probability is $\frac{0.02}{1+0.02} = 0.02$.

[2]If we know $\ln(x) = 60$, we have to infer that x equals e^{60}, because $\ln(e^{60}) = 60$ by definition of the natural logarithm, see previous footnote. Therefore, if we know that $\ln(x) = c$, we know that x equals e^c. The exponent of c, e^c, is often written as $\exp(c)$. So if we know that the logarithm of the odds equals c, $\texttt{logodds} = \ln(\texttt{oddsratio}) = c$, then the odds is equal to $\exp(c)$.

10.2.3 Logistic link function

In previous pages we have seen that logodds have the nice property of having meaningful values between $-\infty$ and $+\infty$. This makes them suitable for linear models. In essence, our linear model for our exam data in children might then look like this:

$$\texttt{logodds}_{pass} = b_0 + b_1\texttt{age} \tag{10.14}$$
$$Y \sim Bern(p_{pass}) \tag{10.15}$$

Note that we can write the odds as $p/(1-p)$, p is a probability (or a proportion). So the logodds that corresponds to the probability of passing the exam, p_{pass}, can be written as $ln\frac{p_{pass}}{1-p_{pass}}$, so that we have

$$\ln\frac{p_{pass}}{1-p_{pass}} = b_0 + b_1\texttt{age} \tag{10.16}$$
$$Y \sim Bern(p_{pass}) \tag{10.17}$$

Note that we do not have a residual any more: the randomness around the predicted values is no longer modelled using a residual e that is normally distributed, but is now modelled by a Y-variable with a Bernoulli distribution. Also note the strange relationship between the probability parameter p_{pass} for the Bernoulli distribution, and the dependent variable for the linear equation $b_0 + b_1 age$. The linear model predicts the logodds, but for the Bernoulli distribution, we use the probability. But it turns out that this model is very flexible and useful in many real-life problems. This model is often called a *logit* model: one often writes that the *logit of the probability* is predicted by a linear model.

$$\text{logit}(p_{pass}) = b_0 + b_1\texttt{age} \tag{10.18}$$
$$Y \sim Bern(p_{pass}) \tag{10.19}$$

In essence, the logit function transforms a p-value into a logodds:

$$\text{logit}(p) = \ln(\frac{p}{1-p})$$

So what does it look like, a linear model for logodds (or logits of probabilities)?

In Figure 10.9 we show a hypothetical example of a linear model for the logit of probabilities of passing an exam. These logits or logodds are predicted by age using a straight, linear regression line.

When we take all these predicted logodds and convert them back to probabilities, we obtain the plot in Figure 10.10. Note the change in the scale of the vertical axis, the rest of the plot is the same as in Figure 10.9.

Here again we see the S-shape relationship between probabilities and the logodds. We see that our model predicts probabilities close to 0 for very young

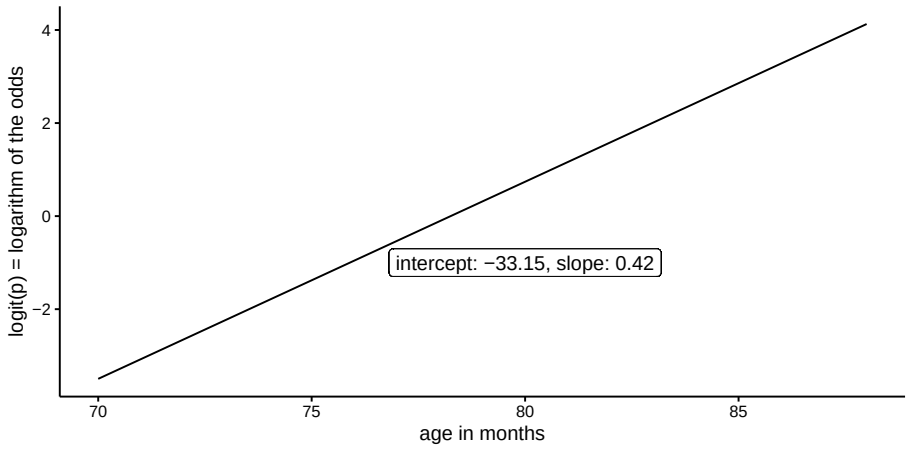


Figure 10.9: Example of a linear model for the logit of probabilities of passing an exam.

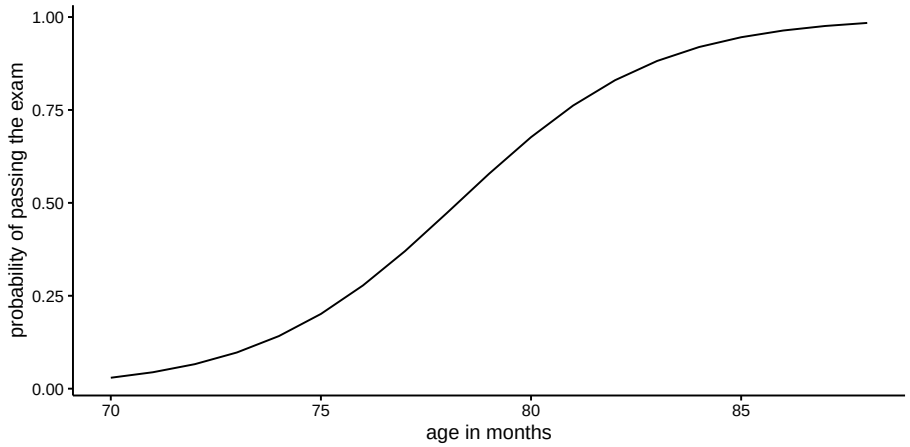


Figure 10.10: Example with logodds transformed into probabilties (vertical axis).

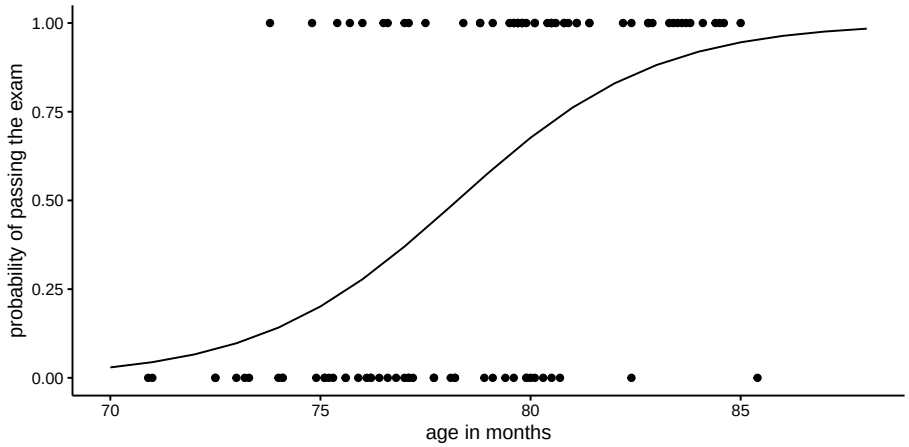


Figure 10.11: Transformed regression line and raw data points.

ages, and probabilities close to 1 for very old ages. There is a clear positive effect of age on the probability of passing the exam. But note that the relationship is not linear on the scale of the probabilities: it is linear on the scale of the logit of the probabilities, see Figure 10.9, but non-linear on the scale of the probabilities themselves, see Figure 10.10.

The curvilinear shape we see in Figure 10.10 is called a *logistic* curve. It is based on the logistic function: here p is a logistic function of `age` (and note the similarity with Equation 10.13):

$$p = \text{logistic}(b_0 + b_1\texttt{age}) = \frac{\exp(b_0 + b_1\texttt{age})}{1 + \exp(b_0 + b_1\texttt{age})}$$

In summary, if we go from logodds to probabilities, we use the logistic function, $\text{logistic}(x) = \frac{\exp(x)}{1+\exp(x)}$. If we go from probabilities to logodds, we use the logit function, $\text{logit}(p) = \ln\frac{p}{1-p}$. The logistic regression model is a generalised linear model with a logit link function, because the linear equation $b_0 + b_1X$ predicts the logit of a probability. It is also often said that we're dealing with a logistic link function, because the linear equation gives a value that we have to subject to the logistic function to get the probability. Both terms, logit link function and logistic link function are used.

If we go back to our data on the third-grade children that either passed or failed the exam, we see that this curve gives a description of our data, see Figure 10.11. The model predicts that around the age of 78 months, the probability of passing the exam is around 0.50. We indeed see in Figure 10.11 that around this age some children pass the exam (`score = 1`) and some don't (`score = 0`). On the basis of this analysis there seems to be a positive relationship between age in third-grade children and the probability of passing the exam in this sample.

What we have done here is a *logistic regression* of passing the exam on age. It

is called logistic because the curve in Figure 10.11 has a logistic shape. Logistic regression is one specific form of a *generalised linear model*. Here we have applied a generalised linear model with a so-called *logit link function*: instead of modelling dependent variable Y directly, we have modelled *the logit of the probabilities of obtaining a Y-value of 1*. There are many other link functions possible. One of them we will see in the chapter on generalised linear models for count data (Chapt. **??**). But first, let's see how logistic regression can be performed in R, and how we should interpret the output.

Table 10.1: Taking the train to Paris data.

train	age	sex_male	income	business
1	35.12	1	7544.00	1
1	66.66	1	7096.00	0
0	42.77	1	29261.00	1
0	72.63	0	24977.00	0
1	76.25	0	876.00	1
0	19.87	1	126943.00	1

10.3 Logistic regression in R

Imagine a data set on travellers from Amsterdam to Paris. From 1000 travellers, randomly sampled in 2017, we know whether they took the train to Paris, or whether they used other means of transportation. Of these travellers, we know their age, sex, yearly income, and whether they are travelling for business or not. Part of the data are displayed in Table 10.1. A score of 1 on the variable `train` means they took the train, a score of 0 means they did not.

Suppose we want to know what kind of people are more likely to take the train to Paris. We can use a logistic regression analysis to predict whether people take the train or not, on the basis of their age, sex, income, and main purpose of the trip.

Let's see whether income predicts the probability of taking the train. The function that we use in R is the `glm()` function, which stands for Generalised Linear Model. We can use the following code:

```
model.train <- data.train %>%
glm(train ~ income,
data = .,
family = binomial(link = logit))
```

`train` is our dependent variable, `income` is our independent variable, and these variables are stored in the data frame called `data.train`. But further we have to specify that we want to use the Bernoulli distribution and a logit link function. So `link = logit`. But why a binomial distribution? Well, the Bernoulli distribution (one coin flip) is a special case of the Binomial distribution (the distribution of several coin flips). So here we use a binomial distribution

for one coin flip, which is equivalent to a Bernoulli distribution. Actually, the code can be a little bit shorter, because the logit link function is the default option with the binomial distribution:

```
model.train <- data.train %>%
glm(train ~ income,
data = .,
family = binomial)
```

Below, we see the parameter estimates from this generalised linear model run on the train data.

```
model.train %>%
  tidy()

## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept) 90.0      32.5          2.77 0.00564
## 2 income      -0.00817   0.00297     -2.75 0.00603
```

The parameter estimates table from a `glm()` analysis looks very much like that of the ordinary linear model and the linear mixed model. An important difference is that the statistics shown are no longer t-statistics, but Z-statistics. This is because with logistic models, the ratio b_1/SE does not have a t-distribution. In ordinary linear models, the ratio b_1/SE has a t-distribution because in linear models, the variance of the residuals, σ_e^2, has to be estimated (as it is unknown). If the residual variance were known, b_1/SE would have a standard normal distribution. In logistic models, there is no σ_e^2 that needs to be estimated (it is by default 1), so the ratio b_1/SE has a standard normal distribution. One could therefore calculate a Z-statistic $Z = b_1/SE$ and see whether that value is smaller than 1.96 or larger than 1.96, if you want to test with a Type I error rate of 0.05.

The interpretation of the slope parameters is very similar to other linear models. Note that we have the following equation for the logistic model:

$$\begin{aligned} \text{logit}(p_{\texttt{train}}) &= b_0 + b_1 \texttt{income} \\ \texttt{train} &\sim Bern(p_{\texttt{train}}) \end{aligned} \tag{10.20}$$

If we fill in the values from the R output, we get

$$\begin{aligned} \text{logit}(p_{\texttt{train}}) &= 90.0 - 0.008 \times \texttt{income} \\ \texttt{train} &\sim Bern(p_{\texttt{train}}) \end{aligned} \tag{10.21}$$

We can interpret these results by making some predictions. Imagine a traveller with a yearly income of 11,000 Euros. Then the predicted logodds equals

$90.0 - 0.00817 \times 11000 = 0.13$. When we transform this back to a probability, we get $\frac{\exp(0.13)}{1+\exp(0.13)} = 0.53$. So this model predicts that for people with a yearly income of 11,000, about 53% of them take the train (if they travel at all, that is!).

Now imagine a traveller with a yearly income of 100,000. Then the predicted logodds equals $90.0 - 0.00817 \times 100000 = -727$. When we transform this back to a probability, we get $\frac{\exp(-727)}{1+\exp(-727)} = 0.00$. So this model predicts that for people with a yearly income of 100,000, close to none of them take the train. Going from 11,000 to 100,000 is a big difference. But the change in probabilities is also huge: the probability goes down from 0.53 to 0.00.

We found a difference in probability of taking the train for people with different incomes in this sample of 1000 travellers, but is there also an effect of income in the entire population of travellers between Amsterdam and Paris? The regression table shows us that the effect of income, -0.00817, is statistically significant at an α of 5%, $Z = -2.75, p < 0.01$. We can therefore reject the null-hypothesis that income is not related to whether people take the train or not. We conclude that in the population of travellers to Paris, a higher income is associated with a lower probability of travelling by train.

Note that similar to other linear models, the intercept can be interpreted as the predicted logodds for people that have values 0 for all other variables in the model. Therefore, 90.0 means in this case that the predicted logodds for people with zero income equals 90.0. This is equivalent to a probability of very close to 1.

Chapter 11

Introduction to big data analytics

11.1 Introduction

Previous chapters looked into traditional or classic data analysis: inference about a *population* using a limited sample of data with a limited number of variables. For instance, we might be interested in how large the effect is of a new kind of therapy for clinical depression in the population of *all* patients, based on a sample of 150 treated patients and 150 non-treated patients on a waiting list.

In many contexts, we are not interested in the effect of some intervention in a population, but in the *prediction* of events in the future. For example, we would like to predict which patients are likely to relapse into depression after an initially successful therapy. For such a prediction we might have a lot of variables. In fact, more variables than we could use in a straightforward linear model analysis.

In this age of *big data*, there are more often too many data then too few data about people. However, this wealth of data is often not nicely stored in data matrices. Data on patients for example are stored in different types of files, in different file formats, as text files, scans, X-rays, lab reports, perhaps even videotaped interviews. They may be stored at different hospitals or medical centres, so they need to be linked and combined without mixing them up. In short: data can be really messy. Moreover, data are not variables yet. *Data science* is about making data available for analysis. This field of research aims to extract knowledge and insight from structured and unstructured data. To do that, it draws from statistics, mathematics, computer science and information science.

The patient data example is a typical case of a set of unstructured data. From a large collection of pieces of texts (e.g., notes from psychiatric interviews and counselling, notes on prescriptions and adverse effects of medication, lab

reports) one has to distil a set of variables that could predict whether or not an individual patient would fall back into a second depressive period.

There are a couple of reasons why big data analytics is different from the data analysis framework discussed in previous chapters. These relate to 1) different types of questions, 2) the $p > n$ problem and 3) the problem of over-fitting.

First, the type of questions are different. In classic data analysis, you have a model with one or more model parameters, for example a regression coefficient, and the question is what the value is of that parameter in the population. Based on sample data, you draw inferences regarding the parameter value in the population. In contrast, typical questions in big data situations are about predictions for future data (e.g., how will the markets respond to the start of the hurricane season), or how to classify certain events (e.g., is a Facebook posting referring to a real event or is it "fake news"). In big data situations, such predictions or classifications are based on training data. In classic data analysis, inference is based on sample data.

Second, the type of data in big data settings allows for a far larger number of variables than in non-big data settings. In the patient data example, imagine the endless ways in which we could think of predicting relapse on the basis of the text data alone. We could take as predictor variables the number of counselling sessions, whether or not a tricyclic antidepressant was prescribed, whether or not a non-tricyclic antidepressant was prescribed, whether or not the word "mother" was mentioned in the sessions, the number of times the word "mother" was used in the sessions, how often the word "mother" was associated with the word "angry" or "anger" in the same sentence, and so on. The types of variables you could distil from such data is endless, so what to pick? And where to stop? So the first way in which big data analytics differs from classic data analysis is that a variable selection method has to be used. The analyst has to make a choice of what *features* of the raw data will be used in the analysis. Or, during the analysis itself, an algorithm can be used that picks those features that predict the outcome variable most efficiently. Usually there is a combination of both methods: there is an informed choice of what features in the data are likely to be most informative (e.g., the data analyst a priori believes that the specific words used in the interviews will be more informative about relapse than information contained in X-rays), and an algorithm that selects the most informative features out of this selection (e.g., the words "mother" and "angry"). One reason that variable selection is necessary is because statistical methods, like for example linear models, do not work when the number of variable is large relative to the number of cases. This is known as the $p > n$ problem, where p refers to the number of variables and n to the number of cases. We will come back to this problem below.

Third, because there is so much information available in big data situations, there is the likely danger of *over-fitting* your model. Maybe you have enough cases to include 1,000 predictor variables in your linear models, and they will run and give meaningful output, but then the model will be too much focused on the data that you have now, so that it will be very bad at predicting or classifying new events correctly. Therefore, another reason for limiting the number

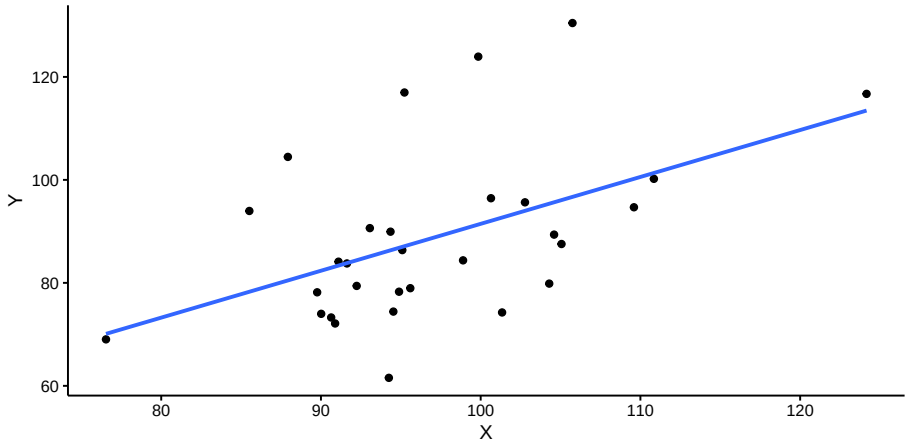


Figure 11.1: Ilustration of overfitting: a data set showing all 50 data points showing a linear relationship between variables X and Y.

of variables in your model is to prevent over-fitting. Limiting the number of variables in a model can be done by various variable selection methods, for example the LASSO in linear regression (Tibshirani, 1996). Over-fitting can also be countered by using so-called ensemble methods like boosting and random forest.

In order to check that you are not over-fitting, one generally splits the data into a *training* data set and a *test* data set (or validation data set). The training data are used to select the variables and fit the model (i.e., determine model parameters). For example, when a simple linear regression model is used to predict height in school children, the least squares estimates are determined based on the training data alone. Next, these estimates are used to predict new values for the test data. That is, we take the test data, use the predictor variables and plug them into the model equation and see what height values we predict for the children in the test data. Next, we compare these predicted values with the actually observed height measurements in these children. In the case of over-fitting, the model will show a very good fit (good predictions) to the training data but a poor fit to the test data (bad predictions). If there is no over-fitting, the predictions for the test data will not be much worse for the test data than for the training data.

As a small data example of the phenomenon of over-fitting, in Figure 11.1 we see a complete data set on variables X and Y for 100 observations. The actual relationship between the two variables can be described by a simple linear regression model $Y = X + e$, where e has a normal distribution with standard deviation 10. But suppose we don't know that, and we want to find a model that gives a good description for these data. We split the data set in half, and use a local polynomial regression fitting algorithm that best describes

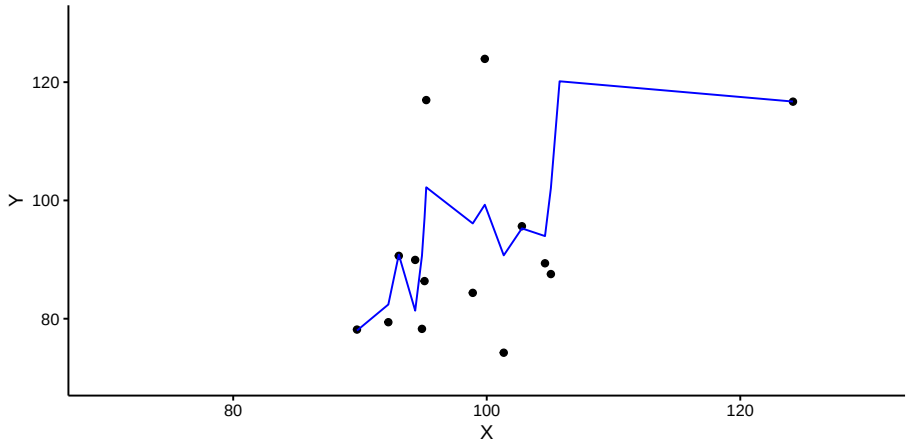


Figure 11.2: Ilustration of overfitting: only showing half the data points (training data) and a local polynomial regression model fitted to them.

the training data. The training data set and the model predictions, depicted as a blue line is given in Figure 11.2. It turns out that the correlation between the observed Y-values and the predicted values based on the model is pretty high: 0.78.

Next, we apply this model to the test data. These are depicted in Figure 11.3. The blue line are the predicted Y-values for the X-values in this data set, based on the model. We see that the blue line is not a good description of the pattern in the test data. This is also reflected in the much lower correlation between the observed Y-values and the predicted values in the test data: 0.71. Thus we see that the model is a good model for the training data, upon which the model was based, but the model is a terrible model for new data, even both data sets have the same origin. The training data were only randomly selected. The model was simply too much focused on the details in the data. Had we used a much simpler model, a linear regression model for example, the relative performance on the test data would be much better. The least squares equation predicts the Y-values in the training data with a correlation of 0.48. That is much worse than the splines, but we see that the model performs much better in the test data: there we see a correlation of 0.39. In sum: a complex model will always give better predictions than a simple model for training data. However, what is important is that a model will also show good predictions in test data. Then we see that often, a relatively simple model will perform better than a very complex model. This is due to the problem of over-fitting. The trade-off between the model complexity and over-fitting is also known as the *bias-variance trade-off*, where bias refers to the error that we make when we select the wrong model for the data, and variance refers to error that we make because we are limited to seeing only the training data.

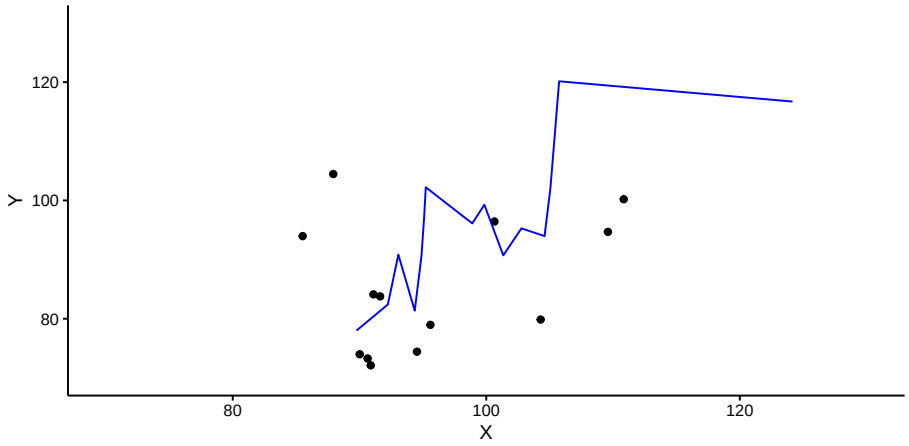


Figure 11.3: Ilustration of overfitting: only showing half the data points (test data) and a local polynomial regression model that does not describe these data well because it was based on the training data.

11.1.1 Model selection

In the preceding chapters, we discussed only the linear model and the extensions thereof: the linear mixed model and the generalized linear model. In big data analytics one uses these models too, but in addition there is a wealth of other models and methods, too. To name but the most well-known: decision trees, support vector machines, smoothing splines, generalized additive models, naive Bayes, and neural networks. Each model or method in itself has many subversions. A very important part of big data analytics is therefore *model selection*. Models vary in their *flexibility*, that is, how well they can fit the data. Remember that when we looked at multiple regression, a model with many predictors usually shows a larger R-squared than a model with fewer predictors. That is, it explains or predicts the sample data better. But now we know that this also brings the danger of over-fitting: it fits the sample or training data better, but not necessarily the test data and the data that we want to predict in future applications. The same is true for simple models and more flexible models. A linear model with two additive predictors is a relatively simple model that might not explain so much variance in the training data. A neural network is an example of a very flexible model that might be much more able to capture variance in training data that shows a large amount of complexity, for example a complicated interaction pattern. We therefore have to make a trade-off between model flexibility and over-fitting that is dependent on each type of problem. If reality is complex we need a complex model to make the right predictions, but if reality is simple and we apply a complex model, we run the risk of over-fitting. If reality is complex and we use a simple model, we run the risk of bad predictions. In order to strike the right balance and find the optimal level of

complexity for our training data, one often uses *cross-validation*.

11.1.2 Cross-validation

Cross-validation is a form of a *re-sampling method*. In re-sampling methods, different subsets of the training data are used to fit the same model or different models, or different versions of a model. There are different forms of cross-validation, but here we discuss k-fold cross-validation. In k-fold cross-validation, the training data are split randomly into k groups (*folds*) of approximately equal size. The model is then fit k times, each time leaving out the data from one of the k groups. Each time, predictions are made for the data in the group that is left out of the analysis. And each time we assess how good these predictions are, for example by determining the residuals and computing the mean squared error (MSE). With k groups, we then have k MSEs, and we can compute the mean MSE. If we do this cross-validation for several models, we can see which model has the lowest mean MSE. That is the model that on average shows the best prediction. This should not lead to over-fitting, because by the random sampling into k sub-samples, we are no longer dependent on one particular subset of the data. Usually, a value of 5 or 10 is used for k.

11.1.3 The $p > n$ problem

Table 11.1: Small data set illustrating the p larger than n problem.

X	Y
0.00	0.64
1.00	2.08
2.00	3.33

Suppose we have a data set consisting of 2 variables, X and Y, with 3 observations. This data set is tabulated in Table 11.1. So we have $n = 3$ and $p = 2$. If we plot the data and fit a regression line with the least squares criterion, we encounter no problem, see Figure 11.4. This is because p is smaller than n. Now let's see what happens when p and n are of equal size. Suppose we omit the first observation and plot only the second and third observations. These are plotted in Figure 11.5. If we now fit a line using the least squares criterion, we see that we can only fit a line without residuals: the line shows a perfect fit. You can probably imagine that for any two data points, a linear line will always show a perfect fit without residuals. The variance of the residuals will be 0. The software that you use for fitting such a model will give you some sort of warning. R will say that there are no residual degrees of freedom. SPSS will say nothing special but will also show a 0 for the number of residual degrees of freedom. More obviously, the output will give you an intercept and a slope, but there will be no standard errors, and hence no t-values and no p-values.

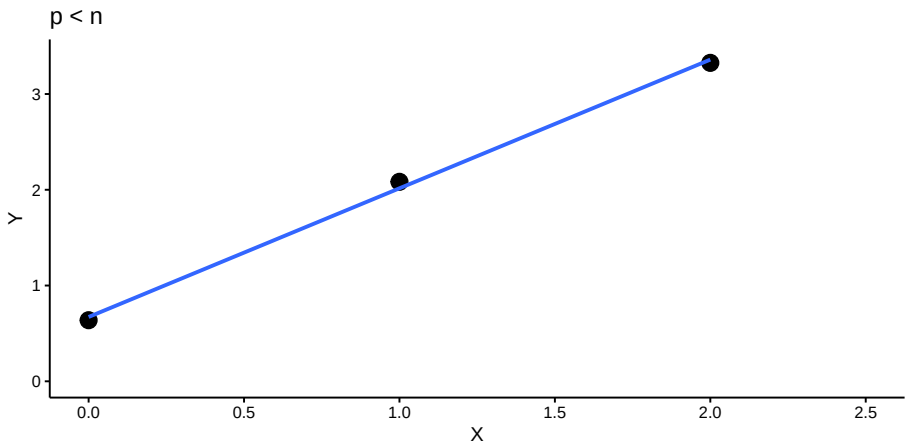


Figure 11.4: $p < n$: There is a unique solution that fits the least squares criterion. No problem whatsoever.

Therefore, you can say something about the intercept and slope in the sample, but you cannot say anything about the intercept and slope in the population.

The situation will be even worse when you have only 1 data point. Suppose we only have the second data point, which we plot in Figure 11.6. If we then try to fit a regression line, we will see that the software will refuse to estimate a slope parameter. It will be fixed to 0, so that an intercept only model will be fitted (see Section 5.15). It simply is impossible for the software to decide what regression line to pick that goes through this one data point: there is an infinite number of regression lines that go through this data point! Therefore, for a two-parameter model like a regression model (the two parameters being the intercept and slope), you need at least two data points for the model to run, and at least three data points to get standard errors and do inference.

The same is true for larger n and larger models. For example, a multiple regression model with 10 predictor variables together with an intercept will have 11 parameters. Such a model will not give standard errors when you have 11 observations in your data matrix, and it will not run if you have fewer than 11 observations.

In sum, the number of data points, n, should always exceed the number of parameters in your model. That means that if you have a lot of variables in your data file, you cannot always use them in your analysis, because you simply do not have enough rows in your data matrix to estimate the model parameters.

11.1.4 Steps in big data analytics

Now that we are familiar with the differences between classical data analysis and big data analytics and the major concepts, we can look at how all the elements of big data analytics fit together. In big data problems, we often see

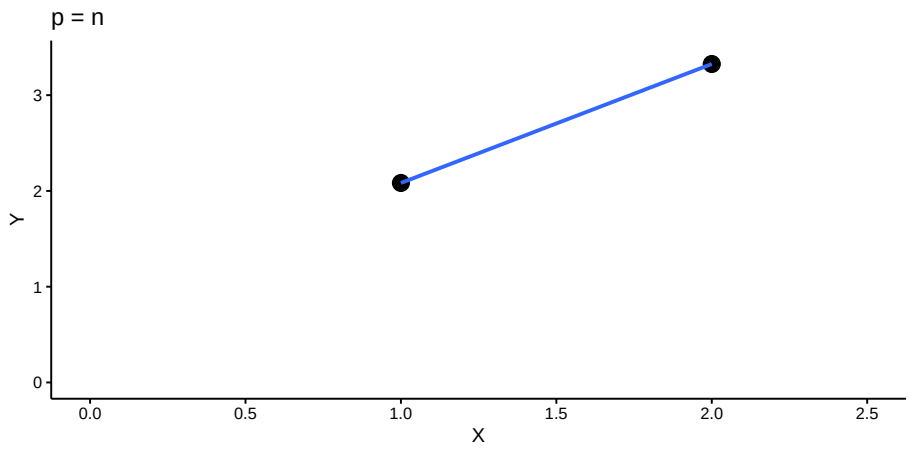


Figure 11.5: $p = n$: There is a unique solution, but there are no degrees of freedom left. Standard errors cannot be determined, so no inference regarding the population parameters is possible.

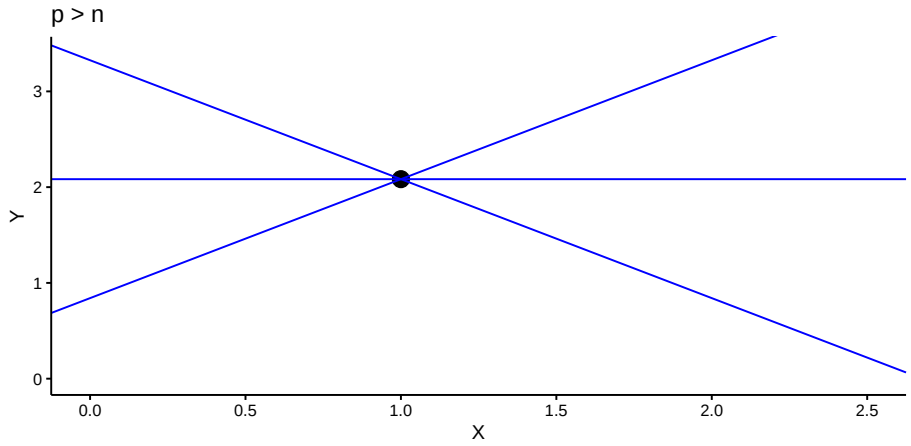


Figure 11.6: $p > n$: There is an infinite number of lines through the data point, but there is no criterion that determines which is best. The problem of the least squares regression line is not defined with only one data point.

the following steps:

1. Problem identification. You need to know what the problem is: what do you want to know? What do you want to predict? How good does the prediction have to be? How fast does it have to be: real-time?

2. Selection of data sources. Depending on what you want to know or predict, you have to think about possible sources of information. Are there any websites that already contain part of the information that you are looking for? Are there any databases you can get access to?

3. Feature selection. From the data sources you have access to, what features are of interest? For example, from spoken interviews, are you mainly interested in the words spoken by the patient? Or perhaps interested in the length of periods of silence, or perhaps in changes in pitch? There are so many features you could extract from data.

4. Construction of a data matrix. Once you have decided what features you want to extract from the raw data, you have to put this information into a data matrix. You have to decide what to put in the rows (your *units of observation*), and in the columns (the features, now variables). So what is now your variable: this could be the length of one period of silence within one interview for a particular patient. But it could also be the average pitch for a 1-minute interval in one interview for one particular patient.

5. Training and test (validation) data set. In order to check that we are not over-fitting, and to make sure that our model will work for future data, we divide our data set (our data matrix) into two parts: training data and test data. This is done by taking a random sample of all the data that we have. Usually, a random sample of 70% is used for training, and the remaining 30% is used for testing (validating) the model. We set the test data aside and will only look at the training data.

6. Model selection. In data science it is very common to try out various models or various sub-models on the training data to see which model fits the data best. To make sure that we do not over-fit the data, we use some form of cross-validation, where we split up the training data even further.

7. Build the model. Once we know what model works best for our training data, we fit that model on the training model. This fitted model is our final model.

8. Validate the model. The final test is whether this final model will work on new data. We don't have new data, but we have put away some of our data as test data (validation data). These data can now be used as a substitute for new data to estimate how well our model will work with future data.

9. Interpret the result and evaluate. There will be always some over-fitting, so the performance on the test data will always be worse than on the training data. But is the performance good enough to be satisfied with the model? Is the model useful for daily practice? If not, maybe the data sources and feature selection steps should be reconsidered. Another important aspect of statistical learning is interpretability. There are some very powerful models and methods around that are capable of very precise predictions. However, the problem with these models and methods is that they are hard to interpret: they are black boxes in that they make predictions that cannot be explained by even the data analysts themselves. Any decisions are therefore hard to justify, which brings ethical issues. For instance, what would you say if an algorithm would determine on the basis of all your life's data that you will not be a successful student? Of course you would want to know on the basis of what data exactly that decision is based. Your make-up? Your height? The colour of your skin? Last year's grades? Of course it would matter to you what variables are used and how. Recent research has focused on how to make complicated models and methods easier to interpret and help data analysts evaluate the usefulness and applicability of their results and communicate them to others.

Appendices

Appendix A

Cumulative probabilities for the standard normal distribution

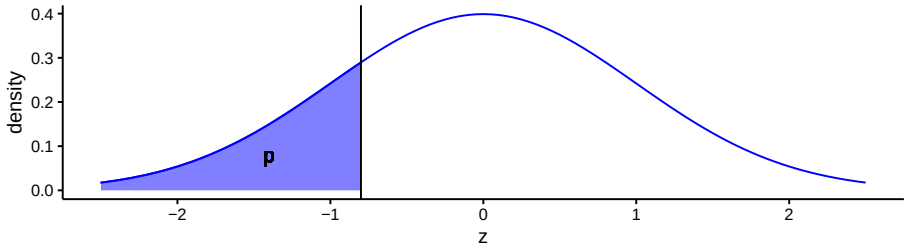


Table A.1: Cumulative proportions (p) for the standard normal distribution.

z	p	z	p	z	p	z	p	z	p	z	p
-4.00	0.0000	-1.43	0.0764	-0.71	0.2389	0.01	0.5040	0.73	0.7673	1.45	0.9265
-3.80	0.0001	-1.42	0.0778	-0.70	0.2420	0.02	0.5080	0.74	0.7704	1.46	0.9279
-3.60	0.0002	-1.41	0.0793	-0.69	0.2451	0.03	0.5120	0.75	0.7734	1.47	0.9292
-3.40	0.0003	-1.40	0.0808	-0.68	0.2483	0.04	0.5160	0.76	0.7764	1.48	0.9306
-3.20	0.0007	-1.39	0.0823	-0.67	0.2514	0.05	0.5199	0.77	0.7794	1.49	0.9319
-3.00	0.0013	-1.38	0.0838	-0.66	0.2546	0.06	0.5239	0.78	0.7823	1.50	0.9332
-2.90	0.0019	-1.37	0.0853	-0.65	0.2578	0.07	0.5279	0.79	0.7852	1.51	0.9345
-2.80	0.0026	-1.36	0.0869	-0.64	0.2611	0.08	0.5319	0.80	0.7881	1.52	0.9357
-2.70	0.0035	-1.35	0.0885	-0.63	0.2643	0.09	0.5359	0.81	0.7910	1.53	0.9370
-2.60	0.0047	-1.34	0.0901	-0.62	0.2676	0.10	0.5398	0.82	0.7939	1.54	0.9382
-2.50	0.0062	-1.33	0.0918	-0.61	0.2709	0.11	0.5438	0.83	0.7967	1.55	0.9394
-2.40	0.0082	-1.32	0.0934	-0.60	0.2743	0.12	0.5478	0.84	0.7995	1.56	0.9406
-2.30	0.0107	-1.31	0.0951	-0.59	0.2776	0.13	0.5517	0.85	0.8023	1.57	0.9418
-2.20	0.0139	-1.30	0.0968	-0.58	0.2810	0.14	0.5557	0.86	0.8051	1.58	0.9429
-2.10	0.0179	-1.29	0.0985	-0.57	0.2843	0.15	0.5596	0.87	0.8078	1.59	0.9441
-2.00	0.0228	-1.28	0.1003	-0.56	0.2877	0.16	0.5636	0.88	0.8106	1.60	0.9452
-1.99	0.0233	-1.27	0.1020	-0.55	0.2912	0.17	0.5675	0.89	0.8133	1.61	0.9463
-1.98	0.0239	-1.26	0.1038	-0.54	0.2946	0.18	0.5714	0.90	0.8159	1.62	0.9474
-1.97	0.0244	-1.25	0.1056	-0.53	0.2981	0.19	0.5753	0.91	0.8186	1.63	0.9484
-1.96	0.0250	-1.24	0.1075	-0.52	0.3015	0.20	0.5793	0.92	0.8212	1.64	0.9495
-1.95	0.0256	-1.23	0.1093	-0.51	0.3050	0.21	0.5832	0.93	0.8238	1.65	0.9505
-1.94	0.0262	-1.22	0.1112	-0.50	0.3085	0.22	0.5871	0.94	0.8264	1.66	0.9515
-1.93	0.0268	-1.21	0.1131	-0.49	0.3121	0.23	0.5910	0.95	0.8289	1.67	0.9525
-1.92	0.0274	-1.20	0.1151	-0.48	0.3156	0.24	0.5948	0.96	0.8315	1.68	0.9535
-1.91	0.0281	-1.19	0.1170	-0.47	0.3192	0.25	0.5987	0.97	0.8340	1.69	0.9545
-1.90	0.0287	-1.18	0.1190	-0.46	0.3228	0.26	0.6026	0.98	0.8365	1.70	0.9554
-1.89	0.0294	-1.17	0.1210	-0.45	0.3264	0.27	0.6064	0.99	0.8389	1.71	0.9564
-1.88	0.0301	-1.16	0.1230	-0.44	0.3300	0.28	0.6103	1.00	0.8413	1.72	0.9573
-1.87	0.0307	-1.15	0.1251	-0.43	0.3336	0.29	0.6141	1.01	0.8438	1.73	0.9582
-1.86	0.0314	-1.14	0.1271	-0.42	0.3372	0.30	0.6179	1.02	0.8461	1.74	0.9591
-1.85	0.0322	-1.13	0.1292	-0.41	0.3409	0.31	0.6217	1.03	0.8485	1.75	0.9599
-1.84	0.0329	-1.12	0.1314	-0.40	0.3446	0.32	0.6255	1.04	0.8508	1.76	0.9608

Continued on next page

Table A.1: Cumulative proportions (p) for the standard normal distribution.

z	p	z	p	z	p	z	p	z	p	z	p
-1.83	0.0336	-1.11	0.1335	-0.39	0.3483	0.33	0.6293	1.05	0.8531	1.77	0.9616
-1.82	0.0344	-1.10	0.1357	-0.38	0.3520	0.34	0.6331	1.06	0.8554	1.78	0.9625
-1.81	0.0351	-1.09	0.1379	-0.37	0.3557	0.35	0.6368	1.07	0.8577	1.79	0.9633
-1.80	0.0359	-1.08	0.1401	-0.36	0.3594	0.36	0.6406	1.08	0.8599	1.80	0.9641
-1.79	0.0367	-1.07	0.1423	-0.35	0.3632	0.37	0.6443	1.09	0.8621	1.81	0.9649
-1.78	0.0375	-1.06	0.1446	-0.34	0.3669	0.38	0.6480	1.10	0.8643	1.82	0.9656
-1.77	0.0384	-1.05	0.1469	-0.33	0.3707	0.39	0.6517	1.11	0.8665	1.83	0.9664
-1.76	0.0392	-1.04	0.1492	-0.32	0.3745	0.40	0.6554	1.12	0.8686	1.84	0.9671
-1.75	0.0401	-1.03	0.1515	-0.31	0.3783	0.41	0.6591	1.13	0.8708	1.85	0.9678
-1.74	0.0409	-1.02	0.1539	-0.30	0.3821	0.42	0.6628	1.14	0.8729	1.86	0.9686
-1.73	0.0418	-1.01	0.1562	-0.29	0.3859	0.43	0.6664	1.15	0.8749	1.87	0.9693
-1.72	0.0427	-1.00	0.1587	-0.28	0.3897	0.44	0.6700	1.16	0.8770	1.88	0.9699
-1.71	0.0436	-0.99	0.1611	-0.27	0.3936	0.45	0.6736	1.17	0.8790	1.89	0.9706
-1.70	0.0446	-0.98	0.1635	-0.26	0.3974	0.46	0.6772	1.18	0.8810	1.90	0.9713
-1.69	0.0455	-0.97	0.1660	-0.25	0.4013	0.47	0.6808	1.19	0.8830	1.91	0.9719
-1.68	0.0465	-0.96	0.1685	-0.24	0.4052	0.48	0.6844	1.20	0.8849	1.92	0.9726
-1.67	0.0475	-0.95	0.1711	-0.23	0.4090	0.49	0.6879	1.21	0.8869	1.93	0.9732
-1.66	0.0485	-0.94	0.1736	-0.22	0.4129	0.50	0.6915	1.22	0.8888	1.94	0.9738
-1.65	0.0495	-0.93	0.1762	-0.21	0.4168	0.51	0.6950	1.23	0.8907	1.95	0.9744
-1.64	0.0505	-0.92	0.1788	-0.20	0.4207	0.52	0.6985	1.24	0.8925	1.96	0.9750
-1.63	0.0516	-0.91	0.1814	-0.19	0.4247	0.53	0.7019	1.25	0.8944	1.97	0.9756
-1.62	0.0526	-0.90	0.1841	-0.18	0.4286	0.54	0.7054	1.26	0.8962	1.98	0.9761
-1.61	0.0537	-0.89	0.1867	-0.17	0.4325	0.55	0.7088	1.27	0.8980	1.99	0.9767
-1.60	0.0548	-0.88	0.1894	-0.16	0.4364	0.56	0.7123	1.28	0.8997	2.00	0.9772
-1.59	0.0559	-0.87	0.1922	-0.15	0.4404	0.57	0.7157	1.29	0.9015	2.10	0.9821
-1.58	0.0571	-0.86	0.1949	-0.14	0.4443	0.58	0.7190	1.30	0.9032	2.20	0.9861
-1.57	0.0582	-0.85	0.1977	-0.13	0.4483	0.59	0.7224	1.31	0.9049	2.30	0.9893
-1.56	0.0594	-0.84	0.2005	-0.12	0.4522	0.60	0.7257	1.32	0.9066	2.40	0.9918
-1.55	0.0606	-0.83	0.2033	-0.11	0.4562	0.61	0.7291	1.33	0.9082	2.50	0.9938
-1.54	0.0618	-0.82	0.2061	-0.10	0.4602	0.62	0.7324	1.34	0.9099	2.60	0.9953
-1.53	0.0630	-0.81	0.2090	-0.09	0.4641	0.63	0.7357	1.35	0.9115	2.70	0.9965
-1.52	0.0643	-0.80	0.2119	-0.08	0.4681	0.64	0.7389	1.36	0.9131	2.80	0.9974
-1.51	0.0655	-0.79	0.2148	-0.07	0.4721	0.65	0.7422	1.37	0.9147	2.90	0.9981
-1.50	0.0668	-0.78	0.2177	-0.06	0.4761	0.66	0.7454	1.38	0.9162	3.00	0.9987
-1.49	0.0681	-0.77	0.2206	-0.05	0.4801	0.67	0.7486	1.39	0.9177	3.20	0.9993
-1.48	0.0694	-0.76	0.2236	-0.04	0.4840	0.68	0.7517	1.40	0.9192	3.40	0.9997
-1.47	0.0708	-0.75	0.2266	-0.03	0.4880	0.69	0.7549	1.41	0.9207	3.60	0.9998
-1.46	0.0721	-0.74	0.2296	-0.02	0.4920	0.70	0.7580	1.42	0.9222	3.80	0.9999
-1.45	0.0735	-0.73	0.2327	-0.01	0.4960	0.71	0.7611	1.43	0.9236	4.00	1.0000
-1.44	0.0749	-0.72	0.2358	0.00	0.5000	0.72	0.7642	1.44	0.9251	4.01	1.0000

Appendix B

Critical values for the t-distribution

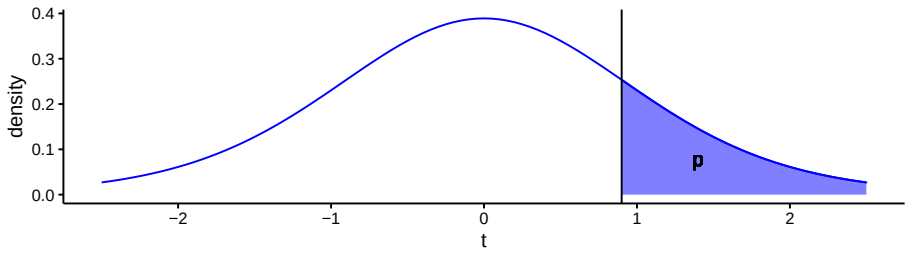


Table B.1: Values of the t-distribution, given the degrees of freedom (rows) and tail probability p (columns). These can be used for critical values for a given confidence level.

df	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.963	3.078	6.314	12.71	15.90	31.82	63.66	127.3	318.3	636.6
2	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.22	12.92
4	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
120	1.041	1.289	1.658	1.980	2.076	2.358	2.617	2.860	3.160	3.373
10000	1.036	1.282	1.645	1.960	2.054	2.327	2.576	2.808	3.091	3.291
	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%

Confidence level